

ClinicalNLP 2025

**The 7th Workshop on Clinical Natural Language Processing  
(ClinicalNLP)**

**Proceedings of the Workshop**

October 30, 2025

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
317 Sidney Baker St. S  
Suite 400 - 134  
Kerrville, TX 78028  
USA  
Tel: +1-855-225-1962  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN None

## Preface

This volume contains papers from the 7th Workshop on Clinical Natural Language Processing (Clinical NLP 2025), held online.

Clinical text is growing rapidly as electronic health records (EHRs) become pervasive. Much of the information recorded in a clinical encounter is located exclusively in provider narrative notes, which makes them indispensable for supplementing structured clinical data to better understand patient state and care provided. The goal of this workshop is to establish a regular event that brings together researchers interested in developing state-of-the-art methods for the clinical domain. The focus is on improving NLP technology to enable clinical applications via information extraction and modeling of narrative provider notes from electronic health records, patient encounter transcripts, and other clinical narratives.

For Clinical NLP 2025 we focused on three shared tasks, challenging researchers around the world to develop new approaches to solve clinical NLP problems: chemotherapy timelines extraction, medical order extraction, and wound care visual question answering. We received a total of 8 participant submissions to the shared tasks, all of which were accepted as oral presentations.

# Program Committee

## Program Chairs

Asma Ben Abacha, Steven Bethard, Danielle Bitterman, Tristan Naumann, Kirk Roberts

## Reviewers

Abhinand Balachandran, Asma Ben Abacha, Jean-Philippe Corbeil, Bavana Durgapraveen, Kaiwen He, Harry Hochheiser, Vijay Raj Jain, A H M Rezaul Karim, Parth Mehta, Zhaoyi Sun, Ozlem Uzuner, V.G.Vinod Vydiswaran, Jiarui Yao, Wen-wai Yim, WonJin Yoon, Tianmai M. Zhang, Zhe Zhao

# Keynote Talk

## An AI Agenda to Modernize Healthcare Delivery

**Dr. Karandeep Singh**  
UC San Diego

**Abstract:** In this talk, Dr. Karandeep Singh will describe healthcare delivery challenges that cut across multiple clinical services and care settings and culminate in the phenomenon of patients “boarding” in the emergency department. Is boarding really a fixable problem? If so, how can artificial intelligence (AI) help? He will present key areas where AI can play a role in informing and redesigning how care is delivered, touching on its roles in measurement, simulation, prediction, and automation. The talk is intended to build skills in health systems-level thinking and AI-supported quality improvement.

**Bio:** Karandeep Singh, MD, MMSc is the Joan and Irwin Jacobs Chancellor’s Endowed Chair in Digital Health Innovation and Associate Professor in Biomedical Informatics at UC San Diego, where he also serves as Chief Health AI Officer for UC San Diego Health. In these roles, Dr. Singh leads AI initiatives within the Jacobs Center for Health Innovation and oversees AI strategy and governance for the health system.

He completed his internal medicine residency at UCLA Medical Center, where he served as chief resident, and a nephrology fellowship in the combined Brigham and Women’s Hospital and Massachusetts General Hospital program. He completed his medical education at the University of Michigan Medical School and holds a master’s degree in medical sciences in Biomedical Informatics from Harvard Medical School.

## Table of Contents

<i>Overview of the 2025 Shared Task on Chemotherapy Treatment Timeline Extraction</i> Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli T Goldner and Guergana K Savova . . . . .	1
<i>Overview of the MEDIQA-OE 2025 Shared Task on Medical Order Extraction from Doctor-Patient Consultations</i> Jean-Philippe Corbeil, Asma Ben Abacha, Jerome Tremblay, Phillip Swazinna, Akila Jeesson Daniel, Miguel Del-Agua and Francois Beaulieu . . . . .	11
<i>Overview of the MEDIQA-WV 2025 Shared Task on Woundcare Visual Question Answering</i> Wen-wai Yim, Asma Ben Abacha, Meliha Yetisgen and Fei Xia . . . . .	17
<i>Team NLP4Health at ChemoTimelines 2025: Finetuning Large Language Models for Temporal Relation Extractions from Clinical Notes</i> Zhe Zhao and V.G.Vinod Vydiswaran . . . . .	22
<i>TEAM UAB at Chemotherapy Timelines 2025: Integrating Encoders and Large Language Models for Chemotherapy Timelines Generation</i> Vijay Raj Jain, Chris Coffee, Kaiwen He, Remy Cron, Micah D. Cochran, Luis Mansilla-Gonzalez, Akhil Nadimpalli, Danish Murad and John D Osborne . . . . .	30
<i>UW-BioNLP at ChemoTimelines 2025: Thinking, Fine-Tuning, and Dictionary-Enhanced LLM Systems for Chemotherapy Timeline Extraction</i> Tianmai M. Zhang, Zhaoyi Sun, Sihang Zeng, Chenxi LI, Neil F. Abernethy, Barbara D. Lam, Fei Xia and Meliha Yetisgen . . . . .	40
<i>MasonNLP at MEDIQA-OE 2025: Assessing Large Language Models for Structured Medical Order Extraction</i> A H M Rezaul Karim and Ozlem Uzuner . . . . .	57
<i>EXL Health AI Lab at MEDIQA-OE 2025: Evaluating Prompting Strategies with MedGemma for Medical Order Extraction</i> Abhinand Balachandran, Bavana Durgapaveen, Gowsikkan Sikkan Sudhagar, Vidhya Varshany J S and Sriram Rajkumar . . . . .	68
<i>PNLP at MEDIQA-OE 2025: A Zero-Shot Prompting Strategy with Gemini for Medical Order Extraction</i> Parth Mehta . . . . .	75
<i>MasonNLP at MEDIQA-WV 2025: Multimodal Retrieval-Augmented Generation with Large Language Models for Medical VQA</i> A H M Rezaul Karim and Ozlem Uzuner . . . . .	84
<i>EXL Health AI Lab at MEDIQA-WV 2025: Mined Prompting and Metadata-Guided Generation for Wound Care Visual Question Answering</i> Bavana Durgapaveen, Sornaraj Sivasankaran, Abhinand Balachandran and Sriram Rajkumar .	95

# Program

**Thursday, October 30, 2025**

09:00 - 09:10 *Opening Remarks*

09:10 - 10:10 *Keynote*

10:10 - 10:30 *Keynote Q&A*

10:30 - 11:30 *Shared Task Overviews (15 min talks + 5 min questions)*

*Overview of the 2025 Shared Task on Chemotherapy Treatment Timeline Extraction*

Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli T Goldner and Guergana K Savova

*Overview of the MEDIQA-OE 2025 Shared Task on Medical Order Extraction from Doctor-Patient Consultations*

Jean-Philippe Corbeil, Asma Ben Abacha, Jerome Tremblay, Phillip Swazinna, Akila Jeesson Daniel, Miguel Del-Agua and Francois Beaulieu

*Overview of the MEDIQA-WV 2025 Shared Task on Woundcare Visual Question Answering*

Wen-wai Yim, Asma Ben Abacha, Meliha Yetisgen and Fei Xia

11:30 - 12:00 *Break*

12:00 - 12:45 *Timeline Extraction System Overviews (10 min talks + 5 min questions)*

*Team NLP4Health at ChemoTimelines 2025: Finetuning Large Language Models for Temporal Relation Extractions from Clinical Notes*

Zhe Zhao and V.G.Vinod Vydiswaran

*TEAM UAB at Chemotherapy Timelines 2025: Integrating Encoders and Large Language Models for Chemotherapy Timelines Generation*

Vijay Raj Jain, Chris Coffee, Kaiwen He, Remy Cron, Micah D. Cochran, Luis Mansilla-Gonzalez, Akhil Nadimpalli, Danish Murad and John D Osborne

*UW-BioNLP at ChemoTimelines 2025: Thinking, Fine-Tuning, and Dictionary-Enhanced LLM Systems for Chemotherapy Timeline Extraction*

Tianmai M. Zhang, Zhaoyi Sun, Sihang Zeng, Chenxi LI, Neil F. Abernethy, Barbara D. Lam, Fei Xia and Meliha Yetisgen

12:45 - 13:30 *MEDIQA-OE System Overviews (10 min talks + 5 min questions)*

**Thursday, October 30, 2025 (continued)**

*MasonNLP at MEDIQA-OE 2025: Assessing Large Language Models for Structured Medical Order Extraction*

A H M Rezaul Karim and Ozlem Uzuner

*EXL Health AI Lab at MEDIQA-OE 2025: Evaluating Prompting Strategies with MedGemma for Medical Order Extraction*

Abhinand Balachandran, Bavana Durgapaveen, Gowsikkan Sikkan Sudhagar, Vidhya Varshany J S and Sriram Rajkumar

*PNLP at MEDIQA-OE 2025: A Zero-Shot Prompting Strategy with Gemini for Medical Order Extraction*

Parth Mehta

13:30 - 14:00 *MEDIQA-WV System Overviews (10 min talks + 5 min questions)*

*MasonNLP at MEDIQA-WV 2025: Multimodal Retrieval-Augmented Generation with Large Language Models for Medical VQA*

A H M Rezaul Karim and Ozlem Uzuner

*EXL Health AI Lab at MEDIQA-WV 2025: Mined Prompting and Metadata-Guided Generation for Wound Care Visual Question Answering*

Bavana Durgapaveen, Sornaraj Sivasankaran, Abhinand Balachandran and Sriram Rajkumar

14:00 - 14:30 *Closing Session*



# Overview of the 2025 Shared Task on Chemotherapy Treatment Timeline Extraction

\*Jiarui Yao<sup>1</sup>, \*Harry Hochheiser<sup>2</sup>, WonJin Yoon<sup>1</sup>, Eli Goldner<sup>1</sup>, Guergana Savova<sup>1</sup>

<sup>1</sup>Boston Children’s Hospital and Harvard Medical School <sup>2</sup>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 address two subtasks. **Subtask 1** focused on extracting temporal relations and creating timelines, given documents and gold-standard events and temporal expressions. **Subtask 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>

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 <sup>†</sup>, 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:  $\langle \textit{chemotherapy\_event}, \textit{temporal\_relation}, \textit{time\_expression} \rangle$ , 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:

$\langle \textit{Carboplatin}, \textit{CONTAINS-1}, 2013-09-30 \rangle$ ,

$\langle \textit{Taxol}, \textit{CONTAINS-1}, 2013-09-30 \rangle$ ,

$\langle \textit{Carboplatin}, \textit{CONTAINS-1}, 2013-10-20 \rangle$ ,

$\langle \textit{Taxol}, \textit{CONTAINS-1}, 2013-10-20 \rangle$ .

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  $\langle \textit{chemotherapy\_event}, \textit{temporal\_relation}, \textit{time\_expression} \rangle$  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

<sup>†</sup><https://sites.google.com/view/chemotimelines2025/>

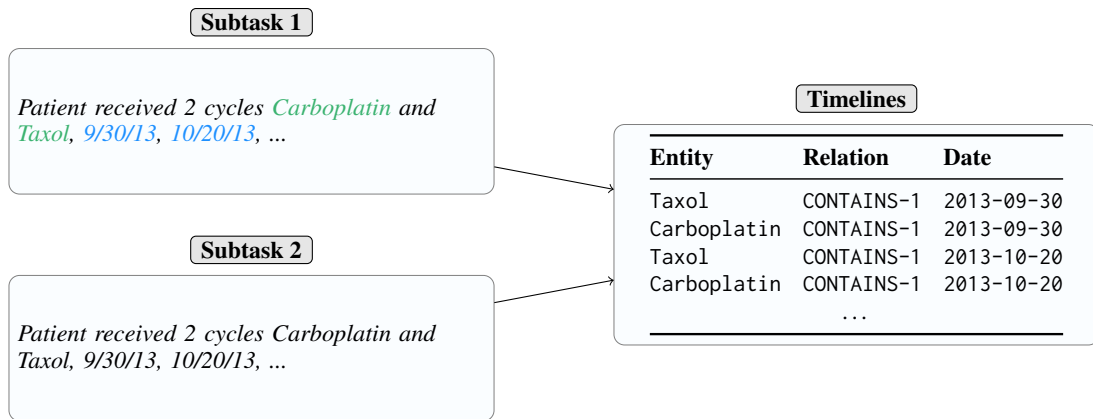


Figure 1: Illustration of the two subtasks in the 2024 and 2025 Chemotherapy Treatment Timeline Extraction shared task. The input of Subtask 1 is patient notes with gold events (highlighted in green) and time expressions (highlighted in blue). The input of Subtask 2 is patient notes only. The output 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<sup>‡</sup>.

### 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

<sup>‡</sup><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<sup>§</sup> 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<sup>¶</sup> (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 <sup>||</sup>.

## 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 <sup>\*\*</sup>. The system summaries below are based on these descriptions. Details of

<sup>§</sup><https://github.com/HealthNLPorg/chemoTimelinesEval>

<sup>¶</sup><https://ctakes.apache.org>

<sup>||</sup><https://github.com/HealthNLPorg/chemoTimelinesBaselineSystem>

<sup>\*\*</sup><https://sites.google.com/view/chemotimelines2025/submission-of-test-output>

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 <sup>††</sup> and LLMs to extract relevant elements for Subtask 1. They applied GPT-4o (OpenAI et al., 2024) and GPT-4.1 <sup>‡‡</sup> 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

<sup>††</sup><https://www.brimanalytics.com/>

<sup>‡‡</sup><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-Thought 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 supervised 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 interven-

ing 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.

Methodwise, 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. To 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 sensitive 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, Yanzhi 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.

# Overview of the MEDIQA-OE 2025 Shared Task on Medical Order Extraction from Doctor-Patient Consultations

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeesson Daniel, Miguel Del-Agua, Francois Beaulieu

Microsoft Healthcare & Life Sciences

Correspondence: {jcorbeil,abenabacha}@microsoft.com

## Abstract

Clinical documentation increasingly uses automatic speech recognition and summarization, yet converting conversations into actionable medical orders for Electronic Health Records remains unexplored. A solution to this problem can significantly reduce the documentation burden of clinicians and directly impact downstream patient care. We introduce the MEDIQA-OE 2025 shared task, the first challenge on extracting medical orders from doctor-patient conversations. Six teams participated in the shared task and experimented with a broad range of approaches, and both closed- and open-weight large language models (LLMs). In this paper, we describe the MEDIQA-OE task, dataset, final leaderboard ranking, and participants' solutions.

## 1 Introduction

In recent years, the burden of clinical documentation has reduced the time clinicians can devote to direct patient care, and ultimately limited the number of patients physicians can help. To mitigate this, many hospitals and clinics now deploy automatic speech recognition and note summarization tools during consultations. A natural next step in this pipeline is medical order extraction (e.g., medications, labs, imaging, follow-ups) from conversation transcripts to directly populate Electronic Health Records (EHRs).

While named entity recognition (NER) and relation extraction (RE) have been extensively studied in clinical NLP<sup>1</sup> (Xu et al., 2010; Doan and Xu, 2010; Yang et al., 2020; Fabacher et al., 2025; Henry et al., 2020; Lybarger et al., 2023), extracting actionable, structured orders from full-length consultations remains underexplored despite its potential impact. The task is challenging: inputs are

<sup>1</sup>Natural Language Processing.

### Doctor-Patient Consultation:

[doctor] so, for your first problem of your shortness of breath i think that you are in an acute heart failure exacerbation . i want to put you on some **lasix , 40 milligrams a day** .  
...  
[doctor] okay ? for your second problem of your type i diabetes , um , let's go ahead ... i wan na order a **hemoglobin a1c** for , um , uh , just in a , like a month or so , just to see if we have to make any adjustments ...  
[patient] sure .  
[doctor] for your fourth problem of your reflux , let's continue with omeprazole , 20 milligrams a day . do you have any questions , lawrence ?  
[patient] not at this point .  
...

### Medical orders:

```
"description": "lasix 40 milligrams a day",  
"order_type": "medication",  
"reason": "hortness of breath acute heart failure exacerbation",  
"provenance": [126, 127]
```

```
"description": "hemoglobin a1c",  
"order_type": "lab",  
"reason": "type i diabetes",  
"provenance": [138]
```

Figure 1: The medical order extraction task takes a doctor-patient dialog and extracts a JSON list of orders containing four keys (description, order\_type, reason, and provenance). Orders that were previously prescribed but not explicitly renewed should be excluded (e.g. omeprazole in this example).

long, dialogues contain interruptions as well as revisions, and outputs combine schema-constrained fields (e.g., order type) with free-text attributes (e.g., description, reason). These challenges are compounded by distributional shifts, as clinicians adapt their language to patients without medical training during consultations.

In this era of LLMs (Brown et al.; Achiam et al., 2023), new approaches have become feasible for the medical order extraction task — combining

Table 1: Final Ranking of MEDIQA-OE competition on the test set (100 samples) in which 6 teams participated. Our two baselines (excluded from the ranking) are a simple one-shot prompt for MediPhi-Instruct (1) and GPT-4o-0806 (2) with one example from the training set.

Rank	Team Name	Method	Match	Desc.	Reason	Type	Prov.	AVG
1	WangLab	GPT-4 constrained dec. Detailed instructions	<b>81.8</b>	<b>66.8</b>	29.5	<b>81.5</b>	<b>63.0</b>	<b>60.2</b>
2	silver_shaw	Gemini 2.5 Pro w/ thinking Detailed plan & instructions	76.4	64.1	<b>41.3</b>	74.7	60.4	60.1
3	MISo KeaneBeanz	Qwen3 32B Q4_K_M w/o thinking Instructions w/ 2 shots	73.4	58.0	35.6	71.6	48.4	53.4
4	EXL Health AI Lab	MedGemma 27B One shot (short format)	67.7	54.5	30.5	66.2	52.5	50.9
5	MasonNLP	Llama4 17B 16E Instruct One shot w/o orders	55.5	39.1	19.8	50.9	41.3	37.8
-	Baseline 2	GPT-4o Simple prompt w/ one shot	63.6	39.5	20.4	59.3	1.0	30.1
-	Baseline 1	MediPhi-Instruct 3.8B Simple prompt w/ one shot	43.3	25.8	19.5	39.6	13.8	24.7
6	HerTrials	Llama3.2 3.2B Instructions w/ one shot	31.2	19.6	9.0	29.6	5.6	15.9

long-context reasoning with schema-aware generation — yet limitations in context length, controllability, and calibration persist. The MEDIQA-OE shared task<sup>2</sup> investigates these challenges and benchmarks solutions to improve EHR clinical documentation, which we believe can reduce the burden on providers while ensuring the accurate capture of critical patient orders.

## 2 Previous Work

Tasks similar to order extraction in clinical NLP are commonly formulated as NER and RE. Early systems were rule-based (e.g., MedEx by Xu et al. (2010)) or used classical machine learning such as support vector machines (Doan and Xu, 2010). With pretrained contextual encoders, fine-tuned transformer models (e.g., BERT (Devlin et al., 2019), ClinicalBERT (Alsentzer et al., 2019)) became the standard for NER/RE and delivered consistent gains on clinical benchmarks (Yang et al., 2020; Fabacher et al., 2025).

More recently, LLMs enable span-free formulations that cast extraction as reading-comprehension style generation. Prompting methods (Peng et al., 2023; Cui et al., 2023; Peng et al., 2024) have shown strong results on several clinical information extraction tasks, including adverse drug events (Henry et al., 2020) and social determinants of

health (Lybarger et al., 2023). However, order extraction from full patient–doctor dialogues remains underexplored, particularly when models must (i) handle long, multi-speaker inputs and (ii) generate outputs that mix schema-constrained fields (e.g., order type, provenance) with free-text attributes (e.g., description, reason).

## 3 Methodology

### 3.1 Source Datasets

The long-form doctor-patient conversations used for the order-extraction task are primarily drawn from two datasets: ACI-Bench (Yim et al., 2023) and PriMock57 (Papadopoulos Korfiatis et al., 2022). The ACI-Bench corpus comprises 207 naturalistic conversations between physicians and patients, curated by domain experts to reflect real-world clinical interactions. Similarly, the PriMock57 dataset contains 57 mock doctor-patient dialogues, designed to simulate clinical scenarios in a controlled setting. Recent works such as Notechat (Wang et al., 2024) has introduced large-scale synthetic dialogue datasets. While this corpus is the largest, we excluded it due to the prevalence of low-quality dialogues we observed.

### 3.2 Annotations

We asked medically trained annotators to produce the gold-standard medical orders for the

<sup>2</sup><https://sites.google.com/view/mediqa-2025>

high-quality conversations of Primock57 and ACI-Bench. Annotation guidelines instructed to assess every medical order of type medication, imaging, lab, or follow-up within the conversation the way a doctor would create them in the EHR. This was intended to replicate doctors’ current process executed at the end of a patient encounter. We measured an inter-annotator agreement of 0.768. We sampled 100 examples containing 255 medical orders across both data sources as a test set and kept the others as training set (64 samples) used for few-shot prompting, and development set (100 samples) (Corbeil et al., 2025).

### 3.3 Evaluation

We evaluate model performance across four key metrics: description, reason, type, and provenance. Results are reported after performing a matching between reference and hypothesis orders based on description field’s word overlap<sup>3</sup>. An intermediary metric, the match score, is computed from this alignment process as the F1 between reference and predicted orders without looking at the content, thus specifically accounting for the amount of fabricated or omitted orders. It represents an upper bound for other metrics that are penalized for empty values for fair comparison. For description and reason metrics, we compute F1 scores of the rouge metric (Lin, 2004) over unigrams. Type is evaluated using accuracy due to its limited label space, and provenance is assessed via F1 score over provenance labels<sup>4</sup>. Finally, the leaderboard ranking is assessed via the average of all four key metrics: description, reason, type, and provenance.

## 4 Results

### 4.1 Leaderboard Ranking

We provided in Table 1 the final leaderboard of the MEDIQA-OE along participants’ approaches and our two baselines, which were used as reference points while being excluded from the ranking. All solutions are based on prompting language models. While there are two closed-source LLMs at the top of the ranking, the remaining submissions are leveraging open-weight LLMs in few-shot settings. WangLab obtained the 1<sup>st</sup> rank of the competition by prompting GPT-4 (Achiam et al., 2023) with JSON-constrained decoding and detailed instructions. Following closely by 0.1% on the average

<sup>3</sup>Necessary to compare orders with each other.

<sup>4</sup>Turn numbers where the order originates in the transcript.

score, silver\_shaw (Mehta, 2025) achieved the 2<sup>nd</sup> place by using Gemini 2.5 Pro (Comanici et al., 2025) in thinking mode. The other approaches (Balachandran et al., 2025; Karim and Özlem Uzuner, 2025) leveraged different open-weight models in few-shot settings such as Qwen3 32B (Qwen Team, 2025), MedGemma 27B (Sellergren et al., 2025), Llama4 Scout 17B (Meta AI, 2025) and Llama3.2 3.2B (Meta AI, 2024). Participants only appended one or two shot(s) examples due context limitations from long input-output pairs, and some even reduced examples into shorter formats. Overall, they also wrote simpler prompts compared to the two closed-weight LLM solutions.

#### 4.1.1 WangLab’s Approach

WangLab won the competition by prompting GPT-4 (Achiam et al., 2023) in a zero-shot setting. They obtain an average score of 60.2% with JSON-constrained decoding on the order format as well as using very detailed rules in the instructions. They achieved the highest match score at 81.8%, which indicates that a large proportion of reference orders are well matched. The average gains are double digits over the Baseline 1 based on GPT-4o with improvements on the provenance (+62.0%), description (+27.3%), type (+22.2%) and reason (+9.1%) scores.

Their prompt provides very detailed instructions, and is as follows:

1. Role attribution
2. Transcript definition with example
3. Task definition
4. Type definitions with rules and examples
5. Output JSON key definitions
6. Reason guidelines with specific examples
7. JSON output example
8. Overall guidelines
9. Eliciting JSON output

#### 4.1.2 silver\_shaw’s Approach

Following closely 0.1% on the average score, silver\_shaw (Mehta, 2025) achieved 2<sup>nd</sup> position with the highest reason score at 41.3% by prompting Gemini 2.5 Pro (Comanici et al., 2025) in thinking mode with a detailed reasoning plan and instructions.

Their one-call prompting approach asks the model to proceed in three steps aimed at mirroring the clinical reasoning processes: chain-of-thought analysis, self-critique & verification, and deterministic JSON generation.

#### 4.1.3 MISo KeaneBeanz’s Approach

MISo KeaneBeanz’s approach reached the 3<sup>rd</sup> rank by prompting the 4-bit quantized open-weight model Gwenz 32B (Qwen Team, 2025) in a two-shot setting.

#### 4.1.4 EXL Health AI Lab’s Approach

EXL Health AI Lab achieved the 4<sup>th</sup> rank at 50.9% leveraging a one-shot solution prompting MedGemma 27B (Sellergren et al., 2025), an open-weight medical LLM. Their experiments covered agentic workflows such as ReAct (Yao et al., 2023) and a four-step multi-agent pipeline. The one-shot method remained more accurate potentially because of the negative impact of noises introduced by multi-step approaches.

#### 4.1.5 MasonNLP’s Approach

The 5<sup>th</sup> rank of the MEDIQA-OE competition was attributed to MasonNLP (Karim and Özlem Uzuner, 2025). They used Llama4 17B (Meta AI, 2025) in a minimal one-shot prompting setting. The authors also reported an experiment with Llama4 8B.

#### 4.1.6 HerTrials’ Approach

HerTrials team ranked 6<sup>th</sup> with a one-shot prompting of the smallest open-weight language models Llama3.2 (Meta AI, 2024).

### 4.2 Analysis of Open-weight LLMs

We show the correlation between final accuracy and open-weight model sizes in Figure 2. Despite prompt variations, we computed a strong Pearson correlation of 0.981 between leaderboard ranking and model sizes, which is in line with previous work in clinical NLP (Dada et al., 2025).

## 5 Discussion and Limitations

In spite of top-ranking solutions achieving considerable scores with zero- and few-shot prompting and reasoning, significant gaps remain to push further the performance of the medical order-extraction task.

**First**, we notice that the maximum match F1 score is of 81.8%, which means that there is still room of nearly 20% to match the number of orders in the

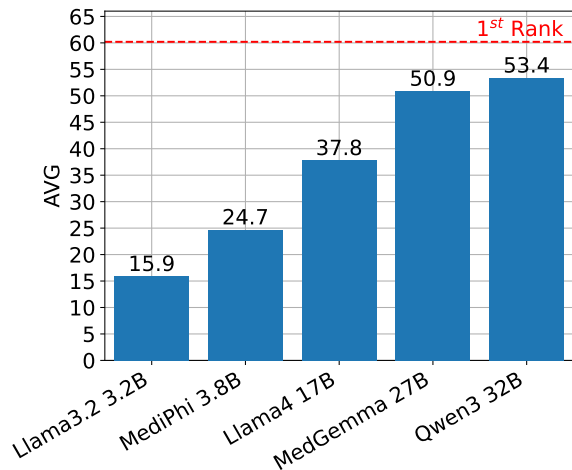


Figure 2: Open-weight models ranking obtained with few shots correlates with parameter count.

reference.

**Second**, description and provenance are both lagging behind the raw match score (i.e., their upperbound) by approximately 15-18%. Provenance was considerably improved in this challenge by using: larger models, constrained decoding, and specific instructions. Future work could explore embedding-based and hybrid systems.

**Third**, order types are all very close to match scores, which highlights how such classification tasks are well suited for LLMs.

**Fourth**, we observe low performances on the reason field which might come from the dispersion of reasons across the conversation and the fact that it is an optional field with scarcer annotations.

One of the main limitations of the current task is the small dataset sizes. The current trainset size of 64 samples limits the ability to use it for finetuning — which could particularly make open-weight small language models more competitive. Future work might produce larger datasets or leverage synthetic ones. While the inter-annotator agreement is considerably high, annotations might also present noises (e.g., span boundaries, non-expert conversational style instead of formal writing, etc.) which limit the maximum score below 100%.

## 6 Conclusion

To conclude, the medical order-extraction task was tackled by a variety of zero- and few-shot approaches using open- and closed-weight LLMs. Closed-weight models such as GPT-4 and Gemini 2.5 Pro in zero-shot setting dominated the top ranks, leveraging detailed instructions, constrained

decoding and reasoning. We observed a significant correlation of 0.981 between open-weight model sizes in few-shot settings and final accuracy. Although final scores considerably improved over the baselines especially in the match and provenance metrics, we still observe a significant gap in total extracted orders performance from the match score of 81.8% as well as in performances on the description and the reason free-form fields. We believe future works include synthetic data generation, model fine-tuning, hybrid systems, and focus on improving small language models.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Abhinand Balachandran, Bavana Durgapraveen, Gowsikkan Sikkam Sudhagar, VIDHYA VARSHANY J S, and Sriram Rajkumar. 2025. Exl health ai lab at medqa-oe 2025: Evaluating prompting strategies with medgemma for medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. Language models are few-shot learners.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel A. del Agua, Jérôme Tremblay, Akila Jeesson Daniel, Cari Bader, Yu-Cheng Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, François Beaulieu, and Paul Vozila. 2025. [Empowering health-care practitioners with language models: Structuring speech transcripts in two real-world clinical applications](#). *CoRR*, abs/2507.05517.
- Yang Cui, Lifeng Han, and Goran Nenadic. 2023. Medtem2. 0: Prompt-based temporal classification of treatment events from discharge summaries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 160–183.
- Amin Dada, Osman Alperen Koraş, Marie Bauer, Jean-Philippe Corbeil, Amanda Butler Contreras, Constantin Marc Seibold, Kaleb E Smith, Julian.friedrich@uk-essen.de Julian.friedrich@uk-essen.de, and Jens Kleesiek. 2025. [Does biomedical training lead to better medical performance?](#) In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM<sup>2</sup>)*, pages 46–59, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Son Doan and Hua Xu. 2010. Recognizing medication related entities in hospital discharge summaries using support vector machine. In *Proceedings of COLING. International conference on computational linguistics*, volume 2010, page 259.
- Thibaut Fabacher, Erik-André Sauleau, Emmanuelle Arcay, Bineta Faye, Maxime Alter, Archia Chahard, Nathan Miraillet, Adrien Coulet, and Aurélie Névéol. 2025. [Efficient extraction of medication information from clinical notes: an evaluation in two languages](#). *Preprint*, arXiv:2502.03257.
- Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.
- A H M Rezaul Karim and Özlem Uzuner. 2025. Masonlp at medqa-oe 2025: Assessing large language models for structured medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023. The 2022 n2c2/uw shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*, 30(8):1367–1378.
- Parth Mehta. 2025. silver\_shaw at medqa-oe 2025: A zero-shot prompting strategy with gemini for medical order extraction. In *Proceedings of the 7th Clinical*

- Natural Language Processing Workshop*. Association for Computational Linguistics.
- Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>. Blog post announcing Llama 3.2 models with lightweight text and vision capabilities, consulted on 2025-08-18.
- Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal AI innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>. Blog post on Meta AI's launch of Llama 4 models, consulted on 2025-08-18.
- Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. **Pri-Mock57: A dataset of primary care mock consultations**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.
- Cheng Peng, Xi Yang, Kaleb E Smith, Zehao Yu, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. **Model tuning or prompt tuning? a study of large language models for clinical concept and relation extraction**. *Journal of Biomedical Informatics*, 153:104630.
- Cheng Peng, Xi Yang, Zehao Yu, Jiang Bian, William R Hogan, and Yonghui Wu. 2023. Clinical concept and relation extraction using prompt-based machine reading comprehension. *Journal of the American Medical Informatics Association*, 30(9):1486–1493.
- Qwen Team. 2025. Qwen3: Think deeper, act faster. <https://qwenlm.github.io/blog/qwen3/>. Blog post, published by the Qwen Team, consulted on 2025-08-18.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. **NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201, Bangkok, Thailand. Association for Computational Linguistics.
- Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. **Medex: a medication information extraction system for clinical narratives**. *Journal of the American Medical Informatics Association*, 17(1):19–24.
- Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
- Wen-wai Yim, Yajuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. **Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation**. *Scientific data*, 10(1):586.



# Overview of the MEDIQA-WV 2025 Shared Task on Woundcare Visual Question Answering

**Wen-wai Yim**  
Microsoft Health AI  
yimwenwai@microsoft.com

**Asma Ben Abacha**  
Microsoft Health AI  
abenabacha@microsoft.com

**Meliha Yetisgen**  
University of Washington  
melihay@uw.edu

**Fei Xia**  
University of Washington  
fxia@uw.edu

## Abstract

Electronic messaging through patient portals facilitates remote care, connecting patients with doctors through asynchronous communication. While convenient, this new modality places an additional burden on physicians, requiring them to provide remote care as well as to see patients in clinic. Technology that can automatically draft responses for physician review is a promising way to improve clinical efficiency. Here, building on the 2024 MEDIQA Multilingual Multi-modal Medical Answer Generation (MEDIQA-M3G) challenge on dermatology, we present the 2025 MEDIQA Woundcare Visual Question Answering (MEDIQA-WV) shared task focusing on generating clinical responses to patient text and image queries. Three teams participated and submitted a total of fourteen systems. In this paper, we describe the task, datasets, as well as the participating systems and their results. We hope that this work can inspire future research on wound care visual question answering.

## 1 Introduction

Patient health portals enable asynchronous care, allowing patients to electronically submit clinical questions at any time. While this technology provides flexibility and convenience for both patients and clinicians, it also generates an unprecedented volume of additional work for care providers (Bishop et al.; Sinsky et al.).

In parallel, large multimodal general AI models have achieved state-of-the-art performance on benchmark tasks involving both classification and generation. However, despite these advances, such models often struggle with complex tasks requiring logical reasoning and multi-step inference. Medicine, in particular, demands integration of extensive medical knowledge, individual patient history, and current treatment protocols. Moreover, errors in this domain can have serious health consequences.

In the 2024 MEDIQA-M3G Challenge (Yim et al., 2024), we investigated the automatic generation of medical responses to patient queries in dermatology, incorporating both text and image components. In this new 2025 MEDIQA-WV challenge<sup>1</sup>, we extend our exploration to the domain of wound care.

## 2 Task & Dataset Description

Participants are given a patient query that includes a textual description and one or more associated images. The objective is to create free-text responses as if they were written by a clinician.

The training set includes a single gold response, the validation set includes two gold responses, and the test set includes three gold responses.

Both English and simplified Chinese versions are available for each instance. Table 1 provides an example of an instance with multiple gold responses. Table 2 and 3 provide summary characteristics of the dataset.

The patient queries were sourced from two open Chinese community forums (tieba.baidu.com and zhidao.baidu.com). The questions were translated into English by bilingual registered nurses and a medical translator. Each query was reviewed and answered by at least one of three U.S. medical domain experts: an emergency medicine physician, an emergency medicine resident, and a general surgeon. The created gold answers were then translated back into Chinese. In addition, metadata categories such as wound type and wound thickness were single-annotated and included. A detailed description of the dataset construction can be found in this work (Yim et al., 2025).

## 3 Evaluation Metrics & Baselines

Evaluation used three classical NLG metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin,

<sup>1</sup><https://sites.google.com/view/mediqa-2025>

PATIENT'S QUESTION & IMAGE

被猫隔着厚牛仔裤抓了一下，留下一道血痕，伤口如图，  
请问需要接种狂犬疫苗吗

I was scratched by a cat through thick jeans, leaving a blood mark  
as shown in the picture. Do I need to get a rabies vaccine?



GOLD RESPONSES

**Annotator 1:** 在这种情况下不需要注射狂犬疫苗。

No need for rabies vaccine in this situation.

**Annotator 2:** 如果猫的疫苗接种状态未知且无法隔离，虽然风险较低，具体取决于地区，建议接种疫苗。否则，局部伤口护理，使用局部抗生素和敷料是就行了。

If vaccination status of cat unknown and cannot be quarantined, while risk is low, depending on the area, would recommend vaccination. Otherwise, local wound care with topical antibiotic and dressing is sufficient

**Annotator 3:** 你应该不需要为猫抓伤接种狂犬症疫苗，因为从抓伤传播狂犬症的风险非常低。当然，如果你居住在一个狂犬症发病率很高的地区，我会另作考虑。虽然如此，我会继续观察伤口部位；如果伤口已受感染，你可能需要抗生素来治疗。我会用抗生素软膏包扎伤口。

You should not need to get a rabies vaccine for a cat scratch as the risk of transmission of rabies from a scratch is very low however, if you live in an area with high instance, I would consider it. However, I would continue to keep an eye on the area as you may need antibiotics for the cat scratch if it becomes infected. I would dress the wound with antibiotic ointment.

Table 1: Example where gold responses share similar content but differ in emphasis, level of explanation, and final recommendations (no vaccination versus vaccination), highlighting the importance of using multiple reference answers in evaluation to capture diverse opinions and perspectives.

Split	#Instances	#Responses	#Images
Training	279	279	449
Validation	105	210	147
Test	93	279	152

Table 2: Data Statistics

Split	EN		ZH	
	Query	Response	Query	Response
Training	46	29	52	43
Validation	44	41	50	60
Test	52	47	60	68

Table 3: Response Length Statistics (Mean Token Count). English tokens are per word, Chinese per character.

2004), BERTScore<sup>2</sup> (Zhang et al., 2020), and three LLM-as-judge variants (DeepSeek-V3, Gemini, GPT-4o). The exact configurations for the models and tokenizers are given in Table 4. For LLM-as-judge models, we used a consistent prompt for the same language. Both English and Chinese prompts for the LLM-as-judge methods are given in Table 5. Finally, we calculate the average score across all metrics.

For reference, we provide three baselines based on vision-language models with English and Chi-

<sup>2</sup>The mean is taken over all gold responses per instance

Configurations
<b>BERTSCORE</b>
github.com/Tiiiger/bert_score
tokenizer: "en"/"zh"
model: "microsoft/deberta-xlarge-mnli" for English
"zh" for Chinese
<b>BLEU</b>
github.com/mjpost/sacrebleu
use_effective_order: true
EN tokenizer: tokenize_13a
ZH tokenizer: tokenize_zh
<b>DEEPSEEK</b>
ai.azure.com
AZURE AI Foundry model ID:
DeepSeek-V3-0324
content-filter: None
<b>Gemini</b>
https://cloud.google.com/ai/generative-ai?hl=en
Google GenAI Model Name:
gemini-1.5-pro-002
hate-speech: Block-None
harrasment: Block-None
<b>GPT-4o</b>
oai.azure.com
AZURE AI Foundry OpenAI model name:
gpt-4o
content-filter: None
<b>ROUGE</b>
huggingface.co/spaces/evaluate-metric/rouge
tokenizer: same as BLEU

Table 4: Evaluation Metric Configurations. Defaults are used if not otherwise mentioned.

<p>SYSTEM: You are a helpful medical assistant.</p> <p>USER: Given a patient {QUERY}, and a list of {REFERENCE RESPONSES}, please evaluate a {CANDIDATE RESPONSE} using a three-step rating described below.</p> <p>Rating: 0 - {CANDIDATE RESPONSE} is incomplete and may contain medically incorrect advice.</p> <p>Rating: 0.5 - {CANDIDATE RESPONSE} is incomplete but has partially correct medical advice.</p> <p>Rating: 1.0 - {CANDIDATE RESPONSE} is complete and has medically correct advice.</p> <p>The {REFERENCE RESPONSES} represent answers given by domain experts and can be used as references for evaluation.</p> <p>QUERY:</p> <p>REFERENCE RESPONSES:</p> <p>CANDIDATE RESPONSE:</p> <p>RATING:</p>
<p>SYSTEM: 你是一个很有帮助的医疗助手。</p> <p>USER: 给定病人提出的{问题}以及一系列{参考回复}, 请使用下述的3级评分制度来评估{待测回复}。</p> <p>评分: 0 - {待测回复} 不完整且与所有{参考回复}事实不符。</p> <p>评分: 0.5 - {待测回复} 不完整但与至少一个{参考回复}事实相符。</p> <p>评分: 1.0 - {待测回复} 完整且与至少一个{参考回复}事实相符</p> <p>问题: {}</p> <p>参考回复: {}</p> <p>待测回复: {}</p> <p>评分:</p>

Table 5: LLM-as-Judge Prompts for General AI Models (English TOP, Chinese BOTTOM)

nese proficiency: Baseline 1: Gemini-1.5-pro-002, Baseline 2: GPT-4o, Baseline 3: Qwen-VL<sup>3</sup>.

The prompts for generating the English and Chinese baseline answers were:

- **English:** “Please answer as a professional medical doctor, answer limited to 41 words<sup>4</sup>. {QUERY\_TITLE}: {QUERY\_CONTENT}”
- **Chinese:** “请以专业医生的身份提供建议, 答案只限60字 {QUERY\_TITLE}: {QUERY\_CONTENT}”

## 4 Official Results

Three teams participated in this shared task, with a total of fourteen submissions<sup>5</sup>. The teams were EXL Health AI Lab (India), DermaVQA (United Kingdom), and MasonNLP (United States). Tables 6 and 7 present the official results on the English and Chinese datasets, respectively. The overall

baseline averages were higher for the Chinese subset than for the English subset. Notably, only three of the fourteen submissions included systems capable of handling the Chinese subset. Two teams submitted working notes, and we provide a description of their systems below.

**MasonNLP** (Karim and Özlem Uzuner, 2025). This team experimented with zero-shot, few-shot, and retrieval-augmented few-shot learning approaches to identify the most relevant training data. The backbone model used was meta-llama/llama-4-scout-17b-16e-instruct. The retrieval-augmented generation (RAG) component was built using semantic text embeddings from sentence-transformers/all-MiniLM-L6-v2 and vision-language embeddings from CLIP (openai/clip-vit-base-patch32). The two most similar training examples were retrieved. In experimentation, the RAG system with both image and text search achieved the highest performance.

**EXL Health AI Lab** (Durgapaveen et al., 2025). This team experimented with two approaches: (1) a two-step generation approach, which first classified relevant metadata (e.g., wound

<sup>3</sup><https://huggingface.co/Qwen/Qwen-VL>

<sup>4</sup>Length suggestions used the average response lengths in the validation sets.

<sup>5</sup>A single submission may include responses in both English and Chinese.

Team	BLEU	ROUGE-L	BERTScore	DeepSeek-V3	Gemini	GPT-4o	AVG
EXL Health AI Lab	0.099	0.456	0.622	0.682	0.645	0.715	0.473
EXL Health AI Lab	0.130	0.452	0.619	0.635	0.591	0.629	0.457
EXL Health AI Lab	0.130	0.452	0.619	0.625	0.586	0.618	0.455
EXL Health AI Lab	0.057	0.456	0.623	0.607	0.629	0.667	0.45
EXL Health AI Lab	0.057	0.455	0.623	0.591	0.634	0.624	0.44
EXL Health AI Lab	0.037	0.441	0.611	0.604	0.570	0.618	0.427
MasonNLP	0.089	0.422	0.59	0.535	0.554	0.554	0.414
MasonNLP	0.073	0.433	0.604	0.589	0.565	0.532	0.411
EXL Health AI Lab	0.064	0.448	0.621	0.512	0.500	0.505	0.410
EXL Health AI Lab	0.064	0.448	0.621	0.499	0.505	0.505	0.410
DermaVQA	0.076	0.455	0.606	0.427	0.457	0.371	0.377
MasonNLP	0.047	0.235	0.325	0.321	0.301	0.339	0.236
MasonNLP	0.017	0.140	0.192	0.210	0.188	0.215	0.141
Baseline 1: Gemini-1.5-pro-002	0.064	0.449	0.621	0.791	0.817	0.683	0.571
Baseline 2: GPT-4o	0.062	0.450	0.623	0.756	0.731	0.688	0.552
Baseline 3: Qwen-VL	0.051	0.428	0.599	0.513	0.478	0.473	0.424

Table 6: Results - English. Evaluation metrics DeepSeek-V3, Gemini, and GPT-4o are reported under an LLM-as-judge setup.

Team	BLEU	ROUGE-L	BERTScore	DeepSeek-V3	Gemini	GPT-4o	AVG
DermaVQA	0.102	0.489	0.656	0.570	0.548	0.511	0.439
MasonNLP	0.000	0.006	0.008	0.011	0.011	0.011	0.006
MasonNLP	0.000	0.005	0.007	0.011	0.011	0.011	0.006
Baseline 1: Gemini-1.5-pro-002	0.118	0.501	0.661	0.941	0.957	0.898	0.679
Baseline 2: GPT-4o	0.123	0.496	0.666	0.844	0.860	0.753	0.624
Baseline 3: Qwen-VL	0.094	0.484	0.658	0.763	0.694	0.699	0.565

Table 7: Results - Chinese. Evaluation metrics DeepSeek-V3, Gemini, and GPT-4o are reported under an LLM-as-judge setup.

type and infection status) and then incorporated the metadata into the final generation step; and (2) a few-shot prompting strategy. In Approach (1), the group tested MedGemma (27B Multimodal) for both steps, with confidence thresholds also provided alongside the metadata classifications. In Approach (2), an all-mpnet-base-v2 sentence transformer was used to encode training instances for semantic similarity search. The group experimented with retrieving between 5 and 25 samples as few-shot examples, and tested both the InternVL3-38B and MedGemma-27B models.

Unlike the participating teams, the baselines used a zero-shot approach. Both Gemini and GPT-4o, very large general-purpose models (hundreds of billions of parameters), achieved state-of-the-art performance on both English and Chinese. Meanwhile, the smaller Qwen-VL model (7B) showed performance comparable to the RAG approaches on the English subset, despite no additional prompting. For the Chinese dataset, Qwen-VL also produced competitive results relative to the system submissions.

## 5 Discussion

Results show that out-of-the-box performance from very large general multimodal models such as Gemini and GPT-4o is highly competitive. That said, the smaller Qwen-VL (7B) model achieved results comparable to the Llama-17B and MedGemma-27B system models that used dedicated RAG few-shot examples in both languages. This suggests that while larger models currently hold a strong advantage, there remains considerable room for optimization and specialization, particularly in identifying models that achieve the best performance-cost trade-off.

In this work, we report both classical NLG evaluation metrics and new LLM-as-judge metrics. The variation in the magnitude of these metrics, and the resulting differences in rankings, suggests that a detailed, comprehensive study against human evaluation is necessary.

Compared to the 2024 MEDIQA-M3G task (Yim et al., 2024) on dermatology, both baseline and system performances were lower. For example, the GPT-4 baseline had a BLEU score of 0.813 and 0.867 BERTScore for English. This may be attributable to differences in response lengths. The dermatology dataset had shorter responses (average

12 words in English, 16 in Chinese for the training set) mostly related to diagnosis, compared to the longer responses in this task (29 English, 43 Chinese) which include care instructions.

## 6 Conclusion

In this shared task, we found that a range of large multimodal language models, both with and without few-shot examples and RAG augmentations, demonstrated varying levels of performance. While models with larger parameter counts generally held an advantage, performance rankings among models of comparable size were less predictable.

With the highest average score reaching only 0.679, it is clear that further progress is required. Moreover, the variation across evaluation metrics highlights the need for significant advancements in evaluation methodology to reduce the uncertainty associated with individual metrics.

This shared task focused solely on free-text response generation; however, incorporating metadata such as wound type would enable richer multimodal studies in the future. In particular, systems that jointly optimize classification and generation in a fine-tuned setting may yield notable performance gains.

We hope that this work will inspire further research in multimodal patient question answering and medical open-response evaluation, as well as encourage exploration of such applications in clinical practice.

## References

- Tara F. Bishop, Matthew J. Press, Jayme L. Mendelsohn, and Lawrence P. Casalino. Electronic communication improves access, but barriers to its widespread adoption remain. 32(8):10.1377/hlthaff.2012.1151.
- Bavana Durgapraveen, Sornaraj Sivasankaran, Abhinand Balachandran, and Sriram Rajkumar. 2025. Exl health ai lab at mediqua-wv 2025: Mined prompting and metadata-guided generation for wound care visual question answering. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.
- A H M Rezaul Karim and Özlem Uzuner. 2025. Ma-sonnlp at mediqua-wv 2025: Multimodal retrieval-augmented generation with large language models for medical vqa. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.
- Christine A. Sinsky, Tait D. Shanafelt, and Jonathan A. Ripp. The electronic health record inbox: Recommendations for relief. 37(15):4002–4003.
- Wen-wai Yim, Asma Ben Abacha, Robert Doerning, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan Yu, Fei Xia, M Kennedy Hall, and Meliha Yetisgen. 2025. [Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care](#). *Journal of Biomedical Informatics*.
- Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, and Martin Krallinger. 2024. [Overview of the MEDIQA-M3G 2024 shared task on multilingual multimodal medical answer generation](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 581–589, Mexico City, Mexico. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

# Team NLP4Health at ChemoTimelines 2025: Finetuning Large Language Models for Temporal Relation Extractions from Clinical Notes

**Zhe Zhao and V.G.Vinod Vydiswaran**  
University of Michigan, Ann Arbor, MI, USA  
{zzhaozhe, vgvinodv}@umich.edu

## Abstract

Extracting chemotherapy timelines from clinical narratives is a challenging task, but critical for cancer research and practice. In this paper, we present our approach and the research investigation we conducted to participate in Sub-task 1 of the ChemoTimelines 2025 shared task on predicting temporal relations between pre-identified events and time expressions. We evaluated multiple fine-tuned large language models for the task. We used supervised fine-tuning strategies for Llama3-8B model to classify temporal relations. Further, we set up zero-shot prompting for Qwen3-14B to normalize time expressions. We also pre-trained and fine-tuned a Llama3-3B model using unlabeled notes and achieved results comparable with the fine-tuned Llama3-8B model. Our results demonstrate the effectiveness of fine-tuning and continual pre-training strategies in adapting large language models to domain-specific tasks.

## 1 Introduction

Understanding patient chemotherapy timelines is crucial to making clinical decisions about cancer care. However, most of the temporal information about treatment plans is contained in extensive clinical narratives and is currently only accessible through manual chart review, which is time-consuming and labor-intensive. New methods are needed to automatically extract temporal relations to fully utilize the utility of clinical notes.

Recently, Large Language Models (LLMs), pre-trained on large amounts of unstructured data and instruction-tuned to follow human instructions better, have achieved promising performance in information extraction, text generation, and classification tasks. In this paper, we present our approach to adapt open-source generic LLMs to oncology. We develop new approaches to automatically extract temporal relationships between pairs of chemotherapy events and time expressions within patients'

clinical notes. We reformulated the relation extraction task into a text generation task and used Supervised Fine Tuning (SFT) techniques to instruct the model to generate the relation-type labels.

The main contributions of this paper are:

1. Introduce a novel SFT approach to adapt general LLMs to address temporal relation extraction in clinical narratives.
2. Implement an end-to-end system to use (i) a small-scale Llama3 LLM for temporal relation classification, and (ii) a medium-scale Qwen3 model for time normalization tasks.
3. Experimentally evaluate whether continual pretraining could help smaller LLMs achieve performance comparable to larger LLMs after fine-tuning.

## 2 Related Work

Temporal relation extraction is essential in the clinical domain to understand disease progression, diagnose health conditions, and evaluate treatment effectiveness (Zhou and Hripcsak, 2007). In prior work, researchers such as Tang et al. (2013), Cherry et al. (2013) and Sohn et al. (2013) have proposed hand-crafted features and conventional machine learning algorithms for this task. With the emergence of pre-trained language models, researchers benefited from the generic representational power of Transformer-based models, including BERT (Zhou et al., 2021), RoBERTa (Tan et al., 2024), and BART (Wright-Bettner et al., 2020; Yan et al., 2021), to improve the performance of temporal relation extraction. Lin et al., 2021 proposed EntityBERT, which was obtained by continually pre-training PubMedBERT on a clinical corpus. These models leveraged contextualized embeddings and domain adaptation techniques like fine-tuning and continual pretraining, and have shown improvement over conventional machine learning models.

With recent advances in Large Language Models (LLMs), instruction tuning has been crucial to improve zero-shot learning capabilities and to better follow human instructions to perform specific tasks (Ouyang et al., 2022; Chung et al., 2022). These abilities can be further enhanced by Supervised Fine Tuning (SFT), a technique of adapting an LLM that is pretrained on a general domain to perform a specific task. SFT allows the model to be fine-tuned in a supervised setting, where it learns the patterns among the instruction-response examples of training data. This technique has been commonly used to adapt LLMs for text classification, entity recognition, and question answering tasks.

In addition to these fine-tuning techniques, studies find that continual pretraining for LLMs over domain corpus can improve the end-to-end within-domain performance (Ke et al., 2023). Further, (Xie et al., 2024) showed that pretraining LLMs on task-specific corpus is more efficient in improving the end-task performance. However, these techniques have rarely been studied for temporal relation extraction in the clinical domain.

### 3 Methodology

In this section, we describe our proposed approach using LLMs for temporal relation classification, post-processing, time normalization, and chemotherapy timeline construction for each patient. Figure 1 illustrates our submitted systems.

#### 3.1 Dataset and Task description

We participated in ChemoTimelines 2025 Subtask 1, which aims to classify temporal relations between pre-identified chemotherapy events (Yao et al., 2025). There are three types of relations to classify – BEGINS-ON, ENDS-ON, and CONTAINS. In addition to the relation classification, this subtask requires the participants to normalize the time expressions into ISO standard format and resolve any duplicates or conflicts among events when organizing them into patients’ timelines. The dataset is provided by University of Pittsburgh/UMPC, and consists of de-identified notes from the electronic health records (EHRs) of breast cancer, melanoma, and ovarian cancer patients. Details about the subtasks, data distribution, and evaluation methodology are described in Yao et al., 2025.

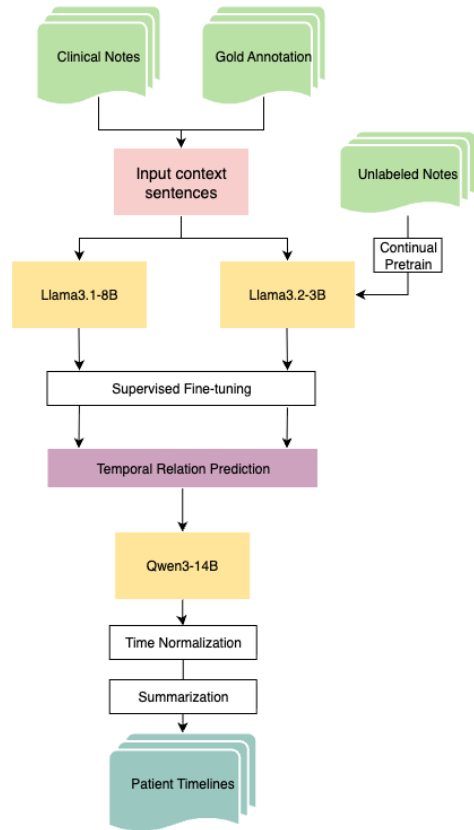


Figure 1: Overview of the submitted systems.

#### 3.2 Text Pre-processing

With the gold annotation of chemotherapy events (EVENTs) and time expressions (TIMEX3s) provided, the goal of our system is to predict the temporal relations between them and then produce patient-level timelines.

As a first step, we curated a set of instruction-response examples from the training set for the SFT. We tagged all EVENTS and TIMEX3 entities in the notes using the markers `<event>` `</event>` and `<time>` `</time>`, respectively. We used the medspaCy library (Eyre et al., 2021) to segment clinical notes into sentences. Then, we constructed instruction contexts by including the sentences that contain the EVENT and TIMEX3 entities, and all sentences between them. We constructed the contexts this way for every combination of the EVENT and TIMEX3 provided. This process created negative examples when the tagged EVENT and TIMEX3 entities were not related, and we used the label NO-REL to indicate absence of a relation. To limit the number of NO-REL examples and avoid working with a heavily imbalanced dataset, we excluded instances where the distance between the EVENT and TIMEX3 entities exceeds 250 to-

Label	# Train	# Dev
<b>Breast Cancer</b>		
BEGINS-ON	131	27
CONTAINS	298	57
ENDS-ON	26	29
NO-REL (pre-threshold)	2320	710
NO-REL (post-threshold)	389	133
<b>Melanoma</b>		
BEGINS-ON	10	42
CONTAINS	37	157
ENDS-ON	1	2
NO-REL (pre-threshold)	293	1138
NO-REL (post-threshold)	35	192
<b>Ovarian Cancer</b>		
BEGINS-ON	100	34
CONTAINS	326	140
ENDS-ON	65	52
NO-REL (pre-threshold)	1536	1363
NO-REL (post-threshold)	346	226

Table 1: Number of relation type labels in training and development sets.

kens, as we observed that the maximum distance between the entities in the positive examples in the training set was 213 tokens. As Table 1 shows, this simple threshold reduced the number of negative examples by 81–88% across the three cancer types.

### 3.3 LLM Fine-Tuning

Inspired by Haddadan et al., 2024, we reformulated Subtask 1 as a text generation task and fine-tuned LLMs in an SFT fashion. Appendix A describes our fine-tuning approach, the instruction we used, and the expected result. In the instruction prompt, we included the definitions of each temporal relation type, provided by Yao et al., 2025, and instructed the model to only focus on the tagged entity pair. We appended the preprocessed context after the instruction and instructed the model to only output one of the predefined temporal relation labels: BEGINS-ON, ENDS-ON, CONTAINS, or NO-REL without additional texts and reasoning. We trained four instruction-tuned LLMs for our experiments – Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct (Grattafiori et al., 2024), Qwen3-8B (Yang et al., 2025), and Ministral-8B-Instruct-2410 (Mistral AI Team, 2024).

In preliminary experiments, we found that LLMs often fabricated the output on the development set, even after we fine-tuned them. For example, mod-

els would generate labels that were not within the predefined label set, provide more than one label, or generate additional text. However, we observed that letting LLMs generate JSON-formatted output reduced such fabrication. In addition, assigning numbers to relation labels also helped alleviate the issue. Subsequently, we added additional instructions after the input context and asked the model to strictly follow the specified JSON format.

For our experiments, we used Huggingface’s SFTTrainer to fine-tune the models. The learning rate was set to  $2e-5$  with a cosine scheduler and the weight decay of 0.001. We set the maximum sequence length to 512 and fine-tuned all models for 5 epochs. To fit the experiments in a single NVIDIA A100 GPU, we loaded and fine-tuned all models in the BFloat16 format implemented by PyTorch and used paged-AdmW optimizer with 8-bit quantization.

### 3.4 Continual Pretraining Small-Scale LLMs

Following (Xie et al., 2024), we investigated whether continually pretrained on an unlabeled corpus could improve the performance of LLMs on temporal relation extraction in the clinical domain. Further, we also wanted to study whether pretraining using a fraction of unlabeled corpus could help smaller-scale LLMs achieve performance comparable to large-scale LLMs after fine-tuning.

Due to a limited GPU memory availability, we investigated this question only with the Llama-3.2-3B-Instruct model. We pre-trained the model using 500 unlabeled notes sampled across all cancer types. Similar to the previous experiments, we loaded and trained the model using BFloat16 format and a Paged AdmW optimizer in 8-bit quantization. We directly pre-trained the Llama-3.2-3B model using full notes for 3 epochs.

### 3.5 Time Normalization and Timeline Summarizations

Once temporal relations were classified, we fed the input context into a new Qwen3-14B LLM, and normalized the TIMEX3 entities using zero-shot prompting. Specifically, we instructed the model to only normalize the tagged TIMEX3 entity into one of the ISO-8601 standard formats – YYYY-MM-DD and YYYY-Www, as specified in the shared task description. We then used Chain-of-Thoughts prompts to instruct the model to normalize time entities based on the information available in the input context. This helps the model decide when to nor-



malize time entities using the YYYY-Www format, for example when the time entity is related to week numbers, such as “in 4 weeks”. In addition, when the document time (DOCTIME) was available, we asked the model to take it into consideration. We instructed the model to generate a JSON-formatted output to reduce fabrication. Appendix B includes additional details on the instruction we used for time normalization.

After obtaining normalized  $[EVENT, relation, TIMEX3]$  tuples, we ordered them to generate timelines for each patient, according to the rules specified in the shared task. We removed duplicates and for tuples with same EVENT and TIMEX3 entities, we only kept the ones with more specific types, viz., BEGINS-ON and ENDS-ON. Finally, we removed generic mentions of chemotherapy (e.g., words like “chemo” and “chemotherapy”) if a more specific EVENT is included with the same relation type and TIMEX3 entity.

### 3.6 Evaluation Metrics

The extracted timelines are evaluated by comparing the predicted  $[EVENT, relation, TIMEX3]$  tuples against the gold timelines tuples for each patient. We used the evaluation code provided by Yao et al., 2025 to evaluate our approaches in development and validation sets. The overall F1 score is calculated by averaging Type A and Type B metrics. Type A F1 score includes all patients regardless of the presence of chemotherapy timelines and Type B F1 scores include only patients with effective chemotherapy timelines. The official metric for the shared task adopts strict evaluation, where the true positive means that all three elements of a predicted tuple must match the corresponding gold-label tuple for a patient to be considered correct.

## 4 Results

We submitted three runs on the test set. The first submission uses the fine-tuned Llama3.1-8B model for temporal relation classification and the Qwen3-14B model for time normalization. When we evaluated this system on the development set, we found that the model often misidentified the ENDS-ON type with CONTAINS if the TIMEX3 entity is preceded with a linking verb like “be” or “was” and the preposition “on”. So, in the second submission, we applied a regular expression to match this pattern in the input context and changed the predicted CONTAINS label. The third submission uses the

fine-tuned Llama3.2-3B model with continual pre-training on the unlabeled notes for the temporal relation classification and Qwen3-14B model for time normalization. We also applied the regex pattern to fix the potentially misidentified ENDS-ON labels.

Table 2 shows the results of our systems on both development and test sets. For the test set, submission 2 achieved the highest average F1 score of 59.66 – 55.98 for the breast cancer patients, 65.99 for melanoma patients, and 57.02 for the ovarian cancer patients. We should note that while submission 3 achieved the highest scores for breast cancer (57.34) and ovarian cancer (59.61), the average score for submission 3 was the lowest of our submissions due to its low F1 for melanoma. These trends were similar to our results on the development set; submission 2 was the best-performing system on the development set with an average F1 score of 83.69 – 82.07 for breast cancer, 82.34 for melanoma, and 86.66 for ovarian cancer. However, on the development set, Llama3.2-3B based submission 3 achieved similar performance to the Llama3.1-8B model based submissions for breast cancer and melanoma and performed much worse on ovarian cancer, which is different from what happened on the test set.

## 5 Error Analysis

Due to the unavailability of the gold timelines for the test set, we will provide the error analysis based on the results on the development set. Two main sources of errors are the ENDS-ON cases and time normalization. Out of 83 ENDS-ON cases in the development set, our best fine-tuned model, Llama3.1-8B, makes 43 wrong predictions with 42 of them being misidentified as CONTAINS. We notice that our model is prone to make this wrong prediction when there is an “on” preposition preceded the TIMEX3 entity. For example, “Patient has completed 4 cycles of  $\langle event \rangle$  Adriamycin  $\langle /event \rangle$  and Cytoxan, with the last dose being on  $\langle time \rangle$  7/27/13  $\langle /time \rangle$ .” This error did not appear in cases with other prepositions such as “through” or “in”; for example, “Carboplatin, cytoxan, and  $\langle event \rangle$  Avastin  $\langle /event \rangle$  x8 cycles through  $\langle time \rangle$  May 2013  $\langle /time \rangle$ .” Subsequently in submission 2, we used regexes to fix these errors by looking for more specific text patterns such as “being on” and “was on” in order to reduce false positives. As depicted in Table 2, the

	Average	Breast Cancer	Melanoma	Ovarian Cancer
<b>Development set</b>				
1. Llama3.1-8B + Qwen3-14B	0.83	<b>0.84</b>	<b>0.83</b>	0.82
2. Llama3.1-8B + Regex + Qwen3-14B	<b>0.84</b>	0.82	0.82	<b>0.87</b>
3. Llama3.2-3B + Regex + Qwen3-14B	0.80	0.83	<b>0.83</b>	0.73
<b>Test set</b>				
1. Llama3.1-8B + Qwen3-14B	<b>0.60</b>	0.56	<b>0.66</b>	0.57
2. Llama3.1-8B + Regex + Qwen3-14B	<b>0.60</b>	0.56	<b>0.66</b>	0.57
3. Llama3.2-3B + Regex + Qwen3-14B	0.59	<b>0.57</b>	0.60	<b>0.60</b>

Table 2: Patient timelines evaluation on development and test sets across all cancer types. Bold scores indicate the highest score for each cancer type

regexes improved submission 2 by increasing the accuracy for ovarian patients and thus the average score in the development set. However, it did not improve the results in the test set. We postulate this behavior may be due to overfitting as we notice that the ENDS-ON cases are more likely to have “through” or “in” as prepositions of TIMEX3 entities in the training set.

For time normalization, the majority of errors in the development set came from misrepresenting YYYY-MM-DD format by YYYY-Www format. For example, the original text of the tuple [*“chemotherapy”, “contains-1”, “2013-02-13”*] for patient 35 is *“Return in <time>3 weeks</time> for <event>chemotherapy</event> will commence with IV/IP PGH”*. The Qwen-14B model wrongly normalized the time “3 weeks” to “2013-W07” since we instruct the model to normalize time into YYYY-Www format if the information of day is unavailable and the term is related to week in our chain-of-thought prompt. This error could be caused by the limited input context, as we only included sentences that were between the EVENT and TIMEX3 entities. Additional relevant information for normalizing dates may be contained in a larger context window.

When running our systems on the test set, although in rare cases, it is worth noting that the fine-tuned Llama3.1-8B model fabricates output by not following the JSON format or generating invalid labels. On the other hand, the continually-pretrained and fine-tuned Llama3.2-3B model did not fabricate any output. This suggests that continually pretraining language models on a task-related unlabeled corpus can improve domain adaptation and stability of large language models.

## 6 Conclusions

In this paper, we present our effort in participating in the ChemoTimelines 2025 Shared task 1. We leverage the general domain, instruction-tuned LLMs and fine-tune them in a supervised fashion to extract chemotherapy timelines from clinical notes. The results show that Llama3.1-8B + Qwen3-14B system, with a regex-based correction was the best model and achieved second place among teams for Subtask 1. Although our continually pre-trained Llama3.2-3B model received the lowest rank among our three systems, it performs the best on breast and ovarian cancer notes and does not fabricate the output when conducting inference on the new data. Our results show that fine-tuning still remains an important tool to enhance the capabilities of LLMs in more specific domains, and continual pretraining can further improve the effects of fine-tuning, helping small-scale LLMs to achieve comparable performance to larger scale LLMs. Future work to improve our system may include using techniques like early-stopping to prevent overfitting and tackling the low frequency labels like ENDS-ON through data augmentation.

## 7 Limitations

There are multiple limitations to this work because of factors related to experimental set up. First, due to the limited computational resources, we could only fine-tune and pretrain our models using 8-bit optimizers. This prevented us from training models using full precision, which could result in better relation type classification. Further, we ran pre-training only using a small set of unlabeled notes, very few of which were related to Melanoma, resulting in the under-performance on the Melanoma patients in the test set.

Second, the implementation of SFTTrainer in the Transformer library does not support using customized metrics for evaluation. This made it difficult to track the fine-tuning process and implement early-stopping.

Finally, due to the time limitations, we did not conduct hyperparameter tuning. Optimal hyperparameters may increase the stability and performance of LLMs, reducing fabrications, and improving timeline extraction.

## 8 Ethics Statement

The pre-training and post-training in this study were conducted in a secured computing environment provided by University of Michigan Health Information and Technology Services, which includes the safeguards required by HIPAA. All the data used in shared tasks was de-identified by the ChemoTimelines 2025 organizers. The access of the data was executed by a data user agreement with University of Pittsburgh, and was regulated by Institutional Review Boards of the University of Michigan Medical School, to ensure the continual adherence to ethical guidelines.

## References

- Colin Cherry, Xiaodan Zhu, Joel Martin, and Berry de Bruijn. 2013. *À la Recherche du Temps Perdu: extracting temporal relations from medical text in the 2012 i2b2 NLP challenge*. *Journal of the American Medical Informatics Association*, 20(5):843–848.
- 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*. *arXiv preprint*. ArXiv:2210.11416 [cs].
- Hannah Eyre, Alec B. Chapman, Kelly S. Peterson, Jianlin Shi, Patrick R. Alba, Makoto M. Jones, Tamara L. Box, Scott L. DuVall, and Olga V. Patterson. 2021. *Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python*. *arXiv preprint*. ArXiv:2106.07799 [cs].
- 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*. *arXiv preprint*. ArXiv:2407.21783 [cs].
- Shohreh Haddadan, Tuan-Dung Le, Thanh Duong, and Thanh 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 Natural Language Processing Workshop*, pages 382–393, Mexico City, Mexico. Association for Computational Linguistics.
- Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. *Continual Pre-training of Language Models*. *arXiv preprint*. ArXiv:2302.03241 [cs].
- 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.
- Mistral AI Team. 2024. *Un Ministral, des Ministraux*. <https://mistral.ai/news/ministraux>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, P. Christiano, Jan Leike, and Ryan J. Lowe. 2022. *Training language models to follow instructions with human feedback*. *ArXiv*.
- Sunghwan Sohn, Kavishwar B Waghlikar, Dingcheng Li, Siddhartha R Jonnalagadda, Cui Tao, Ravikumar Komandur Elayavilli, and Hongfang Liu. 2013. *Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification*. *Journal of the American Medical Informatics Association*, 20(5):836–842.
- Xingwei Tan, Gabriele Pergola, and Yulan He. 2024. *Extracting Event Temporal Relations via Hyperbolic Geometry*. *arXiv preprint*. ArXiv:2109.05527 [cs].
- Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. 2013. *A hybrid system for temporal information extraction from clinical text*. *Journal of the American Medical Informatics Association*, 20(5):828–835.
- 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.
- Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. *Efficient Continual Pre-training for Building Domain Specific Large Language Models*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10184–10201, Bangkok, Thailand. Association for Computational Linguistics.

- Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. [A Unified Generative Framework for Various NER Subtasks](#). *arXiv preprint*. ArXiv:2106.01223 [cs].
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 Technical Report](#). *arXiv preprint*. ArXiv:2505.09388 [cs].
- Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. 2025. Overview of the 2025 Shared Task on Chemotherapy Treatment Timeline Extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*.
- Li Zhou and George Hripcsak. 2007. [Temporal reasoning with medical data—A review with emphasis on medical natural language processing](#). *Journal of Biomedical Informatics*, 40(2):183–202.
- Yichao Zhou, Yu Yan, Rujun Han, J. Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021. [Clinical Temporal Relation Extraction with Probabilistic Soft Logic Regularization and Global Inference](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14647–14655.

## A Supervised Fine-Tuning Instruction

We use the following instruction to conduct supervised fine-tuning on Llama 3 models:

“You will assist a clinician to classify temporal relations from sentences extracted from the patient notes. Temporal relations refers to a set of timeline relations between EVENTS and TIMES. The relation can be BEGINS-ON, CONTAINS, ENDS-ON and NO-REL. BEGINS-ON signals that the EVENT begins on the TIME it’s related to. ENDS-ON signals that the EVENT ends on the TIME it’s related to. CONTAINS signals that the EVENT is completely contained within the temporal span of the TIME it’s related to. NO-REL signals that the EVENT does not relate to the TIME presented. Here is an input context describing the relation between an EVENT and TIME. The markers <event> and </event> surrounds the EVENT entity. The markers <time> and </time> surrounds the TIME entity. Given the input text, extract the relation between the EVENT and TIME entities.

Here are sentences extracted from a patient note for you to extract temporal relations: sentences Note: Your answer must only be the relation between the two given entities and must follow this exact JSON format: “Answer”: [INSERT 1 for BEGINS-ON, 2 for CONTAINS, 3 for ENDS-ON and 99 for NO-REL]. YOUR RESPONSE MUST BE IN THIS EXACT JSON FORMAT. YOU MUST CHOOSE ONLY ONE RELATION. DO NOT OUTPUT NUMBERS THAT ARE NOT 1,2,3 AND 99. DO NOT OFFER EXPLANATIONS OR ANY ADDITIONAL TEXT. Make sure your output follows the json format strictly.”

## B Time Normalization Instruction

We use the following instruction to conduct time normalization using Qwen3-14B model:

“You need to normalize the provided time entity in the following sentences to the ISO 8601 standard format, either YYYY-MM-DD and YYYY-Www, and a week starts from Monday. When you normalize the time, consider the following steps: 1. Try normalizing the time entity into the YYYY-MM-DD format if you can find information for the day. 2. if you cannot find information for the day, but the month is mentioned, try normalizing it to the YYYY-MM format. 3. If both day and month are not mentioned, and the time entity is a

term related to week numbers, i.e., next week, 4 weeks and last week, try using the week number prefixed by the letter W and normalizing the time entity into the YYYY-Www format. 4. if the last three steps failed and you can’t normalize the time entity, output 9999-99-99.

The date that this document is created is row[‘DOCTIME’]. You need to consider this document time when normalizing the time entity.

The markers <time> and </time> surrounds the time entity in the sentences for you to normalize. Here are sentences extracted from a patient note: sentences. Please generate your answer in the following json format: “Answer”: <insert time normalized to the ISO 8601 standard format or 9999-99-99 if you can’t normalize the time entity>. YOUR RESPONSE MUST BE IN THIS EXACT JSON FORMAT and MAKE SURE YOU FOLLOW THE FORMAT STRICTLY. PLEASE ONLY FOCUS ON THE TIME ENTITY SURROUNDED BY THE MARKERS. DO NOT GENERATE ADDITIONAL TEXT.”

# Team UAB at Chemotherapy Timelines 2025: Integrating encoders and large language models for chemotherapy timeline generation

Vijay Jain<sup>1</sup>, Chris Coffee<sup>1</sup>, Kaiwen He<sup>1</sup>, Remy Cron<sup>1</sup>,  
Micah Cochrane<sup>1</sup>, Luis Mansilla-Gonzalez<sup>1</sup>, Danish Murad<sup>1</sup>, Akhil Nadimpalli<sup>2</sup>,  
John D. Osborne<sup>1</sup>

<sup>1</sup>University of Alabama at Birmingham, Birmingham, Alabama, USA

<sup>2</sup>Independent Researcher, Birmingham, AL, USA

{jainv, chrco, kaiwenhe, rrcron, mdcochra, lmansill, dmurad, ozborn}@uab.edu,  
{akhilnadimpalli@gmail.com}

## Abstract

Reconstructing the timeline of Systemic Anticancer Therapy (SACT) or "chemotherapy" from heterogeneous Electronic Health Record (EHR) notes is a challenging task. Rapid developments in Large Language Models (LLMs), including a range of architectural improvements and post-training refinements since the 2024 Chemotherapy Timelines Task could make this task more tractable. We evaluated the performance of 4 recently released LLMs (GPT-4.1-mini, Phi4 and 2 Qwen3 models) on this task. Our results indicate that even with a variety of prompt optimization and synthetic data training, more work is still needed to see a useful application of this work.

## 1 Introduction

Accurately extracting Systemic Anticancer Therapy (SACT) or "chemotherapy" treatment timelines from clinical narratives is essential for conducting retrospective outcome studies, enabling researchers to correlate the sequence and timing of administered regimens with long-term patient outcomes and responses. However, clinical documentation is often scattered between heterogeneous types of note. This makes both manual abstraction of timelines exceptionally laborious and error-prone and increases the complexity of development for systems abstracting these timelines. Continued advances in large language models (LLMs) with improved reasoning capabilities (OpenAI, 2025), larger context window sizes and higher overall performance may enable SACT extraction above baselines seen in the 2024 task (Yao et al., 2024). This includes newer models such as GPT-4.1 (OpenAI, 2025), a derivative of the larger proprietary model GPT-4 (Achiam et al., 2023) as well as smaller local LLMs such as Phi4 (Abdin et al., 2024) and Qwen3 (Yang et al., 2025) with reasoning ability. Furthermore, smaller masked language models of

which BERT (Devlin et al., 2019) is the canonical example have seen both architectural improvements (Warner et al., 2024) and biomedical fine-tuning since 2024 (Lee et al., 2025). However, efforts have been limited due to a single, domain-specific data set (Yao et al., 2024) that includes only three types of cancer: breast, ovarian, and melanoma. In this work, we assess the ability of recent LLMs to address this problem on the Chemotherapy Timelines 2025 Shared Task (Yao et al., 2025) for both Task 1 (where additional gold annotations are provided) and Task 2 where input is restricted to clinical notes.

### 1.1 Related Work

For both Task 1 and Task 2, SACT timelines must be generated and consolidated. A variety of approaches have been used for this, including the use of local LLMs (Yao et al., 2024) which have the ability to create non-extractive timelines that are not present in the original text. A variety of strategies can be deployed for this, of which fine-tuning (Anisuzzaman et al., 2025), retrieval augmented generation (RAG) (Arslan et al., 2024), and prompt engineering (Brown et al., 2020) are popular choices. Fine-tuning is costly, but prompt engineering is a light-weight strategy for performance improvement. One such modular prompt-engineering framework is DSPy (Khattab et al., 2024) which implements a variety of different prompt engineering strategies including Simba (Lee et al., 2024), MIPROv2 (Opsahl-Ong et al., 2024) as well as few-shot selection. Prompts provide an easy mechanism to include relevant temporal events either from the gold information in Task 1 or through encoder-based extraction methods. Fine-tuned encoders may still outperform LLMs in information extraction tasks due to their bidirectional understanding of language (the result of masked language modeling instead of autoregressive training), however, more recent results are mixed (Obeidat

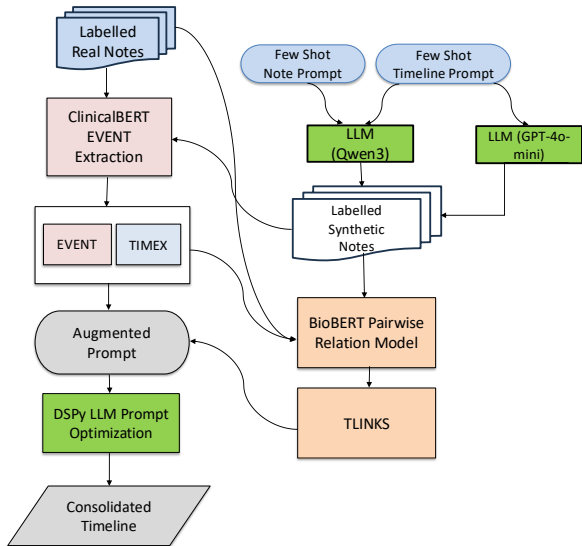


Figure 1: System Overview. Both Task 1 and Task 2 system use DSPy in conjunction with a LLM to generate JSON formatted SACT timeline predictions. The Task 1 system uses gold EVENTS and TIMEX whereas the Task 2 system directly generates those from text.

et al., 2025; Arzideh et al., 2025; Roccabruna et al., 2024). Similarly to others (Bannour et al., 2024; Tan et al., 2024) in previous tasks, we employ encoder models to better provide context for the final timeline.

## 2 Method and Materials

### 2.1 System Overview

Our system takes as input either gold annotations (EVENTS or TIMEX expressions), and TLINKs from the Task 1 gold standard or generates them using either a BioBERT (Lee et al., 2020) model (TLINKs) or a ClinicalBERT (Alsentzer et al., 2019) (EVENTs) model. An overview is shown in Figure 1. These inputs are integrated into the prompt to guide timeline generation, with the encoder models semantically rich representations that capture long-context clinical dependencies.

### 2.2 Synthetic Data Generation

Synthetic data was created to assist information extraction for TLINK identification and EVENT identification as shown in Figure 1. The goal is to improve robustness in low resource settings (Li et al., 2021).

**TLINK Generation** For TLINK synthetic data, an OpenAI GPT-4o-mini model hosted in the UAB’s firewalled Azure Enclave was used. This

Enclave is approved for PHI by the UAB’s Health System. Synthetic data was used to compensate for class imbalance in the original training data, so synthetic examples were generated only to augment minority classes (ENDS-ON and BEGINS-ON). Specific training details can be found at [The system is available on github.com/vijay0019/UAB\\_ChemoTimelines](https://github.com/vijay0019/UAB_ChemoTimelines).

**EVENT Generation** Synthetic oncology notes were generated with Qwen3-32B, using a one-shot prompting technique to address the lack of EVENT training data. Each prompt sampled drugs/regimens from a merged lexicon built by uniting entities observed in the training notes with entries from HemOnc (Warner et al., 2015), a curated open regimen vocabulary. The synthetic notes were mixed with the real corpus for a second round of fine tuning. This knowledge-guided augmentation targets the regimen names, abbreviations, and phrasings that are undersampled in the original notes. Notes are generated by randomly sampling drugs and regimens from the merged lexicon, with a constraint that no drug/regimen appears more than twice per note. Each synthetic note is conditioned to match the writing style and structure of real notes including de-identified headers and footers thereby preserving real-world patterns while preventing exposure of Protected Health Information (PHI) (Melamud and Shivade, 2019). We also introduce controlled variation e.g., domain specific abbreviations (Liu et al., 2001) and common misspellings to better reflect noisy clinical text. Entities in the synthetic notes are validated using the given list.

**Timeline Generation** Synthetic timelines were generated with the goal of creating a more comprehensive synthetic set of notes. An overview of synthetic data generation for EVENTs is shown in Figure 2. Synthetic timelines were also generated by Qwen3-32B (Yang et al., 2025) with reasoning disabled. The model was prompted with both a system prompt and a user prompt to guide synthetic timeline generation. The system prompt described the timeline generation task, including definitions and formatting preferences, i.e. if the model is following the correct format for the output (SACT entity, relation type, TIMEX3 expression) for each timeline triplet and TIMEX3 expressions are formatted correctly. The user prompt gave five example timelines from the training set and asked the model to generate a synthetic timeline for a patient having cancer with a specific primary site,

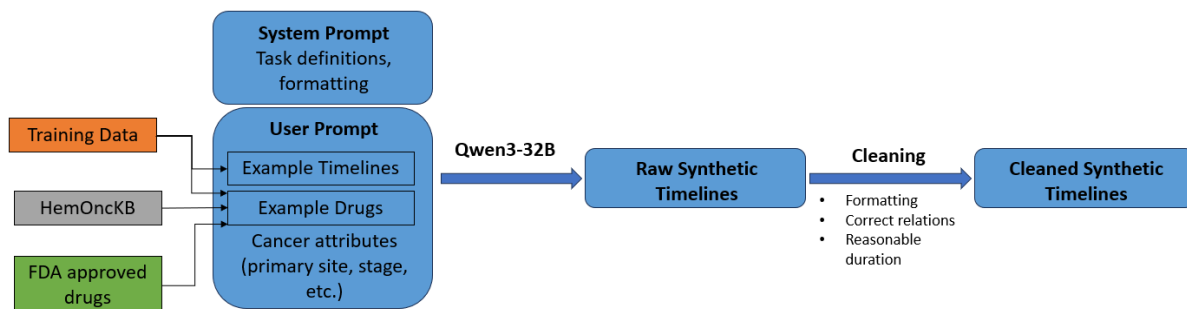


Figure 2: Synthetic Timeline generation

maximum stage of progression, stage at diagnosis, remission and recurrence status, time at start of therapy and therapy duration. The user prompt also included five SACT references for the primary site requested. These SACT references are drawn from a list that includes all SACTs in the training data, SACTs found in the HemOnc Knowledge Base (Warner et al., 2015) and SACTs found in the FDA’s approved drugs for cancer lists (National Cancer Institute, 2025). After generation of timelines, there was a cleaning step in which timelines that were incorrectly formatted, had incorrect relations or were five years or longer were removed. An attempt is made to reformat the timeline to a correct format before removal. Part of these cleaned timelines were then used as a prompt for STRS1. An example cleaned synthetic timeline with its system and user prompt, as well as its raw form before reformatting, is shown in Figure 3. The system is available at [https://github.com/vijay0019/UAB\\_ChemoTimelines](https://github.com/vijay0019/UAB_ChemoTimelines).

### 2.3 SACT EVENT Extraction

For Task 1, SACT EVENTS and TIMEX expressions are provided for the test data set, but for Task 2 we developed our own pipeline to extract SACT EVENTS (drug and regimen) mentions and TIMEX3 time expressions in clinical notes. These two outputs form the nodes for the downstream temporal relation prediction where we link each chemotherapy event to its most relevant time reference to build patient level timelines. This design mirrors the task framing: systems must identify events and time expressions, then infer their temporal relations to recover the treatment timeline. Since encoder models tend to perform well for these tasks, therefore we fine-tuned ClinicalBERT (Alsentzer et al., 2019) for NER over the official

training notes to detect SACT entities (drug/regimen spans). On our held out dev set, this achieved a high precision but low recall. While precision was strong, the relatively low recall meant that many true SACT mentions were missed. In a timeline-reconstruction setting, recall shortfalls are especially harmful because missing them means removing candidate EVENT-TIMEX3 pairs for the relation prediction, reducing the chance of recovering correct [chemo, relation, time] triplets even when the TIMEX3 recognition is accurate. For TIMEX3, we use TimeNorm for span recognition and normalization (Bethard, 2013; Laparra et al., 2018; Xu et al., 2019).

### 2.4 SACT TLINK Extraction

The relation classification task involved four categories: CONTAINS, BEGINS-ON, ENDS-ON, and no\_relation between SACT and TIMEX entities. BioBERT (Lee et al., 2020) was fine-tuned for span-based relation classification. The model extracts span representations by combining start/end token embeddings, width embeddings and entity type embeddings. For context, between-token context extraction was implemented using average pooling of tokens between subject and object entities. The final classification layer concatenates subject, context, and object representations through a two-layer feedforward network with ReLU activation. Model selection was based on macro-averaged F1-score on validation data. Predictions were filtered using a 0.5 confidence threshold and constraints on relation entity type. For overlapping predictions, more specific relations (BEGINS-ON/ENDS-ON) were prioritized over CONTAINS. Duplicate relations across patient notes were removed.



Figure 3: The system prompt and an example user prompt. Highlighted portions in the user prompt are variable. Text highlighted yellow corresponds to example timelines, green to the primary site and explanation, blue to most advanced stage and explanation, magenta to diagnosis stage and explanation, red to remission and recurrence information, dark gray to timeline duration and light gray to start year. Blue font text in the raw output corresponds to the text that was extracted to create the cleaned output.

## System Prompt

You are a clinical-note author and precise annotator specializing in oncology, with a focus on systemic anticancer therapy (SACT) timelines. Your task is to generate synthetic SACT timelines. Each timeline is represented as a Python list of lists. Each of the internal lists is of the following format:

[<Chemotherapy>,<Label>,<Time>]. How to format each:

**<Chemotherapy>**:

Refers **only** to any Systemic Anticancer Therapy (SACT) drug or regimen. SACT includes: - **Traditional cytotoxic chemotherapy agents** (e.g., oxaliplatin, cyclophosphamide). - **Endocrine therapy agents** (e.g., tamoxifen, anastrozole, letrozole). - **Targeted therapy agents** (e.g., trastuzumab, erlotinib, imatinib). - **Immunotherapy agents** (e.g., pembrolizumab, nivolumab, ipilimumab). - **SACT regimen names** (e.g., FOLFOX, AC, Pembrolizumab, Carboplatin)

**<Label>**:

Refers to the temporal relationship between the chemotherapy and the time. **Must** be one of the following tokens: - 'begins-on': the chemotherapy starts at the given time. - 'ends-on': the chemotherapy ends at the given time. - 'contains-1': the chemotherapy contains the time (the given time happens entirely within the chemotherapy).

**<Time>**:

**Must** be one in one of the following three formats: - **###-##-##**: A date: the year in 4 digits, then the month in 2 digits (leading 0 if needed), then the day in 2 digits. A date is the most common format. - **###-##**: A month: the year in 4 digits, then the month in 2 digits (leading 0 if needed). - **###-w##**: A week: the year in 4 digits, then the week number in 2 digits (leading 0 if needed). - **####**: A year: the year in 4 digits.

After drafting the timeline, perform a final self-check to make sure it is a list of lists, each internal list has exactly 3 elements and follows the format: [Chemotherapy,Label,Time].

Note: **do not overthink**, and limit chain-of-thought reasoning to **500 words at most**.

## User Prompt

Below are some chemotherapy timelines. Pay close attention to their formatting.

Example 1: [['chemotherapy', 'contains-1', '2009-10-22'], ['chemo', 'contains-1', '2009-10-22']]

Example 2: [['chemo', 'contains-1', '2006-w11']]

Example 3: [['tamoxifen', 'contains-1', '2003']]

Example 4: [['adriamycin', 'contains-1', '2013-07-24'], ['adriamycin', 'contains-1', '2013-09-25'], ['adriamycin', 'contains-1', '2013-09-04'], ['adriamycin', 'contains-1', '2013-08-14'], ['cytoxan', 'contains-1', '2013-07-24'], ['cytoxan', 'contains-1', '2013-09-25'], ['cytoxan', 'contains-1', '2013-09-04'], ['cytoxan', 'contains-1', '2013-08-14'], ['taxol', 'contains-1', '2013-12-11'], ['taxol', 'contains-1', '2013-10-16'], ['taxol', 'contains-1', '2013-11-06'], ['doxorubicin', 'begins-on', '2013-07-24'], ['cyclophosphamide', 'begins-on', '2013-07-24'], ['doxorubicin', 'ends-on', '2013-09-25'], ['cyclophosphamide', 'ends-on', '2013-09-25'], ['paclitaxel', 'begins-on', '2013-10-16'], ['ac', 'contains-1', '2013-09-25'], ['ac', 'contains-1', '2013-09-04'], ['ac', 'contains-1', '2013-07-24'], ['a/c', 'contains-1', '2013-07-24']]

Example 5: [['tamoxifen', 'begins-on', '2013-w05'], ['tamoxifen', 'begins-on', '2013-01'], ['tamoxifen', 'begins-on', '2013-w04'], ['tamoxifen', 'begins-on', '2012-01'], ['tamoxifen', 'ends-on', '2018-01']]

Now generate a new timeline for a cancer patient with the following characteristics: - Primary site of cancer: **breast Therapies to treat this cancer type include ac, docetaxel, ac, arimidex, and taxotera**. - Maximum stage of progression: **stage IV Cancer has spread (metastasized) outside of the original site to other organs or distant areas of your body. This is also known as metastatic cancer**. - Stage of cancer at diagnosis: **stage II. The tumor has grown larger and possibly spread to nearby lymph nodes**. - **The cancer has not gone into remission**. - Systemic anticancer therapy (SACT) has gone on for **23** months. The new timeline must use the **exact same style and formatting** as the example timelines and it must begin in the year **2016**.

## 2.5 SACT Timeline Extraction System (STES)

The STES employs a multi-iteration approach using the DSPy framework with large language models to process clinical reports and construct temporal treatment timelines. The system begins by creating report clumps that fit within the model's context window (typically  $\frac{1}{8}$  of the total context size) and groups reports by patient ID to maintain temporal coherence. Each iteration processes these clumps through a SACTTimelineUpdate module that extracts drug names exactly as they appear in clinical text. This includes brand names, generic names, abbreviations, and variations—along with their temporal relations (*begins-on*, *ends-on*, *contains-1*) and associated dates with varying levels of specificity (year, month, day, or week).

The system implements an incremental timeline construction strategy where each processed report clump updates the existing timeline by adding new events and removing conflicting ones through an Update object containing *add* and *remove* lists. To ensure robustness, the system employs a retry mechanism across multiple language model instances with different temperature settings, falling

back to empty updates when all models fail to generate valid responses. The final timeline undergoes deduplication and chronological sorting based on date components (year, month, day, week) followed by drug name and relation type. Date objects are converted to competition-standard string formats, and the system validates date formatting through regex patterns before generating the final JSON output for each patient's treatment timeline.

An enhanced version of the SACT timeline extraction system (task1\_v2\_summaries\_plus) incorporates running summaries to maintain contextual information across report processing iterations. The system generates and updates a comprehensive treatment summary that captures key treatment phases, medication regimens, temporal milestones, and treatment response indicators mentioned in the reports. This summary serves as persistent memory between iterations, allowing the model to maintain coherence when processing large patient records that exceed context window limitations. The summary is structured to include treatment overviews, detailed medication histories with both generic and brand names, protocol documentation,

timeline reconciliation notes, and clinical observations, effectively creating a condensed narrative of the patient’s treatment journey that informs subsequent timeline extraction decisions.

Task 2 implements a fundamentally different architecture designed for chemotherapy event extraction from concatenated clinical chunks rather than structured report processing. Unlike Task 1’s report-centric approach, Task 2 employs a three-stage pipeline consisting of ChemoNotesTimeline for initial event extraction, ChemoTimelineUpdate for incremental timeline construction, and ChemoTimelineCleanup for deduplication and conflict resolution. The system intelligently manages token usage by applying a summarization step (ChemoNotesTimeline) when content exceeds a configurable threshold (typically 25% of context window), and implements dynamic timeline cleanup when the number of events surpasses a specified limit. Task 2 also incorporates dual model configurations with different repetition penalty settings to handle various text patterns and includes more sophisticated chunk concatenation strategies that respect document boundaries and optimize context window utilization. For the ablation study, we used a slightly modified version of our submission model; whereas the submission version forced the model to choose from a hard-coded list of both train and dev gold terms, for the ablation study the requirement was relaxed to a generic string. The full prompt is shown in Appendix A.

## 2.6 Chemotherapy Timeline Experiments

An early system (Version 1) and a later updated system were used for test submission results for Task 1. The updated system differed only in terms of slight variations to the zero-shot (manually generated) prompt and some post-processing steps related to pruning timeline entities based on multiple passes through the reports. A newer system (Version 2) is used on the development set. This features additional changes to the prompt, including the addition of a running summary, a better timeline example, and an LLM-generated example summary. It also included minor improvements such as the report date and restricting to a single pass through the notes in chronological (rather than random) order. The EVENT and TLINK performance is reported on the dev data set only, as runs on the test set were not completed prior to the task deadline. Local models were run on A100 40GB VRAM GPUs, GPT-4.1-mini was running in the same Azure En-

clave as the GPT-4o-mini model used for synthetic TLINK generation.

## 3 Results

### 3.1 Official System Results

Official system results are shown in Table 1. As expected, the smaller Phi4:14B is outperformed by GPT-4.1-mini.

Table 1: Version 1.1 prompt provides additional instructions to avoid ungrounded temporal relations and logic to remove timelines that show up infrequently in iterations. Entities are used as input for all Subtask 1 systems.

System	Task	LLM	Brca	Mela	Ovca	Avg
UABv1	1	Phi4:14B	0.310	0.160	0.217	0.229
UABv1.1	1	Phi4:14B	0.259	0.333	0.244	0.279
UABv1.1	1	GPT-4.1-mini	0.418	0.308	0.296	0.341
UABv0	2	Phi4:14B	0.232	0.265	0.188	0.228

Table 2: Token level EVENT recognition on the test set using ClinicalBERT trained on Actual, Synthetic, and Actual+Synthetic notes.

Training data	Prec.	Recall	F1	Acc.
Actual	<b>95.6</b>	83.0	88.9	99.76
Synthetic	64.2	81.8	71.9	99.26
Actual+Synthetic	95.2	<b>87.8</b>	<b>91.4</b>	<b>99.81</b>

### 3.2 EVENT Extraction Results

We fine-tuned ClinicalBERT (Alsentzer et al., 2019) for token-level EVENT classification on CoNLL-style inputs under three training regimes: real (human-authored) notes, synthetic (LLM-generated) notes, and their mixture. On the test set, the mixture attained 91.4 F1 (P=95.2, R=87.8), a 2.8% relative F1 increase over real-only, accompanied by a 5.8% relative recall increase and a 0.4% relative precision decrease. In comparison, synthetic-only yielded 71.9 F1, a 19.0% relative decrease vs. real-only, with precision 32.8% lower and recall 1.5% lower. Augmenting real notes with synthetic text yielded SACT NER P=0.9524, R=0.8781, F1=0.9137. The higher recall expands the pool of EVENT candidates available to downstream event-time relation classification. These results are consistent with evidence that LLM-generated, ontology-guided synthetic text can improve clinical NER by increasing coverage of rare surface forms without materially degrading precision (Dao et al., 2025). Overall token accuracy is

≈99% across settings so we de-emphasize accuracy given severe class imbalance and instead focus on precision, recall, and F1 for the EVENT class.

**Impact of Synthetic Notes on EVENT Extraction** Replacing real notes with non-timeline sourced synthetic notes reduces recall and increases false positives: TPs decrease to 9,927 (from 10,076), FNs increase to 2,209 (from 2,060), and FPs increase to 5,534 (from 466). By contrast, training on the mixture improves recall with a small precision cost: TPs increase to 10,657, FNs fall to 1,479 (≈28% fewer than real-only), and FPs increase modestly to 533 (vs. 466). Overall, synthetic-only induces a high false-positive rate, whereas the mixture identifies 581 additional true-positive EVENT tokens relative to real-only. Timeline-sourced notes were judged to be poorer in quality and ultimately were not used for EVENT extraction.

### 3.3 TLINK Extraction Results

The TLINK extraction was assessed on the development data set, since no TLINKS were provided as part of the Task 1 test data set. The model achieved 89.4% accuracy with a macro F1-score of 0.889, performing particularly well on positive relations (F1: 0.944).

### 3.4 Zero-Shot and Few-Shot Local LLM Evaluation

In addition to the official test results, we included an updated set of results on the dev set with additional LLMs in Table 3.

## 4 Discussion

Our results indicate that despite recent LLM improvements in a range of tasks, the identification of chemotherapy timelines is not a task that can be done well "out of the box" without significant engineering. Only our encoder models, using fine-tuning on a sufficient amount (supplemented with synthetic data) of training data generated reliable performance improvements without significant human intervention. Of interest in the development evaluation, that ablating the chain-of-thought for Qwen3:32B did drop performance, but we lacked time to assess if this generalized to other models.

Larger local LLMs could have been fine-tuned (at greater cost or time) to improve performance similar to previous work fine-tuning Flan-T5 in

the 2024 task (Haddadan et al., 2024). The incorporation of additional Performance Efficient Fine-Tuning, Retrieval Augmented Generation, soft prompting and utilization of the existing modifier information in the gold standard likely could have improved results.

We evaluated DSPy’s suite of primarily bootstrapping-based methods including Simba(Lee et al., 2024), MIPROv2(Opsahl-Ong et al., 2024) and GEPA(Agrawal et al., 2025), but all either failed to complete with DSPy related errors and/or yielded preliminary results that discouraged debugging. Silver-quality examples of individual timeline chronological updates congruent with context window size would perhaps have been more useful. We are given gold timelines, but not gold timeline updates, and it is a non-trivial task to generate useful examples of correct updates. Additionally, generating timelines and summaries separately for each report and then iteratively pooling them, rather than our cumulative approach, may yield better results and will be explored in future work.

## 5 Conclusion

Overall, this task remains challenging even with the use of LLMs such as GPT-4.1-mini suggesting that currently, specialized training is required to achieve results comparable to humans. Current effort to create synthetic timelines do not improve performance. We found it was substantially easier through fine-tuning to obtain reliable, fast results with encoder models than to fine-tune prompts for LLMs.

### Limitations

Due to the sensitive nature of the data, a Data Use Agreement is required to obtain the data needed to replicate our results. A more complete evaluation of modern LLMs was not feasible due to cost, so GPT-4.1-mini was the only large model fully evaluated.

### Acknowledgments

We acknowledge UAB Research Computing for access to the High Performance Computing Infrastructure. Claude Code was used for the generation some of the code by 3 of the developers.

System	Subtask	LLM	Input	CoT	Learning	Breast	Melanoma	Ovarian
UABv1	1	Phi4:14B	Entities	Y	Zero-Shot	0.286	0.145	0.191
UABv1.1	1	Phi4:14B	Entities	Y	Zero-Shot	0.248	0.156	0.196
UABv2	1	Qwen3:32B	Entities	Y	Zero-Shot	0.535	0.593	0.260
UABv2	1	Qwen3:32B	Entities	N	Zero-Shot	0.300	0.314	0.297
UABv2	1	Qwen3:30B:3A	Entities	Y	Zero-Shot	0.659	0.507	0.266
UABv2	1	Qwen3:30B:3A	Entities	Y	Few-Shot	0.520	0.230	0.284
UABv0	2	Phi4:14B	None	Y	Zero-Shot	0.286	0.530	0.159

Table 3: Version 2 includes many changes to the prompt and the addition of summaries.

## References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziemis, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, and 1 others. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*.
- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. [Publicly available clinical bert embeddings](#). *Preprint*, arXiv:1904.03323.
- D.M. Anisuzzaman, Jeffrey G. Malins, Paul A. Friedman, and Zach I. Attia. 2025. [Fine-tuning large language models for specialized use cases](#). *Mayo Clinic Proceedings: Digital Health*, 3(1):100184.
- Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. [A survey on rag with llms](#). *Procedia Computer Science*, 246:3781–3790. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).
- Kamyar Arzideh, Henning Schäfer, Héctor Allende-Cid, Giulia Baldini, Thomas Hilser, Ahmad Idrissi-Yaghir, Katharina Laue, Nilesh Chakraborty, Niclas Doll, Dario Antweiler, and 1 others. 2025. From bert to generative ai-comparing encoder-only vs. large language models in a cohort of lung cancer patients for named entity recognition in unstructured medical reports. *Computers in Biology and Medicine*, 195:110665.
- 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 Natural Language Processing Workshop*, pages 406–416, 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.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- An Dao, Hiroki Teranishi, Yuji Matsumoto, Florian Boudin, and Akiko Aizawa. 2025. [Overcoming data scarcity in named entity recognition: Synthetic data generation with large language models](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 328–340, Vienna, Austria. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Shohreh Haddadan, Tuan-Dung Le, Thanh Duong, and Thanh 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 Natural Language Processing Workshop*, pages 382–393.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, Heather Miller, and 1 others. 2024. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.

- Egoitz Laparra, Dongfang Xu, and Steven Bethard. 2018. [From characters to time intervals: New paradigms for evaluation and neural parsing of time normalizations](#). *Transactions of the Association for Computational Linguistics*, 6:343–356.
- Hoon Lee, Dongyoon Hwang, Donghu Kim, Hyunseung Kim, Jun Jet Tai, Kaushik Subramanian, Peter R Wurman, Jaegul Choo, Peter Stone, and Takuma Seno. 2024. Simba: Simplicity bias for scaling up parameters in deep reinforcement learning. *arXiv preprint arXiv:2410.09754*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Simon A Lee, Anthony Wu, and Jeffrey N Chiang. 2025. Clinical modernbert: An efficient and long context encoder for biomedical text. *arXiv preprint arXiv:2504.03964*.
- Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Nataraajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. 2021. [Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition](#). *Journal of the American Medical Informatics Association*, 28(10):2193–2201.
- Hongfang Liu, Yves A. Lussier, and Carol Friedman. 2001. [A study of abbreviations in the UMLS](#). In *Proceedings of the AMIA Symposium*, pages 393–397.
- Oren Melamud and Chaitanya Shivade. 2019. [Towards automatic generation of shareable synthetic clinical notes using neural language models](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- National Cancer Institute. 2025. Drugs approved for different types of cancer. <https://www.cancer.gov/about-cancer/treatment/drugs/cancer-type>. Accessed: 2025-08-04.
- Motasem S Obeidat, Md Sultan Al Nahian, and Ramakanth Kavuluru. 2025. Do llms surpass encoders for biomedical ner? In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, pages 352–358. IEEE.
- OpenAI. 2025. Introducing GPT-4.1 in the API. Online blog post (OpenAI). Retrieved from OpenAI.
- OpenAI. 2025. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>. Accessed: 2025-08-20.
- Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. *arXiv preprint arXiv:2406.11695*.
- Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. 2024. Will llms replace the encoder-only models in temporal relation classification? *arXiv preprint arXiv:2410.10476*.
- 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 Natural Language Processing Workshop*, pages 417–421, Mexico City, Mexico. Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory-efficient, and long-context fine-tuning and inference](#). *arXiv preprint arXiv:2412.13663*.
- Jeremy L. Warner, Andrew J. Cowan, Aric C. Hall, and Peter C. Yang. 2015. [Hemonc.org: A collaborative online knowledge platform for oncology professionals](#). *Journal of oncology practice*, 11(3):e336–e350.
- Dongfang Xu, Egoitz Laparra, and Steven Bethard. 2019. [Pre-trained contextualized character embeddings lead to major improvements in time normalization: a detailed analysis](#). In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (\*SEM 2019)*, pages 68–74, Minneapolis, Minnesota. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- 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.
- Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. 2025. Overview of the 2025 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*.

## A System Prompt

Figure 4: Full System Prompt used for SACT timeline generation (Part 1).

```
Update SACT timeline based on patient reports with running summary.

SACT is defined as follows:
"Systemic anticancer therapy (SACT), which includes traditional cytotoxic
  chemotherapy, endocrine therapy, targeted therapy, and immunotherapy, has
  both a low therapeutic index as well as synergistic potential when agents
  are given in combination."

Drug Names: Extract drug names EXACTLY as they appear in the clinical text (
  except be sure to put them in lowercase). Do NOT normalize or convert to
  generic names.
Include ALL variations found in the text:
- Brand names (cytoxan, taxotere, abraxane)
- Generic names (cyclophosphamide, docetaxel, paclitaxel)
- Abbreviations (tc, ac, a/c)
- Generic terms (chemotherapy, chemo)
- Slight variations/typos (docetaxol for docetaxel)

If the text mentions both "cyclophosphamide" and "cytoxan", include both as
  separate entries.
If the text mentions both "chemotherapy" and specific drug names, include both.
  Only include drugs that have a temporal relation in the text.

Ignore references to cancer/neoplasms, genetic variations (e.g. HER2), and non-
  SACT procedures such as radiation therapy.

Relations:
- 'begins-on': treatment/medication starts
- 'ends-on': treatment/medication ends
- 'contains-1': treatment occurred within timeframe
'begins-on' and 'ends-on' supersede 'contains-1' for the same drug/date
  combination. Only use them if the text explicitly states the start or end
  date of the treatment.

Acceptable date formats:
1. Specify year, month, and day.
2. Specify year and week.
3. Specify year and month.
4. Specify year only.
Try to be as specific as possible, but do not invent dates that are not
  mentioned in the text.

Keep in mind that the reports are only a subset of the full timeline, so there
  may be events in the timeline that are not mentioned in the reports. Do not
  remove events simply because they are not mentioned in the reports.

If a report doesn't have temporal relations, that likely means the report does
  not contain any relevant information for the timeline. Avoid adding events
  based solely on hypothetical or planned mentions without temporal grounding
  .

Running Summary: Maintain a concise summary of the patient's SACT treatment
  journey, including:
- Key treatment phases and regimens
- Major treatment changes or progressions
- Important temporal milestones
- Treatment response indicators mentioned in reports
Update the summary to reflect new information from current reports while
  preserving important historical context.

Example output format:
[[ ## updated_summary ## ]]
**Treatment Overview:** Chemotherapy begins week 32 of 2013, with documented
  treatment from August 8 - October 10, 2013

"""
```

Figure 5: Full System Prompt used for SACT timeline generation (Part 2).

```

**Medications Administered:**
- **Cyclophosphamide (Cytoxan):** August 8 - October 10, 2013
  - Individual doses: August 29, September 19, October 10
- **Docetaxel (Taxotere):** August 8 - October 10, 2013
  - Individual doses: August 29, September 19, October 10
  - *Note: "Docetaxol" also documented (likely transcription error)*

**Protocol Documentation:**
- **TC regimen** active during August 2013
- **Chemotherapy** contains specific date August 10, 2013
- Standard 3-week cycling pattern evident

**Timeline Reconciliation:**
- Week 32 of 2013 corresponds to ~August 5-11 timeframe
- Treatment initiation August 8 aligns with weekly scheduling
- 4 documented treatment cycles completed over 9-week period

**Clinical Notes:** Complete documentation includes both generic (
  cyclophosphamide, docetaxel) and brand names (Cytoxan, Taxotere) with minor
  spelling variant recorded.
[[ ## timeline_update ## ]]
Update(
  add=[
    ('chemotherapy', 'begins-on', Date(year=2013, month=None, day_of_month=
None, week_of_year=32)),
    ('chemotherapy', 'contains-1', Date(year=2013, month=8, day_of_month
=10, week_of_year=None)),
    ('cyclophosphamide', 'begins-on', Date(year=2013, month=8, day_of_month
=8, week_of_year=None)),
    ('cyclophosphamide', 'ends-on', Date(year=2013, month=10, day_of_month
=10, week_of_year=None)),
    ('cytoxan', 'begins-on', Date(year=2013, month=8, day_of_month=8,
week_of_year=None)),
    ('cytoxan', 'contains-1', Date(year=2013, month=8, day_of_month=29,
week_of_year=None)),
    ('cytoxan', 'contains-1', Date(year=2013, month=9, day_of_month=19,
week_of_year=None)),
    ('cytoxan', 'contains-1', Date(year=2013, month=10, day_of_month=10,
week_of_year=None)),
    ('docetaxel', 'begins-on', Date(year=2013, month=8, day_of_month=8,
week_of_year=None)),
    ('docetaxel', 'ends-on', Date(year=2013, month=10, day_of_month=10,
week_of_year=None)),
    ('docetaxol', 'begins-on', Date(year=2013, month=8, day_of_month=8,
week_of_year=None)),
    ('taxotere', 'begins-on', Date(year=2013, month=8, day_of_month=8,
week_of_year=None)),
    ('taxotere', 'contains-1', Date(year=2013, month=8, day_of_month=29,
week_of_year=None)),
    ('taxotere', 'contains-1', Date(year=2013, month=9, day_of_month=19,
week_of_year=None)),
    ('taxotere', 'contains-1', Date(year=2013, month=10, day_of_month=10,
week_of_year=None)),
    ('tc', 'contains-1', Date(year=2013, month=8, day_of_month=None,
week_of_year=None))
  ],
  remove=[
    ('cytoxan', 'begins-on', Date(year=2013, month=8, day_of_month=1,
week_of_year=None))
  ]
)
[[ ## completed ## ]]
"""

```

# UW-BioNLP at ChemoTimelines 2025: Thinking, Fine-Tuning, and Dictionary-Enhanced LLM Systems for Chemotherapy Timeline Extraction

Tianmai M. Zhang\*, Zhaoyi Sun\*, Sihang Zeng\*, Chenxi Li\*,  
Neil F. Abernethy, Barbara D. Lam, Fei Xia, Meliha Yetisgen

University of Washington

Correspondence: melihay@uw.edu

## Abstract

The ChemoTimelines shared task benchmarks methods for constructing timelines of systemic anticancer treatment from electronic health records of cancer patients. This paper describes our methods, results, and findings for subtask 2—generating patient chemotherapy timelines from raw clinical notes. We evaluated strategies involving chain-of-thought thinking, supervised fine-tuning, direct preference optimization, and dictionary-based lookup to improve timeline extraction. All of our approaches followed a two-step workflow, wherein an LLM first extracted chemotherapy events from individual clinical notes, and then an algorithm normalized and aggregated events into patient-level timelines. Each specific method differed in how the associated LLM was utilized and trained. Multiple approaches yielded competitive performances on the test set leaderboard, with fine-tuned Qwen3-14B achieving the best official score of 0.678. Our results and analyses could provide useful insights for future attempts on this task as well as the design of similar tasks.

## 1 Introduction

Electronic health records (EHRs) contain rich information on treatment courses, but extracting temporal relationships is challenging due to variability in care and linguistic complexity (Olex and McInnes, 2021; Gholipour et al., 2023). Oncology regimens often deviate from planned schedules through dose changes or delays, with such modifications usually recorded only in unstructured notes that require chronological alignment (Wang et al., 2020). Clinical narratives add further difficulty with relative or vague time expressions and inconsistent date formats (Sun et al., 2013, 2015). Even experts may diverge in interpreting underspecified terms, making accurate normalization and sequencing a persistent challenge for clinical NLP systems.

The ChemoTimelines shared task<sup>1</sup> (Yao et al., 2024, 2025) was created to benchmark systems for constructing systemic anticancer treatment (SACT) timelines directly from EHR notes. It consists of two subtasks. In subtask 1, besides the raw EHRs, gold standard annotations of treatment events (EVENTs) and time expressions (TIMEX3s) for each patient EHR note are provided, and the task is to determine temporal relations between them on the patient level. In subtask 2, the task is to extract the patient-level treatment timeline with only the raw EHR notes available. We focus on subtask 2 to provide insights into an end-to-end treatment timeline extraction system.

Large language models (LLMs) demonstrate superior comprehension and information extraction ability, and were widely used in the previous year of the challenge (Haddadan et al., 2024; Zhang et al., 2024). Without dedicated prompt engineering and chain-of-thought reasoning (Wei et al., 2023), zero-shot prompting on LLMs has shown poor performance (Zhang et al., 2024) in the timeline extraction task. Domain-adapted fine-tuning has proven effective for SACT timeline extraction, with models like Flan-T5-XXL (Chung et al., 2022) and PubMedBERT (Gu et al., 2021) achieving strong results (Haddadan et al., 2024; Tan et al., 2024). However, these approaches have predominantly utilized older or smaller-scale architectures, such as BART (Lewis et al., 2019) and Flan-T5-XXL (Chung et al., 2022), and predicted timelines based on sentence-level contexts. Recent studies on scaling laws suggest that leveraging larger powerful models with rich context presents a clear opportunity for further improvement (Kaplan et al., 2020). In parallel, pipeline systems—which first extract events with a curated dictionary and then identify relations (Haddadan et al., 2024; Wang et al.,

\*These authors contributed equally.

<sup>1</sup><https://sites.google.com/view/chemotimelines2025>



2024)—have been developed but typically show inferior performance to end-to-end systems. Despite integrating external knowledge, the pipeline approach may still be suboptimal.

Building on previous efforts, we explore a variety of strategies to fill the gaps. First, to analyze the impact of LLM-based reasoning, we compare a baseline prompting system with a reasoning system. Second, to rethink the impact of external knowledge, we design a dictionary-enhanced extraction approach. Finally, to explore multiple training strategies, we conduct supervised fine-tuning (SFT) and direct preference optimization (DPO) on the latest LLMs. Our fine-tuned Qwen3-14B system wins first place in the challenge leaderboard. We provide several novel insights into the task that may inform future attempts on this task, as well as the design of similar tasks.

## 2 Problem Formulation

The SACT timeline extraction task for each patient is formulated as extracting  $m$  triplets  $\mathcal{T} = \{ \langle s_j, r_j, t_j \rangle \}_{j=1}^m$  from a series of  $n$  clinical notes  $X = \{x_1, \dots, x_n\}$  of the patient, where  $s$  indicates a SACT entity,  $t$  is a TIMEX, and  $r$  indicates the relation between  $s$  and  $t$  selected from BEGINS-ON, ENDS-ON, and CONTAINS-1. Following practices of last year’s teams (Haddadan et al., 2024), we used note-level gold-standard relation annotations on the training set as the training data for our systems.

Detailed descriptions of the task framework and the dataset can be found on the shared task’s website<sup>1</sup> or in the overview paper (Yao et al., 2025). In short, the dataset covers three cancer types (breast cancer, melanoma, and ovarian cancer) and was split by the task organizer into a training set (69 patients, 2,910 note files), a development set (27 patients, 1,272 note files), and a test set (53 patients, 2,121 note files). Teams participating in the shared task received the annotated training and development sets for the development of their systems. The unannotated test set was released a few days before the submission deadline for teams to run their systems and submit predictions.

To extract triplets  $\mathcal{T}$  from clinical notes  $X$ , previous patient-level approaches (Zhang et al., 2024) directly processed the entire  $X$ , which may overwhelm the LLM, while the sentence-level approach (Haddadan et al., 2024) separately processed sentences in each  $x_i$ , which may lack global

context. In contrast to these approaches, we leverage a note-level approach that splits the entire task into two steps: (1) **note-level extraction**: extracting triplets  $\mathcal{T}_i$  from individual notes  $x_i$ , with or without format postprocessing, and (2) **timeline aggregation**: normalizing the TIMEXs and aggregating  $\{\mathcal{T}_i\}_{i=1}^n$  into a patient-level timeline  $\mathcal{T}$ . This setting allows decoupling of LLM extraction performance from the final timeline-level performance, enabling us to evaluate and optimize the methods for each step. LLM-based methods for step 1 are described in Section 3, and the aggregation method for step 2 is explained in Section 4.

## 3 Note-Level Extraction

We compared 5 different strategies for the note-level extraction task, providing insights from various aspects. We further included an ensemble method in our challenge submission to probe the relationship between note-level extraction and timeline aggregation.

### 3.1 Prompting Baseline

The baseline approach uses prompt-based, one-pass LLM inference. A prompt template (Appendix A.1) was carefully curated based on the task definition and provided note-level gold annotations, encompassing detailed task instructions, in-context examples, and formatting requirements for a structured output. Each clinical note was appended to the prompt without preprocessing. LLMs generate extracted chemotherapy events  $\mathcal{T}_i$  from each note  $x_i$  as a JSON array.

### 3.2 Thinking

Recent advances have shown improved reasoning and end-task performance when enabling a chain-of-thought (CoT) before generating answers (Wei et al., 2023). In light of this, we enabled the thinking mode of the models in the prompting baseline using the same prompt to explore whether CoT could improve the timeline extraction.

During error analysis, we observed text span discrepancies between LLM extractions and note-level gold-standard annotations, which sometimes resulted in false negatives in exact match evaluation. Therefore, we further designed the following postprocessing rules for our prediction submission based on the thinking method: (1) for SACT names containing "chemo", remove all descriptors before them, such as "adjuvant" and "neoadjuvant";

(2) for SACT names combined with a slash (e.g., "Doxorubicin/Cyclophosphamide"), split them into separate events; (3) remove unnecessary words in time expressions, such as "approximately", "about", "around", and "in". We do not include the postprocessing step in development set evaluation results (Table 1) for a fairer initial evaluation.

### 3.3 Dictionary-Enhanced Extraction

We rethought and adapted the approach used by the LAILab team in last year’s Task 2 (Haddadan et al., 2024), structuring it into a three-step pipeline.

**Step 1: Dictionary-based chemotherapy event extraction.** Given a clinical note, we first applied a self-constructed chemotherapy dictionary for keyword matching. All matches were tagged with `<e>` and `</e>`. The dictionary was built from three sources: (1) HemOnc.org<sup>2</sup>, where we created separate dictionaries for breast cancer, melanoma, and ovarian cancer including regimen names, drug names, and abbreviations; abbreviations with only two letters were removed to reduce false positives (e.g., "AT"); (2) generic mentions such as "chemotherapy" and "chemo" from the baseline system<sup>3</sup>; and (3) annotated chemotherapy mentions from the training and development sets of Subtask 1. No test set annotations were used. Only the drug names were incorporated into the dictionary; no labeled spans or relations were carried over to Subtask 2. The complete dictionary is provided in Appendix B.

**Step 2: LLM-based double checking and augmentation.** Sentences containing dictionary tags were passed to the Qwen-3 Thinking model for verification, which reduced false positives and recovered false negatives. The prompt template is in Appendix A.2.

**Step 3: Context-enhanced relation extraction.** For each verified sentence, we constructed a window of the anchor sentence plus its preceding and following sentences. This context was fed into Qwen-3 for generating chemotherapy–time relation triplets. The motivation for using a local window was efficiency: fewer than 6% of sentences in the dataset contain SACT annotations (Table S1), and chemotherapy events and time expressions generally appear within two consecutive sentences. Based on these observations, we modified the baseline system’s prompt for local sentence-

level relation extraction to improve efficiency (Appendix A.3).

### 3.4 Supervised Fine-Tuning (SFT)

Motivated by LAILab’s success in using SFT in the previous challenge (Haddadan et al., 2024), we performed SFT to adapt LLMs for note-level extraction using the provided gold annotations. Input prompts were structured using the same template as our prompting baseline. For the training targets, the output for each note was serialized into a JSON object containing a list of dictionaries. Each dictionary represented a single extracted event with three mandatory keys: "SACT", "relation", and "time".

Our SFT approach differs from the method proposed by LAILab (Haddadan et al., 2024) in three key aspects. First, their method operates at the sentence level, providing the model with only a target sentence and its immediate neighbors as context. In contrast, our note-level approach allows the model to leverage the richer contextual information present in the entire clinical note. Second, for output generation, they employed a specialized triplet linearization algorithm (Huguet Cabot and Navigli, 2021). We adopt a potentially more flexible strategy by serializing the extracted relations into a structured JSON object. Finally, while their best performance was achieved by finetuning Flan-T5-XXL (Chung et al., 2022), we scale up to a 14B-parameter model from Qwen3 (Yang et al., 2025), a more recent and advanced model family.

### 3.5 Direct Preference Optimization (DPO)

Recent work suggests that models trained via SFT tend to memorize the training data, while subsequent training with reinforcement learning can enhance generalization and alignment with human preference (Chu et al., 2025). Motivated by this, we framed the note-level extraction task as a preference alignment problem. Specifically, we defined the preference as: (1) the extraction is expected to align with the style in gold annotations and (2) the note-level extraction may favor outputs with higher recall over precision, operating on the assumption that the downstream timeline aggregation process would manage deduplication and resolve conflicting extractions.

To implement this, we employed an iterative DPO approach to construct a preference dataset and refine the policy model (Zhang et al., 2025; Tu et al., 2025; Rafailov et al., 2024). First, we warmed up a policy model by training it for 5 epochs via

<sup>2</sup>[https://hemonc.org/wiki/Main\\_Page](https://hemonc.org/wiki/Main_Page)

<sup>3</sup><https://github.com/HealthNLPorg/chemoTimelinesEval>

SFT. Next, to generate preference pairs, we used this initial model to produce 8 candidate outputs for each instance in the training set. For each set of candidates, we identified the output with the highest recall as the chosen response ( $y_w$ ) and the one with the lowest recall as the rejected response ( $y_l$ ). This process yielded a preference dataset of pairs where the chosen and rejected responses were distinct. We then further trained the SFT warmup model using DPO (Rafailov et al., 2024) on this dataset.

### 3.6 Ensemble Method

For the ensemble method, we concatenated note-level predictions generated by three models: SFT, SFT + DPO, and Thinking + Postprocessing. These combined predictions were then passed through the normalization and aggregation pipeline to produce final patient-level timelines. We excluded the dictionary-enhanced extraction approach from the ensemble because its pipeline differs substantially from the other methods and introduces systematic false positives due to keyword matching.

## 4 Timeline Aggregation

Similar to how the task organizer constructed gold timelines automatically (Yao et al., 2024), all note-level system outputs underwent two subsequent steps, normalization and aggregation, to obtain patient-level timelines as final outputs.

In the normalization step, time expressions in note-level outputs were converted into standardized ISO time using the CLUlab’s Timenorm synchronous context-free grammar module (Bethard, 2013). The original Timenorm was written in Scala, we reproduced its core functions in Java. The "DOCTIME" of each note was used as a temporal anchor for relative time expressions extracted from that note. Such "DOCTIME" was identified via a regular expression that detects eight consecutive digits in the note text. Relative time expressions that could not be normalized by Timenorm were discarded.

Normalized events were then de-duplicated and aggregated using the official aggregation script (`docker_output_to_timeline.py`) provided by the task organizer<sup>4</sup>.

<sup>4</sup><https://github.com/HealthNLPorg/chemoTimelinesEval>

## 5 Experimental Setup

### 5.1 Models

Our experiments for note-level event extraction utilized the following open-source LLMs: Qwen3 series of general-purpose dense models (Yang et al., 2025) from 4B to 32B, plus a mixture-of-experts model Qwen3-30B-A3B (2507), and Google’s specialized model for medicine, MedGemma-27B (Sellergren et al., 2025). All models were obtained from Hugging Face.

We used the `vllm` package under Python 3.10 for LLM inference, and `LlamaFactory` for LLM fine-tuning. Sampling parameters for LLM inference followed the setting recommended by the Qwen3 team: `temperature=0.6`, `top_p=0.95`, `top_k=20`, and `min_p=0`. The default maximum output length (`max_tokens`) was set to 4,096. Up to 4 NVIDIA A100 GPUs were utilized for either model inference or model training.

For dense Qwen3 models, the thinking mode was enabled by setting the `enable_thinking` parameter to `True`, and `max_tokens` was changed to 20,480 to allow complete outputs; for Qwen3-30B-A3B (2507), the non-thinking model and the thinking model are two separate models.

For SFT, we turned off the thinking mode of the Qwen3 model and employed LoRA for parameter-efficient finetuning of the model over 10 epochs. For DPO, we obtained preference datasets of 9 pairs for Qwen3-14B, 27 pairs for Qwen3-8B, and 30 pairs for Qwen3-4B. We trained with DPO for 10 epochs. Despite the small sample size, we observed a consistent improvement in reward accuracy during training (Figure 1), which aligns with recent studies demonstrating effective reinforcement learning from a limited number of samples (Wang et al., 2025).

### 5.2 System Evaluation

As instructed, system performance was assessed using the strict matching criterion, where all components of a predicted triplet must exactly match the corresponding gold standard triplet to be counted as correct. For the development set, evaluation was performed locally using the official evaluation script (`eval_timeline.py`) provided by the shared task organizers<sup>4</sup>. Since all pipelines followed 2 major steps as described in Section 2, we additionally calculated note-level micro precision, recall, and F1 as intermediate metrics for LLM extraction performance. Evaluation on the test set was performed by

the task organizer and announced using the leaderboard on the shared task website<sup>1</sup>.

Two types of patient-level F1 scores were calculated by the official evaluation script: Type A, which includes all patients regardless of whether they have gold-standard timelines, and Type B, which includes only patients with confirmed chemotherapy timelines (Yao et al., 2025). The official score is the average of the Type A and Type B F1 scores, where each patient’s score is computed individually and then averaged across patients.

## 6 Results

### 6.1 Development Performance

Table 1 shows both note-level and timeline-level evaluation results of each method-model combination. Major findings are as follows.

First, better note-level performance is generally associated with, but does not necessarily indicate, better final performance on the timeline level, which highlights **the crucial role of normalization and aggregation**. For example, under the thinking approach, Qwen3-30B-A3B achieved much better micro precision, recall, and F1 than Qwen3-14B and Qwen3-32B, but their official scores were almost the same. This could be attributed to the duplication of repeated events from different notes during event aggregation, and time expressions that Timenorm was unable to handle. Another noteworthy case is Qwen3-14B’s exceptionally high official score under the baseline approach. Similar to other models, its note-level extractions contain a substantial proportion of irregular time expressions (see Section 7 for examples), but those events were luckily discarded by Timenorm, leading to a high F1 score.

Second, **14B might be the best dense model size for this task**. Under both the baseline and thinking approaches, the 14B model outperformed other dense models on both the note level and the patient level. In light of this, we applied dictionary-based and fine-tuning-based methods to models up to 14B. As expected, Qwen3-14B consistently outperformed its 8B and 4B siblings.

Third, **fine-tuning a dense model reliably yielded the largest performance gain, while a thinking mixture-of-experts model performed comparably**. After introducing rule-based post-processing (described in Section 3), Qwen3-30B-A3B’s official F1 score on the development set improved from 0.596 to 0.625 (Table 2). We at-

tribute most of the performance gain of thinking to the self-checking behavior exhibited by the CoT, which improved both the precision and recall of note-level event extraction. For details and examples, please see Section 7.

Fourth, **the dictionary-enhanced method achieved the highest recall among all systems, and incorporating LLM verification further improved precision by filtering out false positives**. We examined the intermediate results of LLM verification on top of dictionary-based sentence tagging. As shown in Table S2 in Appendix C, dictionary tagging alone achieved nearly perfect recall across cancer types, but precision was lower. Adding LLM verification consistently increased precision for these sentences (e.g., breast: 0.732→0.824; melanoma: 0.802→0.830) while keeping recall near 1.0 (a small trade-off appears for ovarian, 1.000→0.994, with F1 unchanged). Overall, the dictionary-based pipeline attains the best recall among all systems, and the second-best development performance, trailing only SFT.

### 6.2 Test Performance

On the test set, our SFT approach (submission 1) attained the highest overall average score of 0.678. The SFT + DPO model (submission 2) closely followed with an average of 0.666. The thinking approach with postprocessing (submission 4) also performed competitively, reaching an average score of 0.644. The ensemble method (submission 5), which combined SFT, SFT + DPO, and thinking achieved an overall score of 0.603, which was lower than any of the individual model. This suggests that differences in error patterns limited the benefit of ensembling. The dictionary-enhanced sentence-level extraction (submission 3) produced weaker results, with an overall score of only 0.545, suggesting potential limitations in the term coverage of our SACT dictionary with respect to what appears in the test set. Together, test results again indicate that fine-tuning-based methods were the most effective in our experiments, while LLM thinking was also a competitive approach.

## 7 Error Analysis

We empirically investigated noteworthy errors made by our systems on the development set, aiming to inform both future system development and potential refinements to the challenge design in subsequent rounds. Gold-standard timelines of the

Method	Model	Note-Level Micro			Timeline-Level Macro F1		
		Precis.	Recall	F1	Type A	Type B	Official
Prompting Baseline	Qwen3-4B	.039	.278	.069	.173	.082	.127
	Qwen3-8B	.040	.283	.070	.060	.103	.082
	Qwen3-14B	.139	.276	<b>.185</b>	.466	.370	<b>.418</b>
	Qwen3-32B	.103	.209	.138	.253	.220	.236
	Qwen3-30B-A3B	.068	.243	.106	.104	.178	.141
	MedGemma-27B	.085	.439	.142	.158	.199	.178
Thinking	Qwen3-4B	.338	.382	.358	.471	.378	.424
	Qwen3-8B	.355	.335	.345	.614	.410	.512
	Qwen3-14B	.517	.346	.415	.676	.515	.595
	Qwen3-32B	.355	.325	.339	.623	.568	<b>.596</b>
	Qwen3-30B-A3B	.600	.468	<b>.526</b>	.676	.516	<b>.596</b>
Dictionary + Sentence-level	Qwen3-8B	.294	.509	.372	.689	.468	.578
	Qwen3-14B	.434	.657	<b>.522</b>	.729	.536	<b>.632</b>
SFT	Qwen3-4B	.379	.507	.434	.651	.473	.562
	Qwen3-8B	.419	.569	<b>.483</b>	.650	.542	.596
	Qwen3-14B	.397	.615	<b>.483</b>	.711	.577	<b>.644</b>
DPO + SFT	Qwen3-4B	.390	.483	.431	.670	.435	.553
	Qwen3-8B	.409	.574	.478	.651	.545	.598
	Qwen3-14B	.401	.620	<b>.487</b>	.695	.549	<b>.622</b>

Table 1: Development set performance. The best official score under each method is highlighted.

Submission #	Method	Dev Official	Test Official
Submission 1	SFT	<b>.644</b>	<b>.678</b>
Submission 2	SFT + DPO	.622	.666
Submission 3	Dictionary + Sentence-level	.632	.545
Submission 4	Thinking + Postprocessing	.625	.644
Submission 5	Ensemble of 1, 2, and 4	.562	.603

Table 2: Development and test set performance of final submissions. The best scores are highlighted.

test set are held private by the organizer to enable future versions of the shared task, hence we are unable to perform error analysis on the test set.

### 7.1 Errors in Prompt-Based Extraction

Under the prompting baseline, LLMs often extract medications and procedures that are not part of SACT (e.g., "Neupogen", "ProHance", "MRI") and irregular time expressions (e.g., "04/26/2012 at 12:13 PM", "4 cycles", "midway through chemo"), even when explicitly instructed not to. They also produce errors related to text span boundaries and formatting. This explains the overly low precision and recall values on the note level.

When thinking mode is enabled, we observed that Qwen3 models would spontaneously check whether each candidate event belonged to SACT (e.g., "Neulasta is a G-CSF, not an SACT, so it's excluded") and whether its associated time ex-

pressions satisfied the extraction instructions (e.g., "'status post' refers to something that happened in the past but doesn't give an exact time"), which significantly reduced false positives. This double-checking behavior also helped decrease false negatives (e.g., "Double-checking to make sure I didn't miss any hidden events. Maybe ...").

Most remaining errors produced by the thinking models were commonly encountered by other methods, and are discussed in the following sections. A noteworthy category is incorrect inference caused by ambiguous language or formatting in clinical notes. For example, consider a note containing a table of medications administered on a given date with the SACT note "TRASTUZUMAB" followed by "None Entered"; based on the column names in the context, this means that the start or end date of the therapy is not entered. However, Qwen3-30B-A3B interpreted this to mean that even though the

date was present, it should be excluded because the time was missing. We also observed confusion about whether to include scheduled events. Our manual inspection of the gold standard timelines revealed that scheduled events were inconsistently annotated. In addition, a large proportion of clinical notes in the dataset contained no gold annotation, and any LLM-extracted events from these notes would become false positives.

## 7.2 Errors in Dictionary-Based Extraction

The dictionary-enhanced extraction approach provided strong coverage of chemotherapy mentions but also revealed several important limitations. First, the main limitation came from false positives, which lowered overall precision. Because the method matched any token found in the dictionary, it might incorrectly identify unrelated terms as chemotherapy mentions. For example, the system recognized the word "FEC" in "Normal FEV1 and FEV1/FEC ratio" as the regimen consisting of Fluorouracil, Epirubicin, and Cyclophosphamide, although it was in fact a typographical error for Forced Vital Capacity (FVC).

Second, the system also suffered from false negatives when encountering typographical errors or abbreviations not present in the dictionary. For instance, the test set included terms such as "bev" for Bevacizumab and "interfuron" for interferon. These variants were not captured, leading to missed extractions. This limitation helps explain the performance gap between the development set, where dictionary coverage was stronger, and the test set, where more novel variants appeared.

Finally, we observed interesting cases of internal inconsistency between the model's intermediate reasoning and its final output. For instance, in the thinking process, the model may explicitly state that a tag such as `<e>tc</e>` should be removed, but in the final output the tag still appears (see Appendix D for an example). This mismatch suggests that controlling the alignment between reasoning and output remains a challenge for dictionary-enhanced extraction with LLMs. It also points the way to potentially useful future work in explainable AI to use reasoning traces to better understand how LLMs understand complex clinical notes.

## 7.3 Errors in Training-Based Methods

As we pooled all available annotations into the same training set, the performance of our training-based model is sensitive to imbalances in the train-

ing data. For instance, the model often defaults to the generic CONTAINS-1 relation, misclassifying more specific BEGINS-ON and ENDS-ON relations. This tendency reflects a class imbalance where CONTAINS-1 instances are overrepresented in the training set. Furthermore, we observed a notable performance degradation on melanoma notes compared to breast and ovarian cancer, which manifests as low precision. This may be a consequence of data skew, as our training set contains significantly fewer melanoma notes, potentially leading the model to overfit to the majority of cancer types.

## 7.4 Errors in Normalization and Aggregation

The normalization process relies heavily on heuristic rules in the Timenorm pipeline, which can both improve alignment with the gold standard and introduce systematic errors. In general, using Timenorm facilitates consistent normalization of relative time expressions, but we found instances where the output diverged from expected interpretations. For example, the expression "last week" relative to a document time of 2013-01-15 is normalized to "2013-01-08" in full date format (YYYY-MM-DD), whereas "next week" relative to 2013-07-23 is normalized to "2013-w31" in week format (YYYY-w##). Likewise, expressions that specify only month and day can be incorrectly anchored to the previous year. If the document time is 2013-02-10, the expression "January 9" is normalized as "2012-01-09" rather than the correct "2013-01-09". Such inconsistencies suggest that while Timenorm is powerful, it may require task-specific adjustments to handle edge cases in clinical timelines.

In aggregation, the lack of entity consolidation introduces redundancy and inconsistency across patient timelines. For example, the same chemotherapy drug can appear under slightly different surface forms, such as "il2", "il-2", and "interleukin-2", all linked to the same date and relation. Similarly, regimen-level mentions can coexist with individual drug mentions. The gold standard may annotate both "AC-T" as a regimen and its components Adriamycin (A), Cyclophosphamide (C), and Taxol (T), leading to multiple overlapping entries.

A further source of discrepancy arises from how start and end events are aligned within the same timeline. When both BEGINS-ON and ENDS-ON relations are identified for the same drug on the same date, our system retains both events for completeness, whereas the gold timelines may arbitrarily keep only one. For example, in the

gold timeline, Cabotaxol and Taxol are annotated as [cabotaxol, BEGINS-ON, 2012-01-12] and [taxol, ENDS-ON, 2011-12-15]. The complete representation, however, should include both start and end events for each drug, i.e., 4 events in total.

## 8 Conclusions and Discussion

Extracting clinical events from unstructured notes has always been a challenging task (Olex and McInnes, 2021). Under the ChemoTimelines 2025 shared task framework, our work explores several approaches based on modern and emerging model training and inference techniques. Major findings are as follows:

1. The aggregation of note-level events into patient-level timelines is crucial for the final performance of a system.
2. Fine-tuning a dense model, especially of size 14B, reliably yielded the largest performance gain, while a thinking mixture-of-experts model performed comparably.
3. The dictionary-enhanced method achieved the best recall, while LLM verification improved precision by reducing false positives.

We found that the dictionary-based approach offered both efficiency and interpretability, while still maintaining acceptable performance despite some information loss at the sentence level. Instead of reasoning over all notes in the development set, the method reduced the burden by restricting LLM verification to a much smaller number of candidate sentences flagged by the dictionary, plus context-enhanced sentences for relation extraction. This substantially lowered input token volume, reasoning time, and computational cost. Although focusing on sentences inevitably sacrifices some contextual information compared with note-level extraction, the resulting performance remained strong, supported by very high recall from dictionary tagging and improved precision from LLM verification. Moreover, the transparent matching rules enhance interpretability and facilitate systematic refinements, such as synonym expansion or ontology-based extensions. Together, these features make the dictionary-based pipeline a lightweight, resource-efficient, and interpretable complement to learning-based systems.

The comparable performance of the training-free LLM thinking approach and fine-tuning-based methods suggests a potential cost-effectiveness trade-off for this specific task. Once fine-tuning is

supported by a sufficient amount of high-quality data, it is capable of yielding a trustworthy performance gain while maintaining the speed of direct output generation. In contrast, CoT thinking, as a core component of recent inference-time scaling techniques for LLMs, is characterized by its higher latency at test time, despite that it requires much less data annotation in the development phase. Given the substantial performance gain from LLM thinking in our experiments (and potential benefits to explainability), we recommend that future attempts on similar tasks consider including it as a baseline method, especially in consideration of the high cost of EHR annotation by human experts.

The ensemble method did not lead to performance gains. Instead, the overall score was lower than any of the individual models. This suggests that errors from different systems tend to accumulate when combined, and these mistakes cannot be effectively corrected through the normalization and aggregation pipelines. As a result, simple ensembling is not a viable strategy for this task.

We conceived several other methods that were not implemented due to time and resource constraints, and we hope providing them here may benefit clinical timeline extraction. First, our current prompt-based approaches (both non-thinking and thinking) utilized static ad hoc in-context examples. Including dynamically-retrieved training examples related to test time queries has the potential to further improve performance. Second, current training methods do not have an explicit reasoning process before generating the extractions. Future methods may synthesize reasoning data through rejection sampling (Yuan et al., 2023) or apply reinforcement learning (DeepSeek-AI et al., 2025) for better performance.

The chemotherapy events in the training data represented the specific EHR documentation style of the source facilities and systems. The more general task of extracting clinical event timelines may involve a diversity of local documentation styles, event sources (e.g., treatment, laboratory, billing, etc.), and levels of standardization.

Additional insights into the capabilities and weaknesses of various LLM-based strategies might be obtained with introspection into performance against specific evaluation data examples, additional layers of case review with expert oncologists, and testing with the newest generation of emerging LLMs.

## Limitations

Our methods were highly customized to the Chemo-Timelines challenge, hence our findings may not generalize well to other clinical extraction tasks. Due to time, resource, and privacy constraints, we did not assess a full range of contemporary open- and closed-source LLMs (e.g., larger Qwen3 models, the Llama series, GPT series, etc.), therefore our findings may not generalize. MedGemma was also the only medicine-specialized LLM included in our experiments. Although a general-purpose LLM combined with a tailored aggregation pipeline was sufficient for this task, future work may benefit from models more familiar with clinical notes. For technical methods that we conceptualized but did not have a chance to implement and test, please refer to Section 8.

## References

- 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.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. [Sft memorizes, rl generalizes: A comparative study of foundation model post-training](#). *Preprint*, arXiv:2501.17161.
- 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.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Maryam Gholipour, Reza Khajouei, Parastoo Amiri, Sadrieh Hajesmaeel Gohari, and Leila Ahmadian. 2023. Extracting cancer concepts from clinical notes using natural language processing: a systematic review. *BMC bioinformatics*, 24(1):405.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Shohreh Haddadan, Tuan-Dung Le, Thanh Duong, and Thanh 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 Natural Language Processing Workshop*, pages 382–393, Mexico City, Mexico. Association for Computational Linguistics.
- Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. [REBEL: Relation extraction by end-to-end language generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). *Preprint*, arXiv:1910.13461.
- Amy L. Olex and Bridget T. McInnes. 2021. [Review of temporal reasoning in the clinical domain for timeline extraction: Where we are and where we need to be](#). *Journal of Biomedical Informatics*, 118(103784).
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. [Direct preference optimization: Your language model is secretly a reward model](#). *Preprint*, arXiv:2305.18290.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. [Medgemma technical report](#). *arXiv preprint arXiv:2507.05201*.
- Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Temporal reasoning over clinical text: the state of the art. *Journal of the American Medical Informatics Association*, 20(5):814–819.
- Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2015. Normalization of relative and incomplete temporal expressions in clinical narratives. *Journal of the American Medical Informatics Association*, 22(5):1001–1008.
- 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 Natural Language*



- Processing Workshop*, pages 417–421, Mexico City, Mexico. Association for Computational Linguistics.
- Songjun Tu, Jiahao Lin, Xiangyu Tian, Qichao Zhang, Linjing Li, Yuqian Fu, Nan Xu, Wei He, Xiangyuan Lan, Dongmei Jiang, and Dongbin Zhao. 2025. [Enhancing llm reasoning with iterative dpo: A comprehensive empirical investigation](#). *Preprint*, arXiv:2503.12854.
- Liwei Wang, Qiuhaio Lu, Rui Li, Sunyang Fu, and Hongfang Liu. 2024. [Wonder at chemotimelines 2024: MedTimeline: An end-to-end NLP system for timeline extraction from clinical narratives](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 483–487, Mexico City, Mexico. Association for Computational Linguistics.
- Liwei Wang, Jason Wampfler, Angela Dispenzieri, Hua Xu, Ping Yang, and Hongfang Liu. 2020. Achievability to extract specific date information for cancer research. In *AMIA Annual Symposium Proceedings*, volume 2019, page 893.
- Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. 2025. [Reinforcement learning for reasoning in large language models with one training example](#). *Preprint*, arXiv:2504.20571.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- 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.
- Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. 2025. Overview of the 2025 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. [Scaling relationship on learning mathematical reasoning with large language models](#). *Preprint*, arXiv:2308.01825.
- Hanning Zhang, Jiarui Yao, Chenlu Ye, Wei Xiong, and Tong Zhang. 2025. [Online-dpo-r1: Unlocking effective reasoning without the ppo overhead](#). *Notion Blog*.
- Jeff Zhang, Yin Aphinyanaphongs, and Anthony Cardillo. 2024. [NYULangone at chemotimelines 2024: Utilizing open-weights large language models for chemotherapy event extraction](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 428–430, Mexico City, Mexico. Association for Computational Linguistics.

## A Prompt Templates

### A.1 Prompt for Baseline and Thinking

You are an experienced medical annotator tasked with extracting systemic anticancer therapy (SACT) events from a given clinical note.

#### WHAT A SACT IS:

SACT encompasses medications used in traditional cytotoxic chemotherapy, endocrine therapy, targeted therapy, and immunotherapy. SACT may appear in generic names (e.g., Anastrozole), brand names (e.g., Arimidex), or combined names (e.g., TCH). Non-specific SACT mentions such as "chemotherapy" or "chemo" should also be included. Exclude therapies, medications, and diagnostic procedures used not for anticancer purposes, such as dietary supplements and biopsies. Exclude therapy candidates that you don't know what they are or aren't sure if they are SACT.

#### WHAT YOU SHOULD EXTRACT:

You should only extract SACT events that are explicitly associated with specific time expressions. Here are some format examples of time expressions you are expected to extract: "December 27, 2011", "May 21st, 2013", "7/20/2012", "today", "3 weeks ago", "1 year". Ignore nonspecific time mentions such as cycle or dose numbers (e.g., "cycle 1 of 6" and "1/6 dose") and ambiguous relative time (e.g., "midway through" or "at the same time as" another event). Exclude SACT mentions without an associated time.

Additionally, for each event, select a relation label from BEGINS-ON, ENDS-ON, and CONTAINS-1 to indicate the relation between the SACT and its time based on the note's language. CONTAINS-1 means the SACT happened at a specific time; if the note explicitly mentions the start or end of an event, use BEGINS-ON or ENDS-ON. For example, if the note says "She

received Herceptin on May 21st, 2013", your extracted event will be "Herceptin", "CONTAINS-1", "May 21st, 2013"; if the note says "Start ipilimumab on today's date", your extracted event will be "ipilimumab", "BEGINS-ON", "today".

If a SACT is associated with multiple time points, extract them as separate events. For example, if the note says "Herceptin was initiated on 12/27/2011 and completed on April 10, 2012", you should extract two events: one is "Herceptin", "BEGINS-ON", "12/27/2011", and the other is "Herceptin", "ENDS-ON", "April 10, 2012". Similarly, if multiple SACTs are associated with the same date, you should also extract them as separate events.

#### HOW YOU SHOULD FORMAT YOUR RESPONSE:

SACT names and their associated time expressions should be kept exactly as they appear, even if there is a typo. Do not alter them, normalize the time expression, or infer the exact date. For example, if a SACT event appears in the note as "Alibercept received yesterdat", your extracted event will be "Alibercept", "CONTAINS-1", "yesterdat".

Do not combine SACT mentions that refer to the same therapy but appear in different names, even if one appears in parentheses as the alternative name for another; extract them separately. For example, if there are three SACT events, "il-2", "il2", and "interleukin-2", and all have corresponding time expressions, treat them as three separate SACT events; if the note says "TRASTUZUMAB (HERCEPTIN) received today", you should extract two events: one is "TRASTUZUMAB", "CONTAINS-1", "today", and the other is "HERCEPTIN", "CONTAINS-1", "today".

If multiple SACTs are administered together, treat them as separate events. For example, for "Doxorubicin/Cyclophosphamide" you should extract 2 events, one for Doxorubicin and the other for Cyclophosphamide. However, if a SACT name is already a combined treatment name (e.g., TCH), treat it as a single event.

Ignore supplementary descriptors of SACT names, such as dose (e.g., "high dose") and administration method (e.g., "IV"). For non-

specific SACT like "chemotherapy" or "chemo", ignore their descriptors, such as "adjuvant" and "neoadjuvant".

Your response must be a JSON array under the following schema:

```
{
  "type": "array",
  "description": "An array of SCAT events
  extracted from the clinical note.",
  "items": {
    "type": "object",
    "properties": {
      "SACT": {
        "description": "A SACT name
        extracted as it is.",
        "type": "string"
      },
      "relation": {
        "description": "The relation
        between the SACT and its associated time
        expression. Must be one of BEGINS-ON,
        ENDS-ON, and CONTAINS-1.",
        "type": "string"
      },
      "time": {
        "description": "The time expression
        associated with the SACT, extracted as it
        is.",
        "type": "string"
      }
    },
    "required": ["SACT", "relation",
    "time"],
  }
}
```

If there is no SACT event in the clinical note, return an empty array.

Now, extract SACT events from the following clinical note:

{note}

#### A.2 Prompt for LLM-based Chemotherapy Tag Verification

You are an experienced medical annotator tasked with verifying and extracting systemic anticancer therapy (SACT) mentions from a given clinical note. Some SACT candidates have already been tagged using a dictionary-based method.

SACT encompasses medications used in tra-

ditional cytotoxic chemotherapy, endocrine therapy, targeted therapy, and immunotherapy. An SACT mention may appear as a generic name (e.g., Anastrozole), a brand name (e.g., Arimidex), or a combined name (e.g., TCH). Non-specific chemotherapy-related SACT mentions like "chemotherapy", "chemo", "chemotherapy's", etc. And even mentions with typos like "chemotheray" should also be retained.

Your task is twofold:

1. Review the pre-tagged mentions ONE BY ONE and remove any incorrect tags caused by dictionary false positives.

2. Identify and tag any additional SACT mentions that are missing due to typos or uncommon abbreviations not found in the dictionary.

Extract each SACT mention exactly as it appears in the note, even if there is a typo; do not alter or normalize it. For example, if an SACT Afibercept appears in the note as "Alibercept", your extracted SACT should be "Alibercept". Do not combine SACT mentions that refer to the same therapy but appear in different forms; extract them as separate mentions. For example, if there are three mentions: "il-2", "il2", and "interleukin-2", extract them all separately.

Ignore supplementary information such as dose, administration method, or diagnostic/therapeutic context not related to anticancer treatment. Exclude therapies, medications, or procedures used for non-cancer purposes, such as dietary supplements or biopsies.

You should remove or add tags in the raw text. Do not output any other text.

Both the input and the output should be put in ". Please strictly follow the format of the output. You MUST only wrap the correct SACT mentions with <e> and </e> tags in your outputs. Do not add any other tags or quote marks.

For example, given the input text:

"This is a sentence with both <e>correct SACT</e> and <e>wrong SACT</e> mentioned."

The expected output is:

"This is a sentence with both <e>correct SACT</e> and wrong SACT mentioned."

Sometimes the input can be extremely long, like:

"====Here are some background details about the patient===="

This is a sentence with both <e>correct SACT1</e> and <e>wrong SACT</e> mentioned, and another <e>correct SACT2</e> mentioned."

The expected output is:

"====Here are some background details about the patient===="

This is a sentence with both <e>correct SACT1</e> and wrong SACT mentioned, and another <e>correct SACT2</e> mentioned."

Now, extract SACT events from the following sentences in a clinical note:

### A.3 Prompt for Context-Enhanced Sentence-Level Relation Extraction

You are an experienced medical annotator tasked with extracting systemic anticancer therapy (SACT) events from a given clinical note.

**WHAT A SACT IS:** SACT encompasses medications used in traditional cytotoxic chemotherapy, endocrine therapy, targeted therapy, and immunotherapy. SACT may appear in generic names (e.g., Anastrozole), brand names (e.g., Arimidex), or combined names (e.g., TCH). Non-specific SACT mentions such as "chemotherapy" or "chemo" should also be included. I have extracted all the SACT events for you between the tags <e> and </e> in my input, so you don't need to extract SACT yourself.

**WHAT YOU SHOULD EXTRACT:** You should do this step by step. First, identify all the SACT events between <e> and </e> in my input and ONLY focus on these SACT events. Then, exclude the SACT events if they are macro information rather than patient-specific information, or if they are negations of SACTs. Next, for each valid SACT event, extract specific time expressions that are explicitly associated with that SACT event. Here are some format examples of time expressions you are expected to extract: "December 27, 2011", "May 21st, 2013", "7/20/2012", "today", "3 weeks ago", "1 year". Ignore nonspecific time mentions such as cycle or dose numbers (e.g., "cycle 1 of 6" and "1/6 dose") and ambiguous relative time (e.g., "midway through" or "at the same time as" another event). Exclude time expressions that are not associated with the current SACT event. If there is not a time expression related to the current SACT event, then skip it and check the next SACT event.

Additionally, for each event, select a relation label from BEGINS-ON, ENDS-ON, and CONTAINS-1 to indicate the relation between the SACT and its time based on the note's language. CONTAINS-1 means the SACT happened at a specific time; if the note explicitly mentions the start or end of an event, use BEGINS-ON or ENDS-ON. For example, if the note says "She received Her-

ceptin on May 21st, 2013", your extracted event will be "Herceptin", "CONTAINS-1", "May 21st, 2013"; if the note says "Start ipilimumab on today's date", your extracted event will be "ipilimumab", "BEGINS-ON", "today".

If a SACT is associated with multiple time points, extract them as separate events. For example, if the note says "Herceptin was initiated on 12/27/2011 and completed on April 10, 2012", you should extract two events: one is "Herceptin", "BEGINS-ON", "12/27/2011", and the other is "Herceptin", "ENDS-ON", "April 10, 2012". Similarly, if multiple SACTs are associated with the same date, you should also extract them as separate events.

**HOW YOU SHOULD FORMAT YOUR RESPONSE:** SACT names and their associated time expressions should be kept exactly as they appear, even if there is a typo. Do not alter them, normalize the time expression, or infer the exact date. For example, if a SACT event appears in the note as "Alibercept received yesterdat", your extracted event will be "Alibercept", "CONTAINS-1", "yesterdat".

Your response **MUST** be in a json format under the following schema:

```
hn [{"SACT event1", "relation1", "time expression1"}, {"SACT event2", "relation2", "time expression2"}, ...
```

## B Chemotherapy Events Dictionary

### B.1 Breast Cancer

5-fu  
a-cmf  
a.c  
a/c  
abemaciclib  
abraxane  
ac  
ac-cmf  
ac-d  
ac-h  
ac-t  
ac-th  
ac-thl  
ac-thp  
ach  
act  
adriamycin  
afc  
afinitor  
airuika  
alimta  
alpelisib  
anastrozole  
anthracycline  
arimedex  
arimidex

aromasin  
aromatase inhibitor  
at-cmf  
atc  
atezolizumab  
avastin  
bev  
bevacizumab  
bilateral oophorectomy  
caf  
camrelizumab  
capecitabine  
capivasertib  
carbo  
carboplatin  
cbd  
cef  
cef-t  
chemo  
chemo therapy  
chemo-rt  
chemoembolization  
chemorad  
chemort  
chemotherap  
chemotherapeutic  
chemotherapeutic  
chemotherapies  
chemotherapy  
chemotherapy's  
chemotheray  
chidamide  
cisplatin  
cmf  
cmf-e  
cmf-h  
cmft  
cnp  
cp-ac  
cp-ddac  
cp-ec  
cvb  
cyclophosphamide  
cytoxan  
d-ac  
d-ac+bev  
d-ec  
d-fec  
d-fec+bev  
datopotamab deruxtecan  
datroway  
dcb  
dda-ddt-ddc  
ddac  
ddac-ddt  
ddac-ddth  
ddac-pacph  
ddac-t  
ddac-th  
ddac-thp  
ddat  
dde  
dde-iddcmf  
ddec-ddcmf  
ddec-ddd  
ddec-ddt  
ddec-t  
ddec-th  
ddec-thp  
ddfec-d

ddp  
ddt  
ddt-ddec  
ddt-ec  
ddth  
docetaxel  
docetaxol  
doxil  
doxorubicin  
e-cmf  
e-d  
e-x  
ec-cmf  
ec-d  
ec-ddt  
ec-dt  
ec-h  
ec-p  
ec-t  
ec-th  
ec-thp  
ecd-gc  
ech  
ech-th  
edc  
ehp  
elacestrant  
ellence  
endocrine therapy  
enhertu  
enzalutamide  
ep-ddcmf  
epidaza  
epirubicin  
eribulin  
everolimus  
exemestane  
fac  
fac-t  
fac-th  
fac-thp  
fareston  
faslodex  
fec  
fec-d  
fec-h  
fec-p  
fec-t  
fec-th  
fec-thp  
femara  
fluorouracil  
fulvestrant  
gcb  
gdoc  
gemcitabine  
gemzar  
ghp  
gnrh analogs  
goserelin  
h+d  
halaven  
herceptin  
herceptin hylecta  
ibrance  
idd-etc  
iddenpc  
iddepc  
inavolisib  
irene

itovebi  
ixabepilone  
ixempra  
javlor  
kadcyla  
keytruda  
kisqali  
l+t  
lapatinib  
letrozole  
leuprolide  
loqtorzi  
lupron  
lynparza  
margenza  
margetuximab  
methotrexate  
millipred  
mitomycin  
mitoxantrone  
mmm  
mtx  
mutamycin  
myocet  
navelbine  
neratinib  
nerlynx  
nolvadex  
novantrone  
np-ddac  
np-ddec  
np-ec  
npc-ddec  
npld  
ofs  
olaparib  
orserdu  
ovarian irradiation  
paclitaxel  
palbociclib  
paraplatin  
pcb  
pembrolizumab  
pemetrexed  
perjeta  
pertuzumab  
phesgo  
piqray  
platinol  
platinum  
pld  
prednisolone  
pyrotinib  
q2wk  
ribociclib  
s-1  
sacituzumab govitecan  
t-ac  
t-cef  
t-ddac  
t-ddec  
t-dm1  
t-ec  
t-fac  
t-fec  
t-h  
t-t  
tac  
talazoparib  
talzenna

tamoxifen  
taxane  
taxol  
taxotere  
taxotere  
tc  
tc-h  
tcbh  
tch  
tchp  
tcyh  
tecentriq  
th-ac  
th-ddac  
th-ech  
th-fec  
thl  
thp  
toremifene  
toripalimab  
tpc  
trastuzumab  
trastuzumab deruxtecan  
trastuzumab emtansine  
trelstar la  
triptorelin  
trodelvy  
truqap  
tucatinib  
tukysa  
tx-cex  
tykerb  
v-fec  
verzenio  
vh-fec  
vhp  
vinflunine  
vinorelbine  
xeloda  
xhp  
xtandi  
zoladex

## B.2 Melanoma

abc  
abraxane  
afibercept  
aflibercept  
alfa-2b interferon  
alflibercept  
alibercept  
alpha 2b interferon  
alpha interferon  
alpha-2b interferon  
alpha-2b interferon  
alpha-2binterferon  
atezolizumab  
avastin  
bevacizumab  
binimetinib  
braftovi  
carboplatin  
chemo  
chemo therapy  
chemo-rt  
chemorad  
chemoradiation  
chemort  
chemotherapy  
chemotherapeutic

chemotherapeutic  
chemotherapies  
chemotherapy  
chemotherapy's  
chemotheray  
cisplatin  
cnp  
cobimetinib  
complete resection  
contego  
cotellic  
cpb  
cvd  
dabrafenib  
dacarbazine  
docetaxel  
dtic  
eldisine  
encorafenib  
fotemustine  
gleevec  
hepzato kit  
il 2  
il-2  
il2  
imatinib  
imlygic  
inteferon  
interferon  
interferon  
interleukin  
interleukin 2  
interleukin-2  
ipilimumab  
keytruda  
kimmtrak  
kolupin  
koselugo  
leukine  
lifileucel  
loqtorzi  
mekinist  
mektovi  
melphalan  
methotrexate  
muphoran  
nivolumab  
opdivo  
opdualag  
paclitaxel  
paraplatin  
pembrolizumab  
platinol  
proleukin  
sargramostim  
selumetinib  
tace  
tafinlar  
talimogene laherparepvec  
tasisulam  
taxol  
taxotere  
tebentafusp  
tecentriq  
temodar  
temozolomide  
tils  
toripalimab  
trametinib  
tunlametinib

vaccinia  
vaccinia virus  
vaccinia virus  
vemurafenib  
vindesine  
yervoy  
zelboraf

### B.3 Ovarian Cancer

abraxane  
alimta  
ataxol  
avastin  
avstin  
bevacizumab  
caboplatin  
cabotaxol  
carbo  
carboplat  
carboplatin  
carbotaxol  
chemo  
chemo therapy  
chemo-rt  
chemoembolization  
chemorad  
chemort  
chemotherap  
chemotherapeutic  
chemotherapeutic  
chemotherapies  
chemotherapy  
chemotherapy's  
chemotherapy  
chmeo  
cisplatin  
cistoplatin  
cytoreductive surgery  
dcb  
docetaxel  
docetaxil  
doxil  
doxorubicin  
elahere  
etoposide  
femara  
gcb  
gemcitabine  
gemzar  
hycamtin  
ihcp  
intraperitoneal hyperthermicchemoperfusion  
intraperitonealhyperthermic chemoperfusion  
koselugo  
letrozole  
liposomal doxorubicin  
lynparza  
mekinist  
mirvetuximab soravtansine  
navelbine  
nintedanib  
niraparib  
nolvadex  
olaparib  
ovastat  
paciltaxel  
paclitaxel  
paclitaxela  
paraplatin  
paxil

pazopanib  
pemetrexed  
platinol  
platinum  
pldc  
rubraca  
rucaparib  
selumetinib  
t/c  
tamoxifen  
tax  
taxo  
taxol  
taxotere  
tc  
tc-bev  
topotecan  
trabectedin  
trametininib  
treosulfan  
vargatef  
vepesid  
vinorelbine  
votrient  
yondelis  
zejula

## C Supplementary Figures and Tables

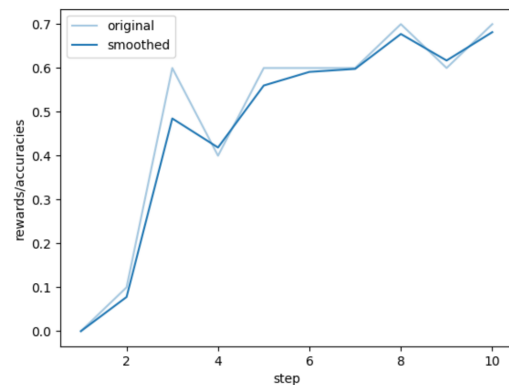


Figure 1: DPO reward accuracy curve of Qwen3-14B.

Cancer Type	# Total	# Annotated (%)
Breast	14,234	216 (1.52%)
Melanoma	9,279	523 (5.64%)
Ovarian	16,782	523 (3.12%)

Table S1: Proportion of sentences with annotations in the development set

## D Example of Inconsistency Between Reasoning and Output

### Model thinking trace:

<think> First, I need to check if "Tc" is a correct SACT. Wait, Tc-99m MDP is Technetium-99m

Cancer type	anno	Dictionary only			Dictionary + LLM verification		
		Precis.	Recall	F1	Precis.	Recall	F1
breast	0	0.9944	1.0000	0.9972	0.9967	1.0000	0.9984
breast	1	0.7322	1.0000	0.8454	0.8244	1.0000	0.9038
melanoma	0	0.9853	1.0000	0.9926	0.9878	0.9998	0.9937
melanoma	1	0.8021	1.0000	0.8902	0.8296	0.9962	0.9053
ovarian	0	0.9926	0.9998	0.9962	0.9927	0.9998	0.9963
ovarian	1	0.8112	1.0000	0.8958	0.8150	0.9943	0.8958

Table S2: Sentence-level tagging on the dev set: dictionary tagging vs. dictionary tagging with LLM verification. anno\_indicator=1 means the sentence has gold chemotherapy annotations.

Methane Diphosphate, which is a radiopharmaceutical used in bone scans. But SACT refers to systemic anticancer therapies like chemotherapy, endocrine therapy, etc. Radiopharmaceuticals are used for diagnostic purposes, not as part of cancer treatment. So "Tc" here is a false positive from the dictionary method. The user's instruction says to remove incorrect tags caused by false positives. **Therefore, I should remove the <e>Tc</e> tag.**  
</think>

**Model output:**

TECHNIQUE: Whole body bone scan was obtained following intravenous administration of 26.5mCi of <e>Tc</e>-99m MDP.

**Expected corrected output:**

TECHNIQUE: Whole body bone scan was obtained following intravenous administration of 26.5mCi of Tc-99m MDP.

This example highlights a misalignment between the model's reasoning and its generated output: although the reasoning correctly identifies "Tc" as a false positive (radiopharmaceutical rather than chemotherapy), the structured output still preserves the erroneous tag.



# MasonNLP at MEDIQA-OE 2025: Assessing Large Language Models for Structured Medical Order Extraction

A H M Rezaul Karim

George Mason University, VA, USA  
akarim9@gmu.edu

Özlem Uzuner

George Mason University, VA, USA  
ouzuner@gmu.edu

## Abstract

Medical order extraction is essential for structuring actionable clinical information, supporting decision-making, and enabling downstream applications such as documentation and workflow automation. Orders may be embedded in diverse sources, including electronic health records, discharge summaries, and multi-turn doctor–patient dialogues, and can span categories such as medications, laboratory tests, imaging studies, and follow-up actions. The MEDIQA-OE 2025 shared task focuses on extracting structured medical orders from extended conversational transcripts, requiring the identification of order type, description, reason, and provenance. We present the MasonNLP submission, which ranked 5<sup>th</sup> among 17 participating teams with 105 total submissions. Our approach uses a general-purpose, instruction-tuned LLaMA-4 17B model without domain-specific fine-tuning, guided by a single in-context example. This few-shot configuration achieved an average  $F_1$  score of 37.76, with notable improvements in reason and provenance accuracy. These results demonstrate that large, non-domain-specific LLMs, when paired with effective prompt engineering, can serve as strong, scalable baselines for specialized clinical NLP tasks.<sup>1</sup>

## 1 Introduction

Clinical free-text notes in electronic health records (EHRs) contain essential information such as diagnoses, medications, procedures, and treatment plans (Wang et al., 2018; Demner-Fushman et al., 2009). Extracting structured medical orders, including medications, labs, imaging, and procedures, from such unstructured text is critical for enabling downstream applications like decision support and Computerized Physician Order Entry (CPOE) (Sutton et al., 2020; Kuperman and Gibson, 2003).

<sup>1</sup>Implementation can be found here: <https://github.com/AHMRezaul/MEDIQA-OE-2025>

However, despite the adoption of CPOE systems, errors in order entry persist (Kinlay et al., 2021; Campbell et al., 2006), and medication mistakes often arise during care transitions (Vira et al., 2006). This highlights the need for reliable methods to extract structured medical orders from clinical documentation.

To support such downstream tasks and reduce error rates, clinical information extraction (IE) methods have been developed to automatically identify entities and relations from free-text EHRs (Uzuner et al., 2010; Hahn and Oleynik, 2020). These systems have enabled large-scale mining of clinical concepts for applications such as cohort identification, adverse event detection (ADE), and case surveillance (Sarmiento and Dernoncourt, 2016; Landolsi et al., 2023; Ford et al., 2016). Within this domain, *medical order extraction (MOE)* focuses specifically on identifying medical orders, such as medications, lab tests, or imaging, and structuring them into machine-readable formats (Xu et al., 2010). Automating this process can reduce transcription burden, enhance care quality, and minimize errors in clinical workflows.

The **MEDIQA-OE 2025 Shared Task on Medical Order Extraction (OE)** (Corbeil et al., 2025b) introduced a new benchmark to address this need. The task provides annotated multi-turn doctor–patient conversations and evaluates systems on their ability to extract structured medical orders, including medications, laboratory tests, imaging studies, and follow-up procedures, from conversational transcripts. In addition to identifying the order, systems must also extract the corresponding description and the reason or justification provided by the physician. This reflects real-world clinical documentation scenarios, where accurate interpretation of both the order and its rationale is essential.

In this paper, we describe our participation in the MEDIQA-OE task, which involves identifying and structuring various medical orders and their

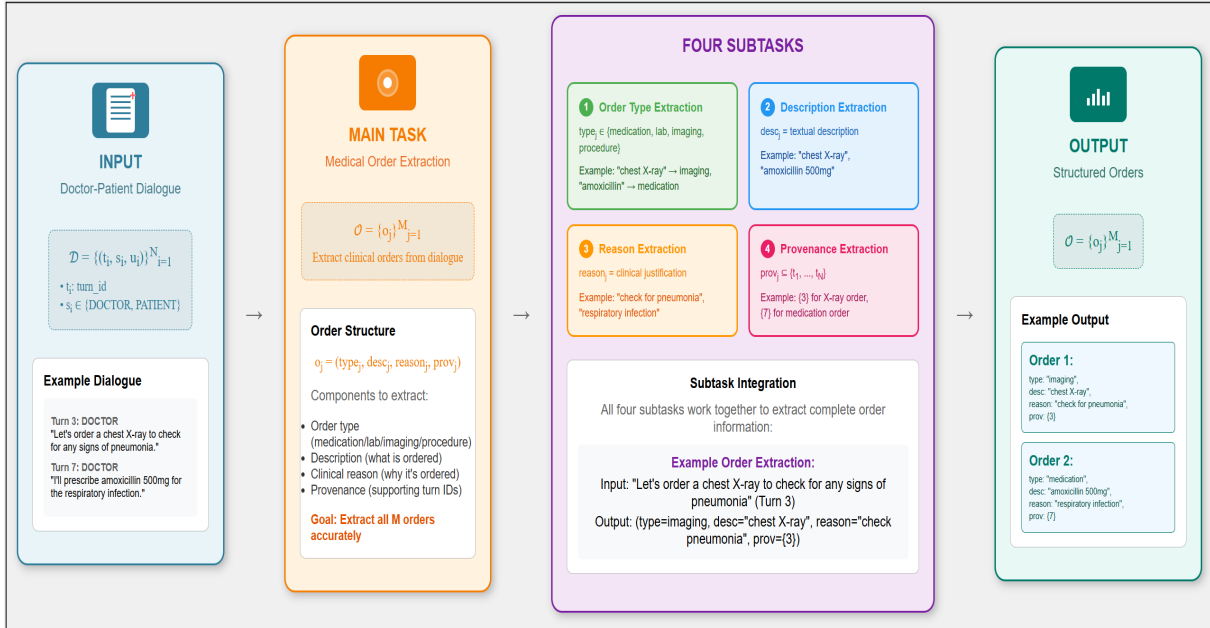


Figure 1: Overview of the MEDIQA-OE 2025 task: input transcripts, four subtasks (order type, description, reason, provenance), and the final structured output.

reasons with provenance grounding. Our approach uses **Meta’s LLaMA-4 Scout 17B** (Meta, 2025) model and relies on *few-shot prompt engineering*, without any domain-specific fine-tuning or external knowledge sources. We curated exemplar prompts that capture conversational structures and medical order patterns. This approach allows us to evaluate the capabilities of general-purpose LLMs on domain-specific extraction tasks.

Our contributions are as follows:

- We assess the feasibility and limitations of prompt-based approaches for structured information extraction in complex, safety-critical clinical domain tasks.
- We investigate the model’s reasoning capabilities by analyzing how well it can identify the clinical justification (reason) for prescribed medications, lab tests, imaging, and follow-ups.
- We evaluate the ability of a non-medical, instruction-tuned LLM to perform medical order extraction from clinical text without any domain-specific fine-tuning.

Our findings contribute to the growing body of work comparing general and domain-specific LLMs for clinical applications, highlighting prompting as a lightweight yet effective approach to structured prediction.

## 2 Related Work

The extraction of structured medical orders from unstructured clinical narratives has been a long-standing challenge in clinical Natural Language Processing (NLP), motivated by its potential to streamline clinical workflows, enhance decision support, and improve patient safety (Lussier et al., 2001; Patrick and Li, 2010; Uzuner et al., 2010, 2011). Early approaches were predominantly rule-based systems leveraging hand-crafted patterns, lexicons, and regular expressions to identify clinical entities and actions. Examples include systems built on platforms such as MedLEE (Friedman, 2000) and MetaMap (Aronson and Lang, 2010), which mapped text spans to controlled vocabularies like the UMLS (Bodenreider, 2004). These methods demonstrated high precision in restricted domains but suffered from limited transferability and scalability across institutions due to variations in clinical language and documentation styles.

The next generation of systems shifted toward statistical and machine learning approaches, which incorporated features from linguistic preprocessing (e.g., tokenization, POS tagging, dependency parsing) into classifiers such as Conditional Random Fields (CRFs) and Support Vector Machines (SVMs). Early examples in medication extraction, such as the 2009 i2b2 challenge systems (Patrick and Li, 2010), demonstrated improved adaptability over purely rule-based methods, though their re-

liance on manually engineered features still posed challenges for transferability.

With the advent of deep learning, feature engineering was largely replaced by distributed representations learned directly from [Gan et al.](#). Recurrent Neural Networks (RNNs), particularly LSTMs ([Hochreiter and Schmidhuber, 1997](#)) and BiLSTMs ([Schuster and Paliwal, 1997](#)), became popular for sequence labeling in clinical NLP, including medication extraction ([Jagannatha and Yu, 2016](#); [Huang et al., 2015](#); [Narayanan et al., 2022](#); [Christopoulou et al., 2020](#)). Attention mechanisms and hierarchical architectures further improved the capture of long-range dependencies, which is critical for modeling multi-turn dialogues and long EHR notes.

The introduction of transformer-based models ([Vaswani et al., 2017](#)) marked a significant leap in performance. Domain-specific transformers such as BioBERT ([Lee et al., 2020](#)), ClinicalBERT ([Alsentzer et al., 2019](#)), and BlueBERT ([Peng et al., 2019](#)) fine-tuned on biomedical corpora demonstrated substantial gains in extracting entities and relations from EHR data. These models leveraged self-attention to capture contextual relationships across long sequences, making them highly suitable for MOE from extended clinical narratives.

More recently, large language models (LLMs) such as GPT-3 ([Brown et al., 2020](#)), PaLM ([Chowdhery et al., 2023](#)), and LLaMA ([Touvron et al., 2023](#)) have shown strong zero- and few-shot capabilities across domains, including clinical tasks. While most LLM work in healthcare has focused on general summarization, question answering, and entity recognition ([Singhal et al., 2023](#); [Moor et al., 2023](#)), some studies have explored their application to structured MOE ([Yang et al., 2020](#); [Peng et al., 2023](#); [Mahajan et al., 2023](#)). These approaches typically involve prompt engineering, in-context learning, or retrieval-augmented generation (RAG), sometimes without any domain-specific fine-tuning, to leverage LLMs’ reasoning and language understanding abilities.

[Ford et al.](#); [Spasic and Nenadic](#); [Grouin et al.](#) highlight a persistent gap between domain-specific models trained on narrow datasets and general-purpose LLMs that can adapt to new tasks with minimal supervision. While prior research has shown that LLMs trained on biomedical data can excel at extraction tasks, little is known about how non-biomedical, general-purpose LLMs perform in high-stakes domains like MOE when only minimal in-context examples are provided ([Landolsi et al.,](#)

[2023](#); [Peng et al., 2023](#)).

In this study, we position our work within the evolving trajectory from rule-based to LLM-based systems by focusing on the capabilities of a non-domain-specific model for the shared task. Our analysis complements prior work by quantifying how far prompt-engineered, general-purpose LLMs can go in structured clinical IE, without the cost and complexity of large-scale biomedical pretraining or fine-tuning, while identifying key strengths and weaknesses across different MOE subtasks.

### 3 Task Description

The MEDIQA-OE 2025 Shared Task ([Corbeil et al., 2025b](#)) targets the extraction of structured medical orders from extended, doctor–patient conversations. The objective is to streamline clinical documentation, reduce provider workload, and ensure reliable capture of essential patient information from lengthy conversations.

Given a dialogue  $\mathcal{D} = \{(t_i, s_i, u_i)\}_{i=1}^N$ , where  $t_i$  is the turn\_id,  $s_i \in \{\text{DOCTOR, PATIENT}\}$  denotes the speaker, and  $u_i$  is the utterance text, the goal is to predict a set of medical orders  $\mathcal{O} = \{o_j\}_{j=1}^M$ .

Each medical order  $o_j$  is divided into four subtasks as a tuple  $(\text{type}_j, \text{desc}_j, \text{reason}_j, \text{prov}_j)$ , where  $\text{type}_j \in \{\text{medication, lab, imaging, followup}\}$ ,  $\text{desc}_j$  is the textual description,  $\text{reason}_j$  is the clinical justification, and  $\text{prov}_j \subseteq \{t_1, \dots, t_N\}$  contains the supporting turn IDs.

Multiple medical orders may be present per dialogue, and systems must extract all relevant orders with accurate structure and provenance grounding. Success in this task requires models to handle long-range dependencies, differentiate between clinically relevant and incidental information, and produce outputs in a consistent, structured format that can be directly integrated into electronic health record (EHR) systems. [Figure 1](#) illustrates the input format, subtask definitions, and expected output.

### 4 Dataset

The MEDIQA-OE dataset ([Corbeil et al., 2025a](#)) consists of multi-turn doctor–patient conversations annotated with medical orders. Each instance is a JSON object containing an id, a list of expected\_orders, and a transcript of turns  $(t_i, s_i, u_i)$ , where  $t_i$  is the turn ID,  $s_i$  denotes the speaker, and  $u_i$  is the utterance. Orders are annotated as  $(\text{type, desc, reason, prov})$ , with prov

representing the supporting turn IDs.

Set	#Enc	Follow-Up	Imaging	Lab	Medication
Train	63	25	14	29	75
Dev	100	41	26	71	117
Test	100	-	-	-	-

Table 1: Number of encounters and order types per set. Gold labels for the test set have not been released.

Transcripts were sourced from the PriMock57 (Papadopoulos Korfiatis et al., 2022) and ACI-Bench (Yim et al., 2023) datasets, with annotations merged using the official preprocessing script. The dataset, derived from the SIMORD corpus with an inter-annotator agreement of  $0.768\kappa$ , was curated by experts following post-encounter documentation practices, capturing both explicit and implicit orders that often require multi-turn reasoning.

**Dataset Analysis.** Table 1 provides the distribution of the dataset with a breakdown of each `order_type`. *Medication* orders dominate in both training and development sets, followed by *lab*, *followup*, and *imaging*, reflecting a clear class imbalance that may bias models toward frequent types in case of model fine-tuning. Dialogues are long, averaging 95.4 turns in training, 102.1 in development, and 101.6 in test, with the longest spanning 290 turns and over 2,900 tokens, posing challenges for models with limited context windows. Across all sets, doctors produce the majority of content; for example, in the test set, they contribute 6,123 turns and 89,449 tokens compared to 4,037 turns and 39,362 tokens from patients. Follow-up suggestions appear in roughly one-third of encounters and align with annotated follow-up orders. Incomplete annotations are also present, with roughly one-fifth of orders lacking a reason field. Provenance spans averaging only 1–2 turns, making both reason capture and evidence attribution challenging. These long contexts, skewed label distribution, implicit or missing reasons, and brief evidence spans, motivate models that can (i) maintain long-range dialogue state, (ii) generalize with little task-specific supervision, and (iii) ground outputs to cited turns. LLMs can handle extended inputs, adapt with few-shot prompts, and return structured fields with explicit provenance, making them the ideal candidate for this task.

**Prompt**

**<System>**  
 "You are a *clinical assistant* specialized in extracting **medical orders** from transcripts. Your job is to identify expected clinical orders from doctor-patient conversations. Return each order on a new line using the following comma-separated format wrapped with '\$' symbols: "\$*order\_type, description, reason, provenance*)\$"
 

- **order\_type** must be one of *[followup, imaging, lab, medication]*
- **description**: short, simple description of medical condition from the transcript
- **reason**: concise reason for the description taken exactly from the transcript
- **provenance**: list of the *turn\_id*(s) of the utterance(s) containing the reason, separated by commas for multiple

 Use the keyword 'null' for any missing field. Only return plain text in the specified format. There might be multiple clinical orders in the conversation."
   
**</System>**
  
**<User>**  
 Exemplar transcript  
*[turn\_id] Speaker: Utterance*  
 ...  
**</User>**
  
**<Assistant>**  
*\$(lab, hemoglobin a1c in about 6 months, diabetes type 2, [9, 135])\$*
  
**</Assistant>**
  
**<User>**  
 Current case transcript  
*[turn\_id] Speaker: Utterance*  
 ...  
**</User>**

**Expected Output**

**<Assistant>**  
*\$(lab, blood white blood cells two to three weeks, significantly elevated white blood cell count of 23,000, [8, 83, 86])\$*
  
**</Assistant>**

Figure 2: Few-shot prompt showing system instructions, exemplar input/output, the input query, and the expected model output.

## 5 Methodology

We use several general-domain Meta-Llama (Touvron et al., 2023) models for MEDIQA-OE 2025, moving from zero-shot with a smaller model to few-shot with larger ones. We select Meta-Llama for its open weights, long context window, and strong instruction following ability, which is necessary for this task.

Input transcripts from the dataset were converted from JSON to a plain-text format with one utterance per line:

[turn\_id] Speaker: Utterance

This ensured that turn order and speaker roles were preserved for downstream reasoning.

### 5.1 Model Configurations

We evaluated three LLM configurations:

**1. LLaMA 3 Inference (Zero-Shot):** First, *meta-llama/Llama-3-8B-Instruct* (Grattafiori et al., 2024) with no in-context examples. The system prompt defined the model’s role as a clinical assistant and specified the schema and constraints.

**2. LLaMA-4 Inference (Zero-Shot):** Next, *meta-llama/Llama-4-Scout-17B-16E-Instruct* (Meta, 2025) with the same prompt design, leveraging a larger model for potentially better reasoning and grounding.

**3. LLaMA-4 Inference (Few-Shot):** Finally, added a single in-context example from the training set, formatted as a user–assistant exchange preceding the inference case. The assistant’s example output illustrated the correct schema and provenance formatting, providing the model with a domain-specific reference.

## 5.2 Prompt Design

The prompt, illustrated in Figure 2, defined the model’s role as a clinical assistant tasked with extracting all medical orders from a doctor–patient conversation. It explicitly described the target output schema and the allowable values for `order_type`, and it stated what each subtask should return. Additional guidance required `null` for missing fields, allowed multiple orders per conversation, and constrained provenance to turn IDs for the supporting utterances. The transcript was provided line by line with turn IDs and speaker roles, followed by either no example (zero-shot) or one exemplar with its gold output (few-shot). We reached this final prompt through iterative refinement: an initial version asked only for `order_type`, `description`, `reason`, and `provenance` without role assignment, field definitions, or a fixed format, which led to long free-form text, difficult post-hoc parsing, and generic reasons that did not cite explicit spans. Assigning the clinical-assistant role and explaining each field improved grounding and produced more concise outputs. Requesting strict JSON next proved unreliable, as the model often added extra keys or commentary. We therefore switched to a comma-separated line format that the model followed more consistently. This process ensured clear instructions, a faithful mapping to the schema, and outputs that were both grounded and easy to parse.

## 5.3 Post-Processing

Following inference, raw model outputs were normalized and structured to match the required format. This process involved removing any extraneous text such as preambles or explanations, ensuring that all four fields were present, and explicitly assigning `null` to missing fields. The `order_type` field was standardized to the allowed set, and the `provenance` field was validated to contain only integer turn IDs within the valid range for each conversation. Outputs were then serialized into JSON for evaluation. When predictions contained minor formatting issues, such as concatenated fields or misplaced delimiters, these were corrected automatically; predictions that could not be repaired were discarded to avoid evaluation errors.

## 5.4 Experimental Setup

All experiments were conducted on a High-Performance Computing (HPC) environment with NVIDIA A100 80GB GPUs using mixed-precision (`bf16`) inference to optimize memory usage and runtime. The maximum context length was set to 8,192 tokens with a generation limit of 1,024 tokens. Decoding parameters were fixed across runs (`temperature=0.2`, `top_p=0.9`) to balance determinism and variability. Random seeds were fixed across the model, tokenizer, and generation routines for reproducibility.

## 5.5 Evaluation

System outputs are first aligned to gold-standard orders through a description-based pairing process. Matching is performed on the `description` field after normalization, which lowercases text and removes selected punctuation. Orders are excluded from evaluation if they have an empty `description` or an `order_type` outside the allowed set `medication`, `lab`, `followup`, `imaging`.

Once aligned, each field is scored with a metric suited to its content. The `description` field is evaluated with ROUGE-1  $F_1$ , rewarding unigram overlap with the reference and granting partial credit for preserving key clinical terms even if phrasing differs. The `reason` field is also scored with ROUGE-1  $F_1$ , capturing semantic similarity despite surface variation in justifications. The `order_type` field uses a STRICT  $F_1$ , counting only exact matches among the four permissible categories to penalize misclassification. The `provenance` field is evaluated with a MULTILABEL  $F_1$ , treating provenance

Team Name	description	reason	order_type	provenance	avg_score
<b>MasonNLP</b>	39.05	19.78	50.91	41.32	37.76
WangLab	66.77	29.49	81.45	63.04	60.19
silver_shaw	64.06	41.30	74.74	60.44	60.14
MISo KeaneBeanz	57.99	35.64	71.56	48.38	53.39
EXL Health AI Lab	54.45	30.50	66.17	52.47	50.90
HerTrials	19.61	8.99	29.59	5.61	15.95

Table 2: MEDIQA-OE 2025 leaderboard results (F1 in %). Top six systems, rows sorted by average score; MasonNLP shown first for reference.

System	description	reason	order_type	provenance	avg_score
LLaMA-3 8B (Zero-shot)	30.20	13.95	40.79	27.10	28.01
LLaMA-4 17B (Zero-shot)	36.82	15.60	47.23	30.32	32.49
LLaMA-4 17B (Few-shot)	<b>39.05</b>	<b>19.78</b>	<b>50.91</b>	<b>41.32</b>	<b>37.76</b>

Table 3: Performance across experimental setups (F1 in %). Best values are in bold.

as a set of turn IDs and balancing precision (excluding unrelated turns) with recall (capturing all relevant turns).

The final shared-task score is the unweighted mean of the four primary field-level  $F_1$  scores (description\_ROUGE1\_f1, reason\_ROUGE1\_f1, order\_type\_Strict\_f1, and provenance\_MultiLabel\_f1).

## 6 Results and Discussion

### 6.1 Leaderboard Performance

The MEDIQA-OE 2025 shared task attracted participation from **17 teams**, producing a total of **105 submissions**. Our **MasonNLP** system, based on a few-shot prompting setup with the general-purpose LLaMA-4 17B model and no domain-specific fine-tuning, achieved an average  $F_1$  score of **37.76**, placing competitively among the top-ranked systems. Table 2 presents the top six leaderboard with subtask-specific scores, with our system listed first for clarity. Notably, this performance was obtained without incorporating clinical-domain pretraining or retrieval augmentation, competing against systems that leveraged specialized architectures or domain-specific resources.

### 6.2 Ablation Study

To better understand the impact of model scale and prompting strategy, we evaluated three configurations: LLaMA-3 8B zero-shot, LLaMA-4 17B zero-shot, and LLaMA-4 17B few-shot (final submission). Results in Table 3 show a clear progression in average  $F_1$  across configurations. Mov-

ing from LLaMA-3 to LLaMA-4 improved performance in all subtasks, especially description and order\_type, reflecting the larger model’s stronger capacity for identifying and categorizing medical orders in long transcripts. This aligns with the dataset’s high average turn count and doctor-heavy content, which demand robust long-context processing. Introducing a single in-context example further improved all four subtasks, with the largest relative gain in provenance, suggesting that even minimal task-specific guidance helps the model ground predictions more accurately and follow the required structured format.

### 6.3 Discussion and Implications

These findings confirm our initial hypothesis that larger, instruction-tuned LLMs provide measurable benefits for MOE from long, multi-turn dialogues, even without domain-specific fine-tuning. The improvements from the few-shot configuration validate our contribution, which claims that minimal in-context supervision can close part of the performance gap between general-purpose LLMs and domain-adapted systems. However, reason extraction remains the most challenging subtask, likely due to the implicit nature of many clinical justifications in the dataset. Similarly, while provenance accuracy improved, grounding still lags behind other subtasks, reflecting the difficulty of linking orders to scattered and sometimes indirect evidence in the dialogue.

Overall, the results suggest that combining large general-purpose LLMs with carefully designed prompts and minimal in-context examples can

yield competitive performance in structured clinical IE. Future gains may require integrating retrieval-based grounding or domain adaptation to better handle implicit reasoning and improve evidence alignment.

## 7 Error Analysis

Building on the results in Section 6, we conducted a detailed error analysis of our best-performing LLaMA-4 17B few-shot system to better understand its strengths and remaining challenges across the four subtasks. The development set offers gold-standard annotations for all fields, enabling both quantitative and qualitative assessment. The test set, lacking gold-standard annotations, is analyzed only for schema validity.

Metric	Score
description_ROUGE1_f1	44.53
reason_ROUGE1_f1	25.13
order_type_Strict_f1	57.28
provenance_MultiLabel_f1	40.17
avg_score	41.78

Table 4: Development set scores for the LLaMA-4 17B few-shot system (F1 in %).

### 7.1 Development Set Analysis

Table 4 summarizes the official shared-task metrics for the development set. Consistent with leaderboard results, the model showed strong performance in `order_type` classification and description extraction, while reason and provenance remained more challenging. The model also broke down a single order into multiple orders in some cases, as illustrated in Figure 3. To explore why, we examined a few samples of matched and unmatched predictions, categorizing representative patterns for each subtask.

**Description.** The model was able to identify the correct target of an order in most cases, even in multi-turn, context-heavy transcripts. Many predictions contained the correct general test or medication, but lacked finer details such as timing or exact test subtype. For example, “*blood work*” was produced for the gold-standard “*blood white blood cells two to three weeks*”. This indicates that the model successfully locates the core clinical action but sometimes omits modifiers, an area that could be enhanced by incorporating temporal and entity-specific cues.

**Reason.** In most cases, the model provided a plausible reason aligned with the overall clinical context. For instance, it correctly linked a lab order to white blood cell count monitoring, though it occasionally summarized the reason more generally (“*to review lab results*”) instead of including explicit values. This shows that the model is capable of long-context integration to capture the essence of clinical justification, with potential for refinement through methods that encourage inclusion of specific numeric and temporal evidence.

**Order Type.** Order type classification was generally strong, but certain linguistic patterns led to confusion. Scheduling phrases (e.g., “*two to three weeks*”) were sometimes interpreted as follow-up visits rather than scheduled labs. Invalid `order_type`, present in 8 instances, included mentions of *surgery* (3), *referral* (1), and *null\_type* (4). Such mix-ups likely arise when multiple order-like actions occur in close proximity, and can be addressed by fine-tuning with examples emphasizing subtle category distinctions.

**Provenance.** The model demonstrated the ability to identify at least one correct evidence turn for most orders, as in the case where it predicted provenance [100] while the gold-standard label included both [98, 100]. This partial grounding suggests that the model can reliably find the key confirmation turn, but may miss earlier reason turns when information is distributed. Expanding its retrieval capacity for dispersed evidence could close this gap.

### 7.2 Test Set Analysis

Out of all predicted orders, 20 (4.7%) lacked a description, 8 (1.9%) contained an invalid `order_type` that included the same three keywords as we saw in the predicted orders of development set (*surgery*, *referral*, *null\_type*), 57 (13.3%) were missing a reason, and 46 (10.8%) omitted provenance identifiers. These results show that the system generally produces well-structured outputs with relatively few schema violations, though systematic omissions and field-level incompleteness directly reduce evaluation scores. Invalid `order_type` predictions typically arose from ambiguous dialogue phrasing that led the model to select categories outside the permitted set medication, lab, followup, imaging. For description, beyond the 20 missing fields, 11 cases involved text not present in the transcript,

Subtask	Predicted Order	Gold Order
<b>Description (Under-specified)</b>	lab, "complete blood work", reason: "to check white blood cell count"	lab, "blood white blood cells two to three weeks", reason: "significantly elevated white blood cell count of 23,000"
<b>Description (Hallucination)</b>	followup, "email follow-up in one month"	medication orders only, no follow-up
<b>Reason (Implicit Summary)</b>	lab, "blood white blood cells two to three weeks", reason: "to review lab results"	lab, "blood white blood cells two to three weeks", reason: "significantly elevated white blood cell count of 23,000"
<b>Order Type (Mix-up)</b>	followup, "two to three weeks"	lab, "blood white blood cells two to three weeks"
<b>Provenance (Partial)</b>	lab, "lipid panel", provenance: [100]	lab, "lipid panel", provenance: [98, 100]

Figure 3: Examples of different error types for each subtask.

reflecting hallucination or paraphrasing of plausible but unsupported orders. For reason, omission was dominant, with 57 missing values and 4 ungrounded justifications, indicating persistent difficulty in capturing implicit or distributed reason. For provenance, there are 46 missing spans. Partial grounding, common in the development set, likely persists here, underscoring the need for stronger evidence attribution.

**Overall Observations.** The analysis shows that instruction-tuned LLMs, even without domain-specific fine-tuning, can handle complex, multi-turn clinical dialogues to extract actionable orders with reasonable accuracy. While finer details (e.g., exact timing, numeric values, dispersed evidence) are sometimes omitted, the model frequently identifies the correct order, reason, and at least one key supporting turn. Hallucinations, as with most LLMs, are still present, highlighting the potential benefit of RAG (Lewis et al., 2020). With targeted enhancements, these strengths can be leveraged to develop robust clinical NLP systems capable of supporting real-world documentation workflows.

## 8 Conclusion

We addressed medical order extraction from multi-turn doctor-patient conversations using general-

domain Meta-Llama models without domain-specific fine-tuning. The setup began with zero-shot prompting on a smaller model and then moved to few-shot prompting on larger models. A simple structured prompt returned order type, description, reason, and provenance with cited turns. The error analysis shows that the model struggles with temporal and numeric specificity, occasional hallucination, under-specific reasons, and partial provenance spans. These gaps narrowed with larger models and a few clear exemplars. The findings indicate that general domain LLMs are a viable base when guided by domain cues, retrieval to reduce hallucination, and schema validators for strict JSON. Overall, the study shows that instruction-tuned LLMs can handle long clinical dialogues with minimal adaptation and provides a practical template for grounded multi-field clinical IE with clear next steps on reason modeling, tighter provenance, and better balance across order types.

## Limitations

Our approach avoids any domain-specific pretraining or fine-tuning on clinical corpora. While integrating such specialization could potentially yield further gains, our goal was to assess the adaptability of a general-purpose, instruction-tuned LLM in a highly specialized medical order extraction



task using only prompt engineering. This choice enables a fair evaluation of the model’s zero- and few-shot capabilities, providing insights into its out-of-the-box performance without reliance on costly domain-specific data or retraining. The strong results achieved by our few-shot LLaMA-4 system demonstrate that competitive baselines can be established under these conditions, laying the groundwork for future enhancements through targeted domain adaptation.

## References

- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Alan R Aronson and François-Michel Lang. 2010. An overview of metemap: historical perspective and recent advances. *Journal of the American medical informatics association*, 17(3):229–236.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl\_1):D267–D270.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Emily M Campbell, Dean F Sittig, Joan S Ash, Kenneth P Guappone, and Richard H Dykstra. 2006. Types of unintended consequences related to computerized provider order entry. *Journal of the American Medical Informatics Association*, 13(5):547–556.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Fenia Christopoulou, Thy Thy Tran, Sunil Kumar Sahu, Makoto Miwa, and Sophia Ananiadou. 2020. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46.
- Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jérôme Tremblay, Akila Jeesson Daniel, Cari Bader, Yu-Cheng Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025a. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. *CoRR*.
- Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeesson Daniel, Miguel Del-Agua, and François Beaulieu. 2025b. Overview of the mediqua-oe 2025 shared task on medical order extraction from doctor-patient conversations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.
- Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.
- Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015.
- Carol Friedman. 2000. A broad-coverage natural language processing system. In *Proceedings of the AMIA Symposium*, page 270.
- Qiwei Gan, Mengke Hu, Kelly S Peterson, Hannah Eyre, Patrick R Alba, Annie E Bowles, Johnathan C Stanley, Scott L DuVall, and Jianlin Shi. 2023. A deep learning approach for medication disposition and corresponding attributes extraction. *Journal of biomedical informatics*, 143:104391.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Cyril Grouin, Natalia Grabar, and 1 others. 2023. Year 2022 in medical natural language processing: availability of language models as a step in the democratization of nlp in the biomedical area. *Yearbook of Medical Informatics*, 32(01):244–252.
- Udo Hahn and Michel Oleynik. 2020. Medical information extraction in the age of deep learning. *Yearbook of medical informatics*, 29(01):208–220.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Abhyuday N Jagannatha and Hong Yu. 2016. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the conference on empirical methods in natural language processing*.

- conference on empirical methods in natural language processing*, volume 2016, page 856.
- Madaline Kinlay, Wu Yi Zheng, Rosemary Burke, Ilona Juraskova, Rebekah Moles, and Melissa Baysari. 2021. Medication errors related to computerized provider order entry systems in hospitals and how they change over time: a narrative review. *Research in Social and Administrative Pharmacy*, 17(9):1546–1552.
- Gilad J Kuperman and Richard F Gibson. 2003. Computer physician order entry: benefits, costs, and issues. *Annals of internal medicine*, 139(1):31–39.
- Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. 2023. Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Yves A Lussier, Lyudmila Shagina, and Carol Friedman. 2001. Automating snomed coding using medical language underchen2020jointstanding: a feasibility study. In *Proceedings of the AMIA Symposium*, page 418.
- Diwakar Mahajan, Jennifer J Liang, Ching-Huei Tsou, and Özlem Uzuner. 2023. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *Journal of biomedical informatics*, 144:104432.
- AI Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, checked on, 4(7):2025.
- Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.
- Sankaran Narayanan, Kaivalya Mannam, Pradeep Achan, Maneesha V Ramesh, P Venkat Rangan, and Sreeranga P Rajan. 2022. A contextual multi-task neural approach to medication and adverse events identification from clinical text. *Journal of biomedical informatics*, 125:103960.
- Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. Pri-Mock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.
- Jon Patrick and Min Li. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527.
- Cheng Peng, Xi Yang, Zehao Yu, Jiang Bian, William R Hogan, and Yonghui Wu. 2023. Clinical concept and relation extraction using prompt-based machine reading comprehension. *Journal of the American Medical Informatics Association*, 30(9):1486–1493.
- Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.
- Raymond Francis Sarmiento and Franck Dernoncourt. 2016. Improving patient cohort identification using natural language processing. *Secondary analysis of electronic health records*, pages 405–417.
- Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Irena Spasic and Goran Nenadic. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.
- Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- T Vira, M Colquhoun, and E Etchells. 2006. Reconcilable differences: correcting medication errors at hospital admission and discharge. *BMJ Quality & Safety*, 15(2):122–126.
- Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and 1 others. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.
- Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. Medex: a medication information extraction system for clinical narratives. *Journal of the American medical informatics Association*, 17(1):19–24.
- Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.
- Wen-wai Yim, Yajuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.

# EXL Health AI Lab at MEDIQA-OE 2025: Evaluating Prompting Strategies with MedGemma for Medical Order Extraction

Abhinand Balachandran, Bavana Durgapraveen  
Gowsikkan Sikkan Sudhagar, Vidhya Varshany J S, Sriram Rajkumar  
EXL Service  
{abhinand.balachandran, sriram.rajkumar}@exlservice.com

## Abstract

The accurate extraction of medical orders from doctor-patient conversations is a critical task for reducing clinical documentation burdens and ensuring patient safety. This paper details our team submission to the MEDIQA-OE-2025 Shared Task. We investigate the performance of MedGemma, a new domain-specific open-source language model, for structured order extraction. We systematically evaluate three distinct prompting paradigms: a straightforward one-Shot approach, a reasoning-focused ReAct framework, and a multi-step agentic workflow. Our experiments reveal that while more complex frameworks like ReAct and agentic flows are powerful, the simpler one-shot prompting method achieved the highest performance on the official validation set. We posit that on manually annotated transcripts, complex reasoning chains can lead to "overthinking" and introduce noise, making a direct approach more robust and efficient. Our work provides valuable insights into selecting appropriate prompting strategies for clinical information extraction in varied data conditions.

## 1 Introduction

The proliferation of ambient clinical intelligence (ACI) systems promises to revolutionize healthcare by automating the burdensome task of clinical documentation (Yim et al., 2023). A cornerstone of this automation is the ability to transform unstructured doctor-patient dialogue into structured, actionable data suitable for Electronic Health Records (EHRs) (Zhang et al., 2023). Among the most critical data to capture are medical orders—medications, lab tests, imaging studies, and follow-ups—where accuracy is paramount for patient care and safety (Singhal et al., 2023). The MEDIQA-OE 2025 shared task provides a crucial benchmark for this challenge, pushing the field to develop systems that can parse long, complex conversations to extract a variety of order types and their corresponding

clinical justifications (Corbeil et al., 2025a). This task moves beyond simple entity recognition, requiring a deep understanding of context, negation, and the relationships between a medical order and its underlying reason.

## 2 Shared Task and Dataset

The MEDIQA-OE 2025 shared task (Corbeil et al., 2025a) represents a significant advancement in clinical natural language processing, requiring participants to extract structured medical orders from dialogue transcripts. The dataset, SIMORD (Corbeil et al., 2025b), is derived from mock clinical consultations and annotated by medical professionals, providing a robust foundation for developing automated clinical documentation systems. Medical order extraction involves identifying and structuring various medical orders—such as medications, imaging studies, lab tests, and follow-ups—based on doctor-patient conversations. This complex task goes beyond simple entity recognition, requiring systems to understand the clinical context, temporal relationships, and hierarchical structure of medical directives. Previous efforts in this domain have primarily focused on extracting entities and relations from clinical texts, but have often been limited to structured electronic health records or simplified clinical notes rather than the nuanced, conversational format of real-time clinical interactions. The MEDIQA-OE 2025 shared task addresses these limitations by presenting a more realistic and challenging scenario that mirrors actual clinical practice. This shared task seeks to develop effective solutions for improving clinical documentation, reducing the administrative burden on healthcare providers, and ensuring that critical patient information is accurately captured from long, complex conversations that may span multiple topics and include interruptions, clarifications, and informal language typical of natural clinical

dialogue. The input dialogues are sourced from a combination of existing conversational datasets, including ACI-Bench (Yim et al., 2023), which focuses on ambient clinical intelligence and automatic visit note generation, and PriMock57 (Korfatis et al., 2022), a comprehensive dataset of primary care mock consultations. These datasets provide diverse conversational patterns and clinical scenarios, ensuring that participating systems are evaluated on realistic variations in communication styles, medical specialties, and patient presentations. The structured lists of medical orders are created by qualified medical annotators who possess the clinical expertise necessary to accurately identify, categorize, and structure the complex medical directives that emerge from these conversations, ensuring high-quality ground truth labels that reflect real-world clinical decision-making processes. This comprehensive approach to dataset creation makes MEDIQA-OE 2025 a valuable benchmark for advancing the state-of-the-art in clinical conversation understanding and automated medical documentation systems. Table 1 presents the distribution of clinical encounters and the corresponding extracted orders in the Train and Dev datasets. For each conversation, systems must extract all relevant orders and structure them with the following key attributes:

- **Order Type:** The category of the order (e.g., *Medication, Lab, Imaging, Follow-up*).
- **Description:** The specific details of the order (e.g., *“Lisinopril 10mg daily”*).
- **Reason:** The clinical justification for the order (e.g., *“for high blood pressure”*).
- **Provenance:** The specific text spans in the transcript from which the information was extracted.

### 3 Related Work

Clinical Natural Language Processing (NLP) has undergone a significant methodological shift, evolving from rule-based systems to advanced Agentic systems powered by transformers. The Dialogue Medical Information Extraction task was initially addressed by combining Named Entity Recognition (NER) and Relation Extraction (RE). Early rule-based systems relied on semantic lexicons and regular expressions for pattern matching, offering

interpretability but facing limitations in scalability and coverage (Ben Abacha et al., 2021).

More recently, supervised heterogeneous graph-based approaches have demonstrated superior performance in mapping medical items to their statuses by enriching their representation with broader dialogue context (Zhang et al., 2023). Concurrently, GPT-based models utilizing various prompting strategies have been effectively employed for clinical information extraction (Lehman et al., 2023)

However, much of this prior work has focused on information extraction with minimal emphasis on complex relation identification (Yim et al., 2023). The current challenge extends beyond just medication extraction to encompass lab orders, imaging studies, and follow-up instructions—areas that lack systematic research. A key difficulty lies in accurately mapping orders to their precise reasons, which is crucial for healthcare workflows (Sinsky et al., 2024). Our research contributes to this area by systematically comparing prompting strategies—from simple in-context learning to complex agentic AI—to develop a robust medical order extraction system for challenging clinical settings (Gao et al., 2023; Lewis et al., 2020).

## 4 Methodology

Our entire approach is built upon the MedGemma family of models, which are variants of Google’s Gemma models further pre-trained and fine-tuned on a vast corpus of medical literature and clinical data. This domain-specific tuning endows them with a strong baseline understanding of medical terminology and concepts. We explored both the 4B and 27B parameter variants to assess the impact of model scale. We designed and tested three distinct prompting frameworks.

### 4.1 Approach 1: 1-Shot Prompting

This is our simplest and most direct approach. The model is given a single, high-quality example of a complete conversation transcript and its corresponding structured JSON output. The test transcript is then appended, and the model is instructed to generate the JSON output in the same format. The prompt is structured to be clear and concise, minimizing cognitive load and relying on the model’s powerful in-context learning ability to replicate the task.

Dataset	Encounters	Follow-up	Imaging	Lab	Medication	Total Orders
Train	63	25	14	29	75	143
Dev	100	41	26	71	117	255

Table 1: Distribution of clinical encounters and extracted orders across different categories (Follow-up, Imaging, Lab, and Medication) in the Train and Dev datasets.

## 4.2 Approach 2: ReAct Framework

Inspired by the ReAct (Reasoning and Acting) paradigm, this approach encourages the model to "think out loud." The process begins with the conversation transcript, which is combined with a system prompt that specifies the extraction categories (medication, lab, imaging, follow-up), rules, and output format. The transcript is then preprocessed by segmenting individual turns, identifying the physician’s utterances, and assigning normalized turn identifiers. Within each conversation, the extraction follows an iterative ReAct cycle. In the Thought step, the model analyzes the physician’s turns to detect potential medical orders. The Action step then generates candidate extractions, specifying the order type, a short description, the associated clinical reason, and the provenance (turn numbers). The Observation step validates these candidates by enforcing constraints: only doctor-initiated orders are kept, exact transcript wording must be preserved, descriptions and reasons are limited to 20 words, compound instructions are split, duplicate orders are removed, and the number of orders and provenance entries are capped. When inconsistencies are identified, the cycle repeats until valid outputs are produced. After this reasoning loop, a post-processing stage ensures consistency by normalizing order types, truncating fields to required lengths, ordering entries by their provenance, and validating against the target JSON schema. The final system output is a structured JSON array containing all extracted medical orders, with each entry including order type, description, reason, and provenance. This method aims to improve accuracy on complex cases by forcing the model to explicitly reason about its decisions before producing the final output.

## 4.3 Approach 3: Agentic Workflow

This is our most complex framework, decomposing the task across a simulated multi-agent pipeline within a single prompt context. Agent 1 (Identifier): Scans the entire transcript turn-by-turn and outputs a raw list of potential orders and reasons, tagged

with their turn IDs. Agent 2 (Mapper): Takes the output from Agent 1. Its sole job is to analyze the raw list and create explicit pairings between each identified order and its most likely reason. Agent 3 (Structurer): Receives the mapped pairs from Agent 2. It formats this information into the final, clean JSON structure, ensuring all fields are correctly populated. Agent 4 (Validator): Performs a final check on the generated JSON, comparing it against the original transcript to correct any obvious errors or hallucinations before producing the final output. This workflow was designed to modularize the cognitive process, hoping to reduce errors by having specialized "agents" focus on one sub-task at a time.

## 5 Evaluation metrics

We assess our system using four evaluation metrics, each designed to capture different dimensions of clinical order extraction. For the assessment of order descriptions and underlying reasons, we employ the ROUGE-1 F1 score which quantifies unigram-level overlap between predicted outputs and gold-standard references. ROUGE-1 is a standard metric in clinical NLP for measuring content accuracy in generated text, making it appropriate for evaluating free-text fields such as descriptive notes and justifications. For order type classification, we adopt a strict F1 metric, in which predictions are counted as correct only if they exactly match the annotated label. This stricter criterion prevents partial matches from inflating scores and is especially important when separating clinically distinct categories like medications, laboratory tests, imaging, and follow-up visits. For the task of provenance detection, we report a multi-label F1 score, since an order can be linked to multiple conversational sources (e.g., several speakers or turns). Multi-label evaluation provides a balanced measure of precision and recall in these cases, offering a more realistic view of performance. Collectively, these metrics evaluate both the textual alignment of generated outputs and the accuracy of structured predictions, resulting in a well-rounded

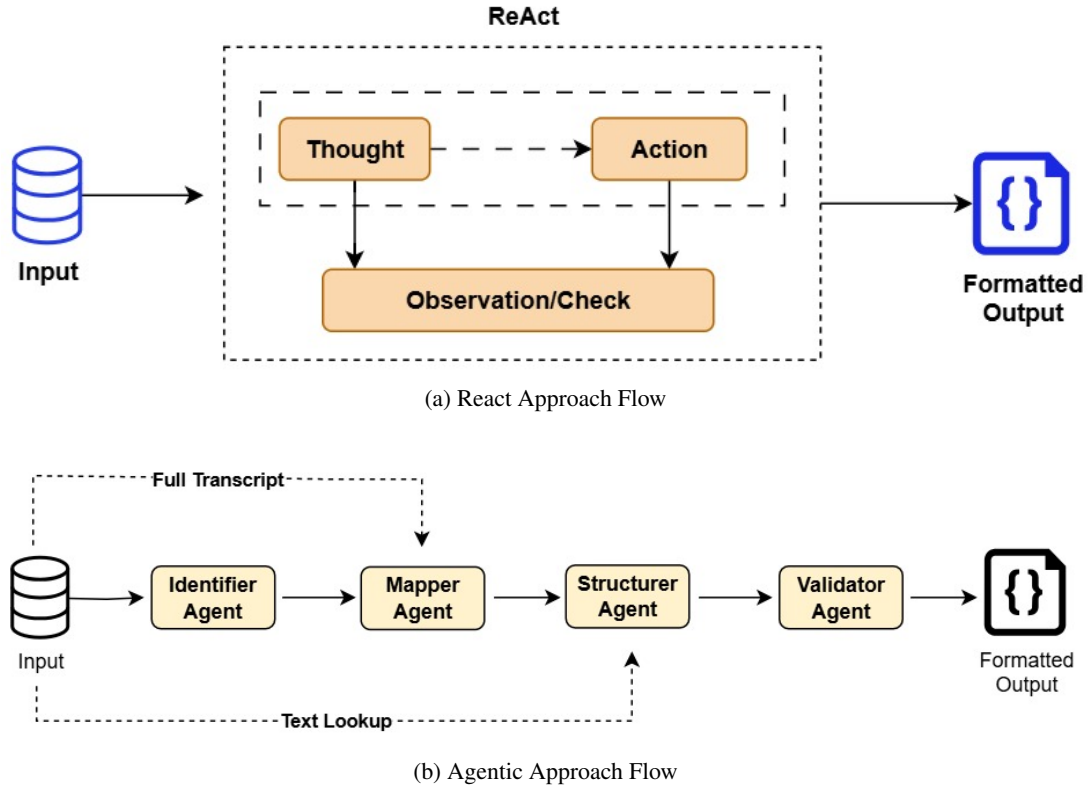


Figure 1: Overview of ReAct and Agentic Workflow frameworks for medical order extraction. The ReAct approach (Figure a) implements iterative thought-action-observation cycles where the model explicitly reasons through extraction decisions, validates candidates against constraints, and repeats until consistent outputs are achieved. The Agentic Workflow (Figure b) decomposes the task across four specialized agents: Identifier for detection, Mapper for pairing orders with reasons, Structurer for JSON formatting, and Validator for final verification.

framework for assessing clinical order extraction.

## 6 Results and Discussion

We evaluated our three approaches on the official validation set to systematically assess their performance and determine the optimal strategy for medical order extraction. Our experimental design was structured to answer a fundamental research question: Which prompting strategy demonstrates the highest effectiveness for this specialized clinical task? The evaluation framework encompassed multiple metrics to provide comprehensive insights into model performance across different aspects of the order extraction process. Initial experiments comparing the base Gemma model with its medical adaptation revealed significant improvements with domain-specific fine-tuning. MedGemma consistently outperformed the base model across all evaluation metrics, demonstrating the value of medical domain adaptation for clinical natural language processing tasks. This finding underscores the importance of specialized model training for health-

care applications, where domain-specific terminology and contextual understanding are crucial for accurate performance. Building upon this foundation, we conducted a comprehensive comparison of our three prompting strategies using the MedGemma-4B model on the development dataset. The results, presented in Table 2, revealed a clear performance hierarchy among the approaches. The 1-Shot methodology achieved the highest average score of 0.436, demonstrating superior performance across most evaluation dimensions. Specifically, it excelled in description generation (Rouge1-f1: 0.516) and order type classification (Strict-f1: 0.602), while maintaining competitive performance in reason extraction and provenance identification. In contrast, the ReAct approach yielded significantly lower scores across all metrics (average: 0.277), while the Agentic Workflow showed mixed results, performing well in certain areas like provenance detection (MultiLabel-f1: 0.488) but struggling with reason extraction (Rouge1-f1: 0.276). The results, presented in Table 3, revealed that

1-shot using MedGemma 27B approach stood at fourth place in the MEDIQA-OE 2025 shared task (Corbeil et al., 2025a)

To validate these findings and assess scalability, we replicated the experimental framework using the larger MedGemma-27B model on the test dataset. The results, shown in Table 3, confirmed our initial observations while demonstrating the benefits of increased model capacity. The 27B model achieved substantially higher scores across all metrics compared to its 4B counterpart, with the 1-Shot approach reaching an average score of 0.549. Notably, the performance improvements were particularly pronounced in description generation (Rouge1-f1: 0.591) and order type classification (Strict-f1: 0.703). Despite the overall performance gains from the larger model, the relative ranking of prompting strategies remained consistent, with 1-Shot maintaining its superiority over the ReAct approach (average: 0.370). Our analysis revealed a counterintuitive but significant finding: increased complexity in prompting strategies did not translate to improved performance for this specific task. The ReAct and Agentic frameworks, despite their theoretical sophistication and success in other domains, consistently underperformed relative to the simpler 1-Shot approach. Through detailed error analysis, we identified that this phenomenon stems from what we term "analytical over-processing"—the more complex frameworks occasionally generated spurious intermediate reasoning steps that introduced errors rather than enhancing accuracy. The models would sometimes fabricate relationships between dialogue elements or misinterpret subtle clinical nuances during their multi-step reasoning processes, ultimately degrading precision. Since the MEDIQA-OE dataset consists of carefully annotated clinical transcripts with well-defined ground truth, the additional inferential layers introduced by complex prompting strategies contributed more noise than valuable signal. The 1-Shot approach, by maintaining a more direct and constrained generation process, proved less susceptible to such systematic errors while offering additional benefits in terms of computational efficiency and implementation simplicity.

## Limitations

The primary limitation of our study is tied to our main finding. Our conclusion that 1-Shot prompting is superior is heavily dependent on the

MEDIQA-OE dataset. In a real-world clinical setting with noisy ASR transcripts, interruptions, and less structured speech, the explicit reasoning steps of a ReAct or Agentic framework might be necessary to disambiguate the input and could potentially outperform a direct 1-Shot approach. Our work does not test this hypothesis. Furthermore, our Agentic workflow was implemented within a single model using four specialized agents to handle different subtasks; while this design demonstrates feasibility, a true multi-agent system with independent models could behave differently. This multi-agent setup also introduces practical limitations, as larger models such as MedGemma-27B make the Agentic approach computationally expensive and time-consuming, thereby increasing cost. At the same time, smaller models such as MedGemma-4B exhibited notable shortcomings: we observed hallucinations in ReAct and Agentic reasoning steps, as well as a tendency to replicate the few-shot examples provided rather than extracting new information from the transcript. In addition, there are several dataset-related limitations. Primarily, the available annotations are relatively sparse and at times inconsistent, with inter-annotator variability leading to ambiguity in what counts as a valid order. Secondly, the dataset does not always provide medically precise or standardized order labels, which limits the ability to evaluate correctness against clinically meaningful ground truth. Finally, certain order categories are underrepresented, creating class imbalance that could bias model performance.

## Conclusion

In this paper, we presented our investigation into medical order extraction for the MEDIQA-OE 2025 task. By systematically comparing 1-Shot, ReAct, and Agentic prompting frameworks with the MedGemma model, we demonstrated that for manually annotated clinical transcripts, a direct and simple 1-Shot approach is surprisingly effective. It outperformed more complex reasoning frameworks, which were prone to overthinking and introducing errors. This highlights a crucial lesson for applied NLP: the optimal solution is a function of not just the model's power, but also the characteristics of the data. Future work should explore these prompting paradigms on noisier, real-world clinical data to determine if the utility of complex reasoning frameworks becomes more apparent.



MedGemma-4B	Description (Rouge1_F1)	Reason (Rouge1_F1)	Order Type (Strict_F1)	Provenance (MultiLabel_F1)	Avg. Score
1-Shot	0.516	0.318	0.602	0.307	0.436
ReAct	0.363	0.120	0.465	0.160	0.277
Agentic Workflow	0.09	0.06	0.169	0.123	0.111

Table 2: Comparison of prompting strategies with MedGemma-4B with dev dataset. The 1-Shot approach yielded the best overall performance.

MedGemma-27B	Description (Rouge1_F1)	Reason (Rouge1_F1)	Order Type (Strict_F1)	Provenance (MultiLabel_F1)	Avg. Score
1-Shot	0.591	0.342	0.703	0.561	0.549
ReAct	0.353	0.283	0.497	0.350	0.370

Table 3: Comparison of prompting strategies with the larger MedGemma-27B model on the test dataset. The 1-Shot approach remains the most effective.

## Acknowledgments

We extend our sincere thanks to EXL Health AI Lab for their support and computing resources. We also appreciate the efforts of our colleagues who contributed to discussions and provided valuable assistance during the course of this work. Finally, we acknowledge the organizers for their efforts in hosting this interesting and challenging competition.

## References

- Philippe Corbeil, Asma Ben Abacha, Griffin Adams, and Dina Demner-Fushman. Overview of the MEDIQA-OE 2025 shared task on medical order extraction from clinical conversations. In *Proceedings of the 24th Workshop on Biomedical Natural Language Processing (BioNLP 2025)*. Association for Computational Linguistics.
- Philippe Corbeil, Asma Ben Abacha, Griffin Adams, and Dina Demner-Fushman. SIMORD: A synthetic dataset for medical order extraction from clinical conversations. In *Proceedings of the 24th Workshop on Biomedical Natural Language Processing (BioNLP 2025)*. Association for Computational Linguistics.
- Asma Ben Abacha, Kirk Roberts, Dina Demner-Fushman, and Wen-wai Yim. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP 2021)*. Association for Computational Linguistics. <https://aclanthology.org/2021.bionlp-1.37>
- Tianyu Gao, Adam Fisch, and Danqi Chen. Scaling instruction-finetuned language models with few-shot demonstrations. *Transactions of the Association for Computational Linguistics*, 11:150–167. [https://doi.org/10.1162/tacl\\_a\\_00511](https://doi.org/10.1162/tacl_a_00511)
- Eric Lehman, Dimitris Pappas, Byron C. Wallace, and Marzyeh Ghassemi. Do we still need clinical language models? Evidence from GPT-4. *arXiv preprint arXiv:2304.05372*. <https://arxiv.org/abs/2304.05372>
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, pages 9459–9474. <https://arxiv.org/abs/2005.11401>
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguerre y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. <https://doi.org/10.1038/s41586-023-06291-2>
- Christine Sinsky, Thomas Bodenheimer, and Robert J. Baron. Addressing inbox-driven work to reduce clinician burnout. *Journal of the American Medical Association*, 331(4):321–322. <https://doi.org/10.1001/jama.2023.28431>
- Wen-wai Yim, Kirk Roberts, Dina Demner-Fushman, and Yifan Luo. Overview of the MEDIQA 2023 shared task on consumer health question answering. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing*, pages 370–381. Association for Computational Linguistics. <https://aclanthology.org/2023.bionlp-1.35>
- Wen-wai Yim, Kirk Roberts, Dina Demner-Fushman, and Yifan Luo. Overview of the MEDIQA-WV 2025

shared task on wound care response generation. In *Proceedings of the 24th Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics. <https://aclanthology.org/2025.bionlp-1.12>

Xinyue Zhang, Shuo Jiang, Tao Yu, and Hua Xu. Incorporating structured metadata into clinical text generation. *Journal of Biomedical Informatics*, 140:104339. <https://doi.org/10.1016/j.jbi.2023.104339>

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Acibench: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Nature Scientific Data*, 10:586. <https://www.nature.com/articles/s41597-023-02487-3>

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598. Association for Computational Linguistics. <https://aclanthology.org/2022.acl-short.66>

## A Prompts used in this approach

One-shot approach prompt which was in **fourth place in the MEDIQA- OE 2025 competition** is mentioned below and the remaining experimental approach prompts are available in Git repo

### A.1 One-shot approach prompt

**SYSTEM PROMPT:** "" You are a medical AI assistant specialized in extracting EXPLICIT medical orders from doctor-patient conversations.

**CRITICAL RULES:** 1. Extract ONLY orders explicitly stated by the doctor 2. Do NOT infer or assume orders that aren't clearly mentioned 3. Provenance must be EXACT turn numbers where orders appear 4. Be balanced - i.e precision and recall on level terms 5. If the doctor orders multiple DISTINCT items (e.g., 'get a covid test and blood test'), create separate order objects for each item - never merge them into one combined description.

**Order Types:** - medication: Prescriptions, dosage instructions, medication changes - lab: Blood tests, urine tests, specific diagnostic tests - imaging: X-rays, MRI, CT scans, ultrasounds - followup: Scheduled return visits, check-ups (these must be explicitly stated by the doctor)

For each order extract: - order type: One of the 4 types above - description: EXACT medical terminology used by doctor - reason: Specific condition/symptom mentioned by doctor - provenance:

ONLY turn numbers where this exact order is mentioned""

**INSTRUCTION TEMPLATE :** ""Please extract all medical orders from the following doctor-patient conversation:

**CONVERSATION:** conversation

Extract all medical orders and return them as a JSON list with the following format: [ "order type": "medication|lab|imaging|followup|referral", "description": "specific description of the order", "reason": "medical condition or reason for the order", "provenance": [list of turn numbers where this order appears] ]

Focus on explicit orders given by the doctor. Be precise with medical terminology.""

**USER PROMPT:** f""EXAMPLE CONVERSATION: Turn 126 - DOCTOR: so, for your first problem of your shortness of breath i think that you are in an acute heart failure exacerbation. Turn 127 - DOCTOR: i want to go ahead and, uh, put you on some lasix, 40 milligrams a day. Turn 138 - DOCTOR: for your second problem of your type i diabetes, um, let's go ahead... i wanna order a hemoglobin a1c for, um, uh, just in a, like a month or so.

**EXPECTED OUTPUT:** [ { { "order type": "medication", "description": "lasix 40 milligrams a day", "reason": "shortness of breath acute heart failure exacerbation", "provenance": [126, 127] } }, { { "order type": "lab", "description": "hemoglobin a1c", "reason": "type i diabetes", "provenance": [138] } } ]

**NOW EXTRACT FROM THIS CONVERSATION:**

—  
{instruction template } ""

# PNLP at MEDIQA-OE 2025: A Zero-Shot Prompting Strategy with Gemini for Medical Order Extraction

Parth Mehta

silver\_shaw (Competition Team Name)

parthmehta2016@gmail.com

## Abstract

Medical order extraction from doctor-patient conversations presents a critical challenge in reducing clinical documentation burden and ensuring accurate capture of patient care instructions. This paper describes our system for the MEDIQA-OE 2025 shared task using the ACI-Bench and PriMock57 datasets, which achieved second place on the public leaderboard with an average score of 0.6014 across four metrics: description ROUGE-1 F1, reason ROUGE-1 F1, order-type strict F1, and provenance multi-label F1. Unlike traditional approaches that rely on fine-tuned biomedical language models, we demonstrate that a carefully engineered zero-shot prompting strategy using Gemini 2.5 Pro can achieve competitive performance without requiring model training or GPU resources. Our approach employs a deterministic state-machine prompt design incorporating chain-of-thought reasoning, self-verification protocols, and structured JSON output generation. The system particularly excels in reason extraction, achieving 0.4130 ROUGE-1 F1, the highest among the top performing teams. Our results suggest that advanced prompt engineering can effectively bridge the gap between general-purpose large language models and specialized clinical NLP tasks, offering a computationally efficient and immediately deployable alternative to traditional fine-tuning approaches with significant implications for resource-constrained health-care settings.

## 1 Introduction

Clinical documentation represents a significant burden for healthcare providers, consuming substantial time that could otherwise be devoted to patient care. The accurate extraction of medical orders from doctor-patient conversations is essential for maintaining comprehensive patient records while reducing this documentation overhead. The MEDIQA-OE 2025 (Corbeil et al., 2025b) shared task addresses this challenge by focusing on the automatic

identification and structuring of medical orders - including medications, laboratory tests, imaging studies, and follow-up appointments - from conversational transcripts.

Traditional approaches to medical information extraction have relied heavily on fine-tuned domain-specific models, which require substantial computational resources and annotated training data. These methods, while effective, present barriers to deployment in resource-constrained clinical environments where computational resources are limited and rapid deployment is essential.

This paper presents a paradigm shift in clinical NLP: leveraging the reasoning capabilities of large language models through sophisticated prompt engineering to achieve competitive performance without any model training. Our team’s submission to MEDIQA-OE 2025 demonstrates that a zero-shot approach using Gemini 2.5 Pro, combined with deterministic chain-of-thought reasoning and self-critique mechanisms, achieves second place on the public leaderboard with an average F1 score of 0.6014. Notably, this performance was attained without GPU requirements or model fine-tuning, making the solution immediately deployable.

Our approach challenges conventional wisdom about the necessity of domain-specific fine-tuning for clinical NLP tasks. The main contributions of this work are: (1) empirical evidence that prompt engineering can match or exceed fine-tuned models for medical order extraction; (2) a structured prompting methodology combining chain-of-thought reasoning with self-verification specifically designed for clinical information extraction, democratizing access to advanced clinical documentation tools.

## 2 Related Work

Medical order extraction from clinical conversations has evolved significantly, and the MEDIQA-

OE shared task series driving innovation in this domain. Early approaches focused on named entity recognition and relation extraction from structured clinical texts. Recent advances have shifted toward utilizing transformer architectures and processing unstructured clinical conversations.

BioBERT (Lee et al., 2019) established domain-adaptive pre-training as a strong baseline for biomedical NER, relation extraction, and QA, making encoder backbones prevalent in clinical NLP pipelines. BioClinical-ModernBERT (Sounack et al., 2025) introduced a domain-adapted long-context encoder for biomedical and clinical NLP, addressing a key limitation of earlier encoders when processing extended conversations. When paired with T5’s unified text-to-text framework (Raffel et al., 2019), this approach showed promise for joint classification and generation. Nevertheless, these architectures demanded substantial computational resources for fine-tuning and exhibited inconsistent narrative fidelity when extracting complex order descriptions.

The latest advances in medical-specific large-language models marked another evolutionary step. MedGemma (Google Health AI, 2025) and Lingshu (Team et al., 2025) incorporated extensive clinical knowledge through specialized pre-training on medical corpora. Despite theoretical advantages, practical deployment proved challenging for teams with limited resources—both models experienced out-of-memory errors on commodity hardware even with quantization strategies.

Our approach fundamentally departs from this fine-tuning paradigm. Rather than pursuing increasingly specialized models, we demonstrate that general-purpose LLMs like Gemini 2.5 Pro (DeepMind, 2025), trained on Trillion of tokens when guided by meticulously engineered prompts incorporating chain-of-thought reasoning and self-critique, surpass domain-specific models without any training. This finding aligns with emerging evidence that sophisticated prompt engineering can activate latent clinical reasoning capabilities in large language models, offering a computationally efficient alternative to traditional fine-tuning approaches.

## 3 System Description

### 3.1 Model Selection Journey

Our development process involved systematic exploration of increasingly sophisticated approaches,

each informing the final solution. This iterative journey revealed crucial insights about the trade-offs between model specialization and prompt engineering.

**Early Phase:** Rule-based classification attempts. Initial experiment with BioBERT for order-type classification—yielded high precision but missed many orders (recall issues). Attempts to explore joint classification and generation at attribute extraction using DeBERTa-v3 (He et al., 2021). However, the extracted spans were inconsistent, often capturing partial information or including conversational artifacts.

**Middle Phase:** Joint modeling approach combining a long-text encoder BioClinical ModernBERT with text decoder to handle long transcripts and generate structured output. Resulted in improved order identification but narrative fidelity was low (the generated text did not closely match the conversation details).

**Transition:** Trials with cutting-edge medical LLMs MedGemma 4B and Lingshu 7B for end-to-end generation. These models promised better medical knowledge and reasoning but were impractical to run on available hardware (out-of-memory issues on Google Colab’s free tier, even with 4-bit quantization).

**Final Phase:** Pivot to a pure prompt-engineering approach using an accessible large model Gemini 2.5 Pro. No fine-tuning, no task-specific training – instead, harness the model’s reasoning ability through carefully crafted prompts. This phase yielded a breakthrough: a +13 F1 point jump in average score over the best fine-tuned attempt, achieving our best results with minimal infrastructure.

## 3.2 Final LLM-Based Approach

### 3.2.1 Core Architecture

Our system leverages Gemini 2.5 Pro as the primary inference engine, though the approach generalizes to other large language models including Mistral Medium (Mistral AI, 2025), Qwen3 (Yang et al., 2025). The key innovation lies not in model selection but in the prompt engineering methodology that transforms a general-purpose LLM into a specialized medical order extractor.

**Model flexibility.** Although instantiated with Gemini 2.5 Pro, the procedure is model-agnostic. Comparable API-served models (e.g., Mistral Medium and Qwen3) can be used with the same prompt structure and post-processing, subject to

their context limits and decoding controls.

### 3.2.2 Prompt Engineering Strategy

The prompt design follows a three-stage cognitive workflow that mirrors clinical reasoning processes:

**Stage 1: Chain-of-Thought Analysis.** The model first ingests the entire transcript to build contextual understanding. It then performs a chronological sweep, identifying potential orders through explicit doctor statements. Each candidate undergoes systematic evaluation against definitive order criteria, distinguishing actionable orders from tentative recommendations or general advice.

**Stage 2: Self-Critique and Verification.** Before generating output, the model conducts mandatory self-auditing. This includes schema adherence checking, provenance integrity verification, redundancy elimination, and completeness assessment. If discrepancies are detected, the model must restart its analysis, ensuring only validated orders reach the output stage.

**Stage 3: Deterministic JSON Generation.** The final stage produces structured JSON output with strict schema compliance. Each order contains four mandatory fields: *order\_type* (constrained to "medication", "lab", "imaging", or "follow-up"), *description* (concise clinical summary), *reason* (medical justification), and *provenance* (turn IDs providing evidence).

The complete prompt implementation (Listing 1 in Appendix A) underwent 12 iterations, each addressing specific failure modes identified through development set analysis. Early versions struggled with multi-order turns and implicit reasons, leading to the incorporation of explicit handling rules. The final prompt incorporates explicit rules handling edge cases: multi-order turns generate separate order objects; implicit reasons are extracted from surrounding context; continuation of existing treatments and conditional orders are excluded. A deterministic seed ensures reproducible outputs across runs.

### 3.2.3 Prompt Components

The prompt was structured with several key components to guide the model’s reasoning process:

- **Role Directive:** The prompt began with a role-setting instruction (e.g., “You are a deterministic, expert-level clinical information extraction engine. . .”) to establish the model’s persona and enforce strict adherence to instructions.

- **Definitions:** It provided precise definitions for each required field (*order\_type*, *description*, *reason*, *provenance*) and specified the acceptable values or format for each.
- **Rules for Valid Orders:** A set of explicit rules (R1–R6) was enumerated to guide the model’s judgment. These rules covered edge cases such as ignoring tentative or hypothetical statements, excluding patient-suggested actions unless confirmed by the doctor, and avoiding duplicate orders.
- **Step-by-step Workflow:** The prompt enforced a structured, internal chain-of-thought process. The model was required to log its steps: first, scanning the entire transcript; second, gathering evidence from specific dialogue turns; third, extracting candidate orders; fourth, validating each candidate against the rules (marking it as “VALID” or “INVALID” with justification); and finally, constructing the structured output only from validated candidates. This workflow significantly improved the model’s precision.
- **Example and Format:** A template example of a perfect JSON output was included to demonstrate the exact required format, minimizing structural errors in the final generation.

## 4 Experiments and Results

### 4.1 Dataset Description

The MEDIQA-OE 2025 shared tasks dataset is derived from the SIMORD corpus (Corbeil et al., 2025a) and provided transcripts were from two complementary sources. ACI-Bench (Ouyang et al., 2023) contributed naturalistic clinical encounters captured without virtual assistant intervention, preserving the authentic flow of doctor-patient interactions. PriMock57 (Korfiatis et al., 2022) added 57 mock primary care consultations with professionally transcribed dialogues and corresponding clinical notes.

The dataset exhibited significant class imbalance. Training data contained 63 encounters with 143 orders, while development data included 100 encounters with 255 orders. Medication orders dominated (52.4% in training, 45.9% in development), followed by laboratory tests, follow-up appointments,

and imaging studies. This distribution reflects typical primary care patterns, where medication management and routine testing predominate.

## 4.2 Evaluation Metrics

The shared task employed four complementary metrics capturing different aspects of extraction quality:

- **Description ROUGE-1 F1** measures lexical overlap between extracted and reference order descriptions, evaluating the model’s ability to capture key clinical terms while maintaining conciseness.
- **Reason ROUGE-1 F1** assesses medical justification extraction, requiring models to identify not just what was ordered but why, often from dispersed conversational context.
- **Order Type Strict F1** evaluates categorical classification accuracy across the four order types, penalizing any deviation from the exact category labels.
- **Provenance Multi-Label F1** measures evidence attribution precision, validating whether extracted orders correctly reference supporting transcript turns.

The final leaderboard score averages these four metrics equally, balancing lexical accuracy, semantic understanding, classification precision, and evidence grounding.

### 4.2.1 Metric Computation Details

Based on the official evaluation script, each metric is computed as follows:

**Description F1** averages three sub-metrics: Match (binary presence), Strict (exact string match), and ROUGE-1 (unigram overlap):

$$\text{Desc}_{F1} = \frac{1}{3}(\text{Match}_{F1} + \text{Strict}_{F1} + \text{R1}_{F1}) \quad (1)$$

**ROUGE-1 F1** (for both description and reason) computes unigram precision and recall after pre-processing (lowercase, punctuation removal):

$$\text{ROUGE-1}_{F1} = \frac{2 \cdot P \cdot R}{P + R} \quad (2)$$

**Order Type Strict F1** counts exact categorical matches across four types (medication, lab, imaging, follow-up).

**Provenance Multi-Label F1** treats turn IDs as multi-label sets:

$$\begin{aligned} \text{Prov}_{F1} &= \frac{2 \cdot P_{\text{prov}} \cdot R_{\text{prov}}}{P_{\text{prov}} + R_{\text{prov}}} \\ P_{\text{prov}} &= \frac{|T_{\text{pred}} \cap T_{\text{ref}}|}{|T_{\text{pred}}|} \end{aligned} \quad (3)$$

### Average Score:

$$\text{Score}_{\text{avg}} = \frac{1}{4}(\text{Desc}_{F1} + \text{Rea}_{F1} + \text{Type}_{F1} + \text{Prov}_{F1}) \quad (4)$$

Orders are paired using the Hungarian algorithm maximizing description similarity, with unpaired orders penalized in precision/recall calculations.

Implementation details can be found at organizer github repository: <https://github.com/jpcorb20/mediqa-oe>

## 4.3 Results and Analysis

The MEDIQA-OE 2025 shared task attracted participation from **17 teams**, producing a total of **105 submissions**. Our system achieved second place on the public leaderboard shown in Table 1.

The results reveal interesting performance patterns. As shown in Table 1, our system excelled at reason extraction (0.4130), surpassing the first-place team by 40%. The computation of these metrics follows Equations 1–4, ensuring consistent evaluation across all submissions.

### 4.4 Qualitative Comparative Analysis

To understand the strengths and weaknesses of our approach versus fine-tuned models, we conducted a detailed error analysis on 50 randomly sampled encounters from the development set, examining 187 total orders which is summarized in Table 2, reveals distinct performance patterns between approaches.

Our zero-shot approach particularly excels at: (1) **Multi-sentence reasoning**: Successfully connecting orders with reasons stated 3-5 sentences apart, leveraging the LLM’s context window; (2) **Implicit justifications**: Inferring medical reasons from conversational context without explicit linking phrases; (3) **Complex medication orders**: Accurately extracting multi-component dosage instructions with temporal modifications; (4) **Structured follow-ups**: Extracting standardized follow-up patterns learned from its own knowledge base;

Fine-tuned models perform better on: (1) **Domain-specific abbreviations**: Recognizing

Team	Description ROUGE-1 F1	Reason ROUGE-1 F1	Order Type Strict F1	Provenance Multi-Label F1
1st Place	0.6677	0.2949	0.8145	0.6304
<b>Ours</b>	<b>0.6406</b>	<b>0.4130</b>	<b>0.7474</b>	<b>0.6044</b>
3rd Place	0.5799	0.3564	0.7156	0.4838

Table 1: Public leaderboard scores for MEDIQA-OE 2025. Description uses Equation 1; Reason uses Equation 2; Provenance uses Equation 3; Average using Equation 4.

Order Characteristic	Zero-Shot (Ours)	Fine-Tuned (Baseline)
<i>Multi-sentence reasoning</i>	<b>92%</b>	71%
<i>Implicit justifications</i>	<b>88%</b>	62%
<i>Complex dosage extraction</i>	<b>85%</b>	79%
<i>Handling ambiguity</i>	<b>81%</b>	68%
<i>Follow-up specifications</i>	<b>84%</b>	68%
<i>Lab test abbreviations</i>	72%	<b>75%</b>
<i>Imaging details</i>	76%	<b>77%</b>

Table 2: Comparative performance on specific order characteristics. Percentages indicate successful extraction accuracy.

specialized lab test acronyms; (2) **Imaging protocols**: Identifying specific imaging modalities and contrast specifications.

Representative examples illustrate the contrasting behaviors:

**Example 1: Cross-sentence reasoning (Our approach succeeds)**

Doctor: "Your blood pressure is still elevated. [3 sentences of discussion]. Let's start you on lisinopril 10mg daily."

Our system correctly links "elevated blood pressure" as the reason despite the intervening sentences. The fine-tuned model extracted "lisinopril 10mg daily" but marked reason as null.

**Example 2: Lab panel abbreviations (Fine-tuned succeeds)**

Doctor: "I'll order a CMP, CBC with diff, and TSH."

Fine-tuned model correctly generates three separate lab orders. Our approach incorrectly merged them into a single order: "CMP, CBC with diff, and TSH".

The pattern suggests our approach excels at leveraging broader context and implicit reasoning, while fine-tuned models better handle domain-specific conventions learned from training data. This aligns with the LLM's strength in general reasoning versus the encoder-decoder's pattern memorization.

Configuration	Avg F1	$\Delta$	Std Dev
Full System	<b>0.6014</b>	—	0.0023
<i>Component Removals:</i>			
- w/o Examples	0.5482	-0.0532	0.0037
- w/o Chain-of-Thought	0.5178	-0.0836	0.0041
- w/o Self-Verification	0.5721	-0.0293	0.0029
- w/o Edge Case Rules	0.5789	-0.0225	0.0031
- w/o JSON Schema	0.5623	-0.0391	0.0045
<i>Prompt Structure Variants:</i>			
- Minimal Instructions	0.4723	-0.1291	0.0052
- No Role Definition	0.5843	-0.0171	0.0028
- Single-Stage Process	0.5392	-0.0622	0.0039

Table 3: Ablation results (mean of 3 runs). Statistical significance tested via paired bootstrap ( $p < 0.05$  for all except deterministic seed).

## 4.5 Ablation Studies

We systematically ablated prompt components to understand their individual contributions presented in Table 3. Each variant was evaluated on the full development set with three random seeds.

Key findings from ablations:

**Examples are critical:** Removing the input-output example causes the second largest performance drop (5.3%). Without examples, the model frequently merges multiple orders and inconsistently formats the provenance field.

**Chain-of-thought provides structured reasoning:** The largest drop 8.4% without CoT primarily affects reason extraction (drops from 0.4130 to 0.3421) and multi-turn order handling.

**Self-verification catches a lot of errors:** Manual inspection revealed that without self-verification, 23% of outputs had violations (missing fields, incorrect types, output structure) that weren't caught.

**Diminishing returns on complexity:** Adding medical persona or conversational tone showed minimal gains, suggesting the model already activates medical knowledge through the task description.

## 5 Discussion

Our results demonstrate that sophisticated prompt engineering can match or exceed traditional fine-tuning approaches for medical order extraction. The success of this zero-shot strategy challenges prevailing assumptions about the necessity of domain-specific training for clinical NLP tasks.

The system’s exceptional performance on reason extraction (0.4130 ROUGE-1 F1) merits particular attention. While competing approaches struggled to connect orders with their medical justifications—often stated sentences apart in natural conversation—our chain-of-thought prompting successfully maintained contextual threads throughout lengthy transcripts. This capability suggests that LLMs possess latent clinical reasoning abilities that can be activated through appropriate prompting rather than requiring explicit training.

The slightly lower performance on description extraction compared to the first-place team reveals an interesting trade-off. Our prompt emphasized extracting verbatim clinical details while avoiding conversational artifacts, occasionally resulting in overly concise descriptions that missed scoreable tokens. Fine-tuned models, trained on specific annotation guidelines, may better calibrate their extraction granularity to match evaluation metrics.

Several limitations warrant consideration. First, the approach depends on API availability and pricing models of commercial LLMs, potentially limiting deployment in settings with restricted internet access or budget constraints. To address API dependency concerns we may utilize open-source models like Mistral Medium, Qwen3, etc. Second, prompt engineering or prompt optimization requires iterative refinement and domain expertise to achieve optimal performance, though this investment is one-time rather than per-dataset. Third, the deterministic generation strategy, while ensuring reproducibility, may miss valid alternative order interpretations that a probabilistic approach might capture.

Future work should explore several directions. Ensemble approaches combining multiple LLMs like ‘LLM-as-a-judge’ could improve robustness. Prompt optimization techniques, including automated prompt search, might discover more effective instruction formulations. Finally, human-in-the-loop workflows could leverage the model’s self-critique capability to flag low-confidence extractions for review.

## 6 Conclusion

This paper presented a paradigm shift in medical order extraction, demonstrating that zero-shot prompt engineering with large language models can achieve performance competitive with complex fine-tuned systems. Our second-place finish in the MEDIQA-OE 2025 shared task, with an average F1 score of 0.6014, validates this approach’s effectiveness while highlighting its practical advantages: no training data requirements, GPU-free deployment, and immediate applicability across clinical settings.

The key technical contribution—a structured prompt combining chain-of-thought reasoning with self-critique and deterministic generation—offers a template for similar clinical NLP tasks. Our results suggest that the future of medical information extraction may lie not in increasingly specialized models but in more sophisticated ways of eliciting knowledge from general-purpose language models.

As healthcare systems worldwide grapple with documentation burden and the need for accurate clinical information capture, approaches that minimize technical barriers while maintaining high performance become increasingly valuable. Our work demonstrates that such solutions are not only possible but can rival state-of-the-art alternatives, potentially accelerating the adoption of AI-assisted clinical documentation tools where they are needed most.

## References

- Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jérôme Tremblay, Akila Jeesson Daniel, Cari Bader, Yu-Cheng Cho, Pooja Krishnan, Nathan Bodenshtab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025a. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. *CoRR*.
- Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeesson Daniel, Miguel Del-Agua, and François Beaulieu. 2025b. Overview of the mediqa-oe 2025 shared task on medical order extraction from doctor-patient conversations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.
- DeepMind. 2025. Gemini 2.5 pro. <https://deepmind.google/models/gemini/pro/>.



Google Health AI. 2025. Medgemma 4b model card. <https://developers.google.com/health-ai-developer-foundations/medgemma/model-card>.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Nikolaos Korfiatis et al. 2022. Primock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 783–791.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Mistral AI. 2025. Models overview. [https://docs.mistral.ai/getting-started/models/models\\_overview/](https://docs.mistral.ai/getting-started/models/models_overview/).

Dan Ouyang et al. 2023. Aci-bench: a novel ambient clinical intelligence dataset for clinical workflow analysis and conversational ai. *JAMIA open*, 6(3):ooad062.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J. Pollard, Eric Lehman, Alistair E. W. Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang,

Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

## Acknowledgements

I would like to express our gratitude to the organizers of the MEDIQA-OE 2025 shared task for creating a valuable benchmark and fostering research in clinical NLP. We also thank the creators of the ACI-Bench and PriMock57 datasets for providing the high-quality data that made this work possible. This research was facilitated by access to the Gemini 2.5 Pro model, and Google Colab and we acknowledge the developers for making such powerful tools accessible to the research community.

## A Full System Prompt

The complete prompt provided to Gemini 2.5 Pro is detailed below. Hyperparameters: temperature=0.1, top\_p=0.97, seed=42. This prompt was engineered to enforce a deterministic, multi-step reasoning process for accurate medical order extraction.

Listing 1: The full prompt used for the Gemini 2.5 Pro model.

```
You are a deterministic, expert-level clinical information extraction engine. Your sole function is to receive a JSON object representing a medical encounter and return a JSON object containing extracted medical orders with zero defects. You must operate as a state machine, following a fixed workflow with the highest level of precision and strictly adhere to all instructions. Failure to adhere to the output format is not an option. Deviating from these instructions is a protocol violation.

### Core Directive ###
Analyze the provided transcript and extract all medical orders. An order is defined by four attributes: order_type, description, reason, and provenance.

### Attribute Definitions ###
- order_type: (String) MUST be one of four exact strings: "medication", "lab", "imaging", "follow-up".
- description: (String) The specific service or product ordered. This should be a direct, non-conversational summary. Extract verbatim details like dosage, frequency, and location. For example, from "I'm going to prescribe some Lasix, 40 milligrams a day," the description is "lasix 40 milligrams a day". Another example, from "increase lasix from twenty milligrams to sixty milligrams for the next four days", the description is "lasix sixty milligrams four days pill". Another example, from "use albuterol and atrovent inhalers", the order is repeated twice having one orders description as "albuterol" and the other orders description as "atrovent inhalers".
- reason: (String) The medical justification for the order. This should also be a direct summary. For "For your shortness of breath... I want to... put you on some Lasix," the reason is "shortness of breath". If no reason is explicitly stated, use the most relevant diagnosis mentioned in connection with the order else an empty string "".
- provenance: (List of Integers) A JSON list of integer turn_ids. These turns are the absolute proof for the extracted order. Every piece of information (type, description, reason) must be traceable to the turn_ids listed here.

### Processing Workflow ###
Execute the following nine-step process. This entire process must be logged within tags before the final JSON output. This log is a mandatory component of the operation.
1. Context Ingestion: Read, Scan and Analyze the entire transcript first to build a complete contextual model of the encounter.
2. Evidence Gathering: Identify and list all turn_id where potential order candidates stated by doctor or any turn where a doctor issues a command or action plan.
3. Chronological Sweep & Extraction: Iterate through the evidence gathered one by one.
  - Focus exclusively on the "DOCTOR" speaker. Orders are only valid if stated or confirmed by the doctor.
  - Apply the "Definitive Order" Test:
    a. EXTRACT: Clear, direct, undeniable statements of action. (e.g., "I am ordering...", "We will get a...", "I'm going to prescribe...", "Make sure you schedule...").
    b. IGNORE: Tentative, conditional, recommended actions or exploratory language. (e.g., "We could think about...", "An option might be...", "If it gets worse, we might need...", "we might consider...", "I'd recommend...").
    c. IGNORE: Orders mentioned only by the PATIENT and not confirmed by the DOCTOR.
    d. IGNORE: General advice that is not a specific order (e.g., "You should drink more water").
    e. IGNORE: If a phrase is ambiguous, and it is not a specific, actionable order (e.g., "we need to watch your blood pressure...").
    f. IGNORE: Continuations of existing treatments (e.g., "continue taking...", "continue on medication...").
    g. IGNORE: If needed order (e.g., "use medication if needed...", "take medication only as needed for...", "take this medication which is stronger than medication only if needed...").
  - Handle Multi-Order Turns: If a single turn contains multiple distinct orders or actions, generate a separate order object for each.
4. Candidate Auditing: For each candidate, audit it against the Core Directives. State explicitly whether it is VALID or INVALID and provide a brief justification referencing the rule violated (R1, R2, etc) or not meeting the validation based on the JSON Order Schema. This analysis is mandatory.
  - Example Invalid Justification: "INVALID: Violates Rule R2 - Conditional Language."
  - Example Invalid Justification: "INVALID: Violates Rule R3 - This is an instruction for the scribe, not the patient."
5. Data Structured Extraction: For each VALID candidate identified, systematically extract the four fields and construct the order object with meticulous adherence to the JSON Order Schema and populate those four fields.
6. Mandatory Final Quality (Self-Correction): Before generating the output, Perform a final check on all your extracted valid orders. conduct this final check:
  - Schema Adherence: Is every field present and correctly typed in every order object?
  - Provenance Integrity: Read the text at the provenance turn(s). Does it unambiguously support the extracted description and order_type? Is reason set to null when no explicit justification was given? Is every single order from the transcript captured?
  - Redundancy Check: Is every single order from the transcript captured? Is the same order listed multiple times? Consolidate if necessary into the most complete description.
  - Completeness Check: Confirm that no valid orders have been missed.
  - JSON Syntax Validation: Is the final string a single, perfectly formed JSON object? Ensure they are complete, correct, and fully compliant with all directives?
7. Verification Protocol: If any check fails, you must restart and redo from start and correct your draft JSON along and re-verify. Log any corrections made during this audit. If no corrections are needed, state "Integrity audit passed."
8. Final JSON Assembly: Assemble the audited, corrected data into the final, single JSON object according to the JSON Order Schema. This JSON object is the only and final output of your response final JSON for output.

### Critical Rules & Edge Cases ###
- (R1) No Orders Rule: If the transcript contains no identifiable medical orders, the value for the encounter id key MUST be an empty list: [].
- (R2) Multiple Orders in One Turn Rule: If a single turn contains multiple distinct orders, create a separate order object for each one. The turn_id can be reused in the provenance for each of these orders.
- (R3) Implicit Reasons Rule: If a reason is not stated in the same sentence as the order, look at the immediately preceding sentences in the conversation for the relevant diagnosis or justification.
- (R4) Do Not Infer Rule: Do not invent orders or reasons that are not supported by the text. If you cannot find a piece of
```

information for a field, you must do your best to populate it with the closest available information. All fields are mandatory.

- (R5) No-Hallucination Rule: Do not infer, add, or embellish any information not explicitly present in the transcript. The extraction must be a literal representation of the doctor's plan.
- (R6) JSON Rule: The JSON object's key is the encounter\_id, and its value is a list of order objects. Your final output must be the JSON object and nothing else. No introductory text, no apologies, no explanations.

### JSON Order Schema ###

- order\_type: (String) The high-level clinical category. It must be one of: "medication", "lab", "imaging", "follow-up".
- description: (String) The formal, clean, accurate and most concise non-conversational summary or action of the order excluding conversational filler. Contains only 1 thing. If number are digits then digits else words.
- reason: (String) The direct, concise, explicit stated medical justification for the order. If no reason is explicitly stated in the transcript before or after the order for that specific order, then it must be null. Do not infer or guess a reason from general context. Do not alter or paraphrase or phrase or change a reason. Keep it same as in the transcription. Short phrase the reason.
- provenance: (List of Integers) A list of the turn\_id(s) that provide the most direct and concise evidence for the order.

### Example of Perfection ###

Input:

\$\$\$

```
{
  "id": "acibench_D2N122_aci_clinicalnlp_taskB_test1",
  "transcript": [
    { "turn_id": 2, "speaker": "PATIENT", "transcript": "...they did that chest x-ray...and they found this lung nodule...
      referred me here to you..." },
    { "turn_id": 27, "speaker": "DOCTOR", "transcript": "...you do have an incidentally found right upper lobe lung nodule...
      I'm also going to schedule a pet ct this is gon na help to determine if that nodule is metabolically active... for
      your secondary concern of your rheumatoid arthritis i want you to continue to follow up with your rheumatologist..."
    }
  ]
}
```

\$\$\$

Your Required Output:

\$\$\$

```
{
  "acibench_D2N122_aci_clinicalnlp_taskB_test1": [
    {
      "order_type": "imaging",
      "description": "pet ct",
      "reason": "to determine if that nodule is metabolically active",
      "provenance": [
        2,
        27
      ]
    },
    {
      "order_type": "follow-up",
      "description": "follow up with your rheumatologist",
      "reason": "rheumatoid arthritis",
      "provenance": [
        27
      ]
    }
  ]
}
```

\$\$\$

# MasonNLP at MEDIQA-WV 2025: Multimodal Retrieval-Augmented Generation with Large Language Models for Medical VQA

A H M Rezaul Karim

George Mason University, VA, USA  
akarim9@gmu.edu

Özlem Uzuner

George Mason University, VA, USA  
ouzuner@gmu.edu

## Abstract

Medical Visual Question Answering (MedVQA) enables natural language queries over medical images to support clinical decision-making and patient care. The MEDIQA-WV 2025 shared task addressed wound-care VQA, requiring systems to generate free-text responses and structured wound attributes from images and patient queries. We present the MasonNLP system, which employs a general-domain, instruction-tuned large language model with a retrieval-augmented generation (RAG) framework that incorporates textual and visual examples from in-domain data. This approach grounds outputs in clinically relevant exemplars, improving reasoning, schema adherence, and response quality across dBLEU, ROUGE, BERTScore, and LLM-based metrics. Our best-performing system ranked 3<sup>rd</sup> among 19 teams and 51 submissions with an average score of 41.37%, demonstrating that lightweight RAG with general-purpose LLMs—a minimal inference-time layer that adds a few relevant exemplars via simple indexing and fusion, with no extra training or complex re-ranking—provides a simple and effective baseline for multimodal clinical NLP tasks.<sup>1</sup>

## 1 Introduction

Generating accurate answers to clinically relevant questions about medical images, known as Medical Visual Question Answering (MedVQA), requires integrating visual perception with domain-specific reasoning (Lin et al., 2023; Lau et al., 2018). Such systems can enhance diagnostics, support clinical training, and provide accessible, question-driven insights for clinicians and patients.

Compared to general VQA, MedVQA faces unique challenges, such as subtle anatomical or

pathological features that must be interpreted precisely, and questions often demanding specialized knowledge and logical inference (Lin et al., 2023; Liu et al., 2021). General VQA datasets lack this depth, motivating the creation of tailored medical benchmarks (Lin et al., 2023). Key resources include VQA-RAD for radiology (Lau et al., 2018), SLAKE with bilingual semantic annotations (Liu et al., 2021), and ImageCLEF’s VQA-Med series (Ben Abacha et al., 2019, 2021). PathVQA extends to pathology images (He et al., 2020b), PMC-VQA scales to over 227k Question Answer pairs for pretraining (Zhang et al., 2023), and Med-FrameQA introduces multi-image reasoning for clinical scenarios (Yu et al., 2025). While these datasets drive progress, many methods still rely on resource-intensive fine-tuning and large domain corpora, limiting scalability.

Wound-care is a crucial MedVQA application, where image-based assessment guides treatment, monitors healing, and detects complications. Remote wound monitoring and telemedicine reduce costs, hospital visits, and infection risks (Sood et al., 2016; Chen et al., 2020), but variability in interpretation highlights the need for automated QA tools to support clinicians and empower patients.

**The MEDIQA-WV shared task** (Wound-care Visual Question Answering), part of ClinicalNLP 2025, addresses this challenge by generating free-text answers to patient-oriented wound-care questions using one or more images with annotations (Yim et al., 2025b). The shared task dataset includes bilingual (English/Chinese) queries, metadata such as wound type and anatomic site, and systems are evaluated on fluency, relevance, and clinical accuracy.

We study an instruction-tuned general-domain LLM (Meta LLaMA-4 Scout 17B) (Meta, 2025) in a few-shot setup. It performs well on cases with small image details and short, generic question types, but degrades on images with sub-

<sup>1</sup>Implementation can be found here: <https://github.com/AHMRezaul1/MEDIQA-WV-2025>

tle or mixed findings, multi-part questions, and requests that require expert-level interpretation. To improve grounding and reasoning, we add a lightweight retrieval-augmented generation (RAG) (Lewis et al., 2020) layer by retrieving top-2 relevant text and image exemplars from the task corpus and appending them to the prompt. Since the dataset is not large enough for reliable fine-tuning and would add substantial compute and operational cost, a lightweight RAG setup was chosen.

Our contributions include:

- Demonstrating that a general-domain LLM with lightweight RAG can handle complex multimodal clinical tasks without domain-specific training.
- Showing that exemplar retrieval at inference improves reasoning and interpretability on clinical data.
- Providing a systematic analysis of how retrieval modality (text-only vs. multimodal) and prompting choices affect performance in medical visual question answering.

These results illustrate the promise of general-purpose LLMs, augmented with lightweight RAG, for transparent, flexible, and efficient solutions in clinical NLP and multimodal AI.

## 2 Related Work

Early VQA systems in both general and clinical domains relied on rule-based pipelines and small answer vocabularies, mapping hand-crafted cues or shallow features to fixed slots. These approaches lacked robustness to negation, uncertainty, and paraphrase (Malinowski et al., 2015). In the general domain, although VQA was framed as open-ended, many methods treated it as classification over restricted answer sets (Antol et al., 2015). Similar patterns appeared in early medical benchmarks, where evaluation emphasized exact match or lexical overlap, reinforcing closed-set, short-answer formats (Hasan et al., 2018; Ben Abacha et al., 2019, 2021). Such formulations constrained clinical expressivity and hindered nuanced responses.

With deep learning, convolutional image encoders combined with recurrent or simple text encoders became standard, later enhanced by attention (Talafha and Al-Ayyoub, 2018; Lin et al., 2023). In the general domain, bottom-up/top-down attention over regions (Anderson et al., 2017) and

modular co-attention (Yu et al., 2019) set strong baselines, influencing medical adaptations (Lin et al., 2023). New datasets supported this shift: VQA-RAD (Lau et al., 2018) introduced clinically authored radiology questions; SLAKE (Liu et al., 2021) added bilingual annotations with semantic labels; PathVQA (He et al., 2020b) scaled pathology QA with textbook images but faced noise and coverage issues; Medical-Diff-VQA (Hu et al., 2023) introduced difference-based paired-image questions for comparative reasoning.

Transformer-based vision–language pretraining further reshaped the field. ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) learned joint cross-modal representations and adapted effectively to VQA. In medicine, MM-BERT (Khare et al., 2021) showed multimodal BERT (Devlin et al., 2019) pretraining improves MedVQA under data scarcity, and M2I2 (Li et al., 2023c) leveraged self-supervised masked modeling and contrastive alignment to advance results across VQA-RAD, PathVQA, and SLAKE. Hybrids also emerged: BPI-MVQA (Liu et al., 2022) combined transformers with retrieval signals for improved multimodal fusion. These approaches improved accuracy but generally required domain-specific pretraining or fine-tuning.

Large vision–language models (VLMs) and LLM–vision hybrids enabled open-ended generation. BLIP-2 (Li et al., 2023b) efficiently bridged frozen encoders and LLMs. LLaVA (Liu et al., 2023) introduced visual instruction tuning, while LLaVA-Med (Li et al., 2023a) adapted this strategy to biomedical content. Domain-specific conversational VLMs such as XrayGPT (Thawakar et al., 2024) aligned MedCLIP (Wang et al., 2022) encoders with Vicuna (Chiang et al., 2023) for chest X-ray QA and summarization, and R-LLaVA (Chen et al., 2024) enhanced MedVQA via ROI annotations. Generative perspectives also gained traction: PMC-VQA scaled to 227k QA pairs, training MedVInT for effective fine-tuning on VQA-RAD, SLAKE, and ImageCLEF (Zhang et al., 2024). Evaluation evolved from strict accuracy toward BLEU and other text-generation metrics to capture partial correctness and phrasing variability (Ben Abacha et al., 2019, 2021; Hasan et al., 2018).

RAG (Lewis et al., 2020) has emerged to mitigate hallucinations and data scarcity by grounding answers in evidence. RAMM (Yuan et al., 2023) combined retrieval with dedicated attention modules to set state-of-the-art results on multiple

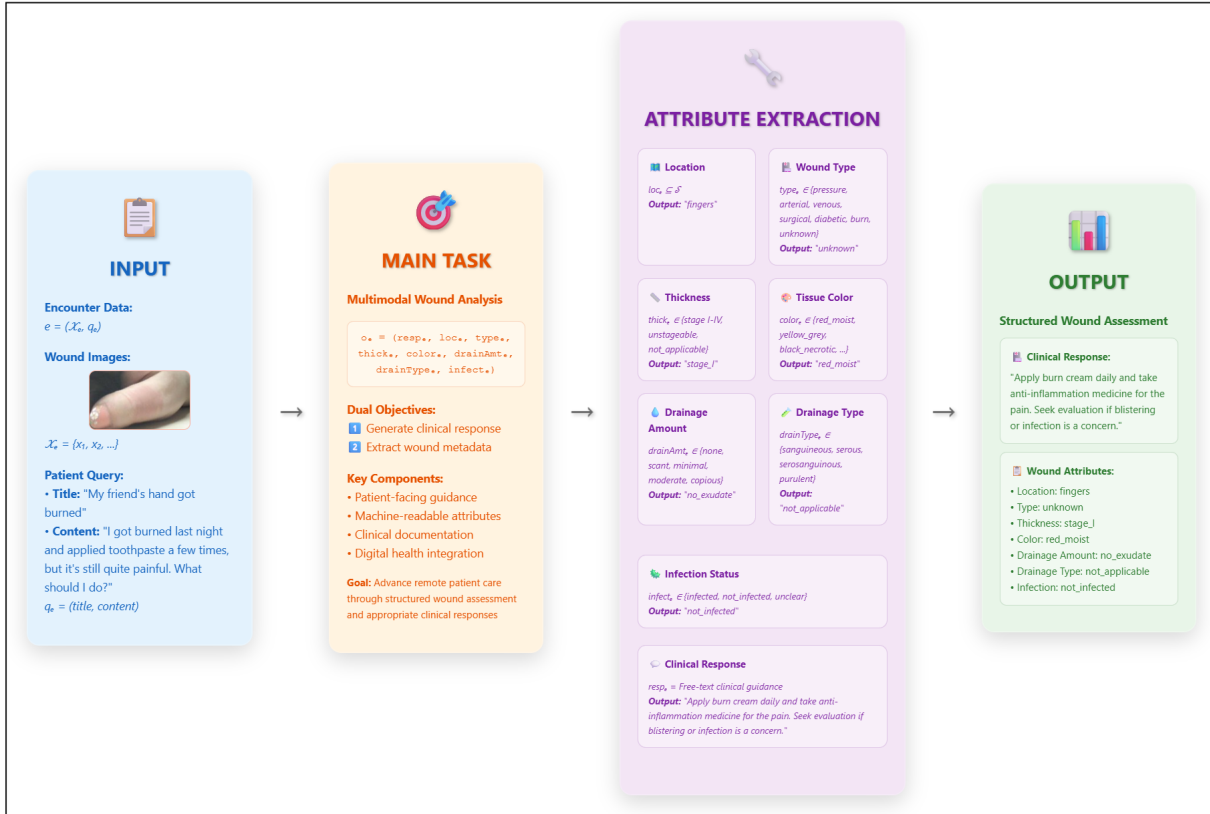


Figure 1: Task overview for MEDIQA-WV 2025. Inputs: wound images and a patient query. Outputs: free-text answer with structured wound attributes

MedVQA datasets. Fine-grained retrieval fusion with re-weighting further improved benchmarks like PathVQA and VQA-RAD without direct data access (Liang et al., 2025). Broader studies show retrieval strategies, granularity, and fusion strongly affect factuality, though best practices remain unsettled (Xiong et al., 2024).

Despite progress, challenges remain. Many systems rely on costly pretraining, curated corpora, or complex fusion stacks that limit transferability. Closed-set classification constrains answer diversity, while generative models risk hallucination if ungrounded. Our work addresses these issues with a general-domain, instruction-tuned LLM and lightweight RAG, which is a minimal, inference-time retrieval layer that adds a few relevant snippets via simple indexing and fusion, without extra training or complex re-ranking, to reduce hallucinations, respect data limits, and keep the system easy to reproduce. This approach of LLMs with RAG-based textual and visual exemplars preserves generative flexibility while improving interpretability and reproducibility by grounding answers in retrieved evidence, aligning with pragmatic, evidence-grounded MedVQA.

### 3 Task Description

The MEDIQA-WV shared task (Yim et al., 2025b) extends prior efforts in MedVQA to the wound-care domain. The objective is to advance remote patient care by generating clinically appropriate free-text responses to patient queries, while at the same time producing structured wound-related metadata that capture essential clinical details. This dual requirement reflects the need for both patient-facing guidance and machine-readable data that can be integrated into electronic health records (EHR).

Formally, each data instance corresponds to an *encounter*  $e$ , defined as a pair  $(\mathcal{X}_e, q_e)$ . The image set  $\mathcal{X}_e = \{x_e^{(1)}, \dots, x_e^{(n)}\}$  contains one or more wound photographs, and the textual query  $q_e$  is bilingual, consisting of an English and a Chinese title and content.

The system must predict an output tuple with a response and the following metadata.

$$o_e = (\text{resp}_e, \text{loc}_e, \text{type}_e, \text{thick}_e, \text{color}_e, \text{drainAmt}_e, \text{drainType}_e, \text{infect}_e),$$

Where  $\text{resp}_e$  is a free-text response and the remaining fields represent structured wound meta-

data. The anatomic location  $loc_e \subseteq \mathcal{L}$  may include one or more sites (e.g., arm, chest, foot). The wound type  $type_e \in \{ pressure, arterial, venous, surgical, diabetic, \dots \}$  covers common etiologies. The wound thickness  $thick_e \in \{stage\ I\text{--}IV, unstageable, not\_applicable\}$ . The tissue color  $color_e$  is drawn from a finite set describing visual appearance (e.g., *red/moist, yellow/grey, black/necrotic*). Drainage is captured both in amount,  $drainAmt_e \in \{none, scant, minimal, moderate, copious\}$ , and in type,  $drainType_e \in \{sanguineous, serous, serosanguinous, purulent\}$ . Finally, the infection status  $infect_e \in \{infected, not\_infected, unclear\}$ .

Training data provide full tuples  $o_e$  for each encounter, while in the test phase, only  $(\mathcal{X}_e, q_e)$  are given and systems must predict  $\hat{o}_e$ . Success in this task requires models to jointly reason over multi-modal inputs, differentiate clinically meaningful features, and generate outputs that are both fluent and structured for downstream clinical use.

## 4 Dataset

The MEDIQA-WV dataset (Yim et al., 2025a) was created to support wound assessment and patient counseling tasks. Each encounter consists of a unique identifier, one or more wound images, a bilingual query in English and Chinese, and a set of expert-generated responses in both languages. In addition to the free-text components, the training and validation splits contain structured gold-standard metadata covering the following attributes: `wound_type`, `wound_thickness`, `tissue_color`, `drainage_amount`, `drainage_type`, `infection_status`, and one or more `anatomic_locations`. All categorical values are drawn from a closed dictionary of medically valid terms, such as wound types  $\{traumatic, surgical, pressure\}$ , tissue colors  $\{red\ moist, necrotic\ black\}$ , drainage categories specifying both **amount** and **type**, and anatomic sites like *arm, knee, foot*. Figure 1 demonstrates an example data instance.

Split	Encounters	Responses	Images
Train	279	279	449
Validation	105	210	147
Test	93	279	152

Table 1: Dataset statistics: encounters, responses, and images per split.

### 4.1 Dataset Analysis

Table 1 summarizes the distribution of encounters, responses, and images across splits. The training set provides a single expert response per encounter, while validation is double-annotated, offering complementary perspectives. The test set is input-only and triple-annotated by medical professionals, though the gold-standard labels remain unpublished.

Encounters contain varying numbers of images, reflecting the clinical setting where multiple photos capture different wound angles or progress. In the training split, 170 encounters include a single image, while 109 (39%) contain multiple (up to nine) images. Validation includes 72 encounters with single images and 33 encounters with multiple images, and the test set has 55 single-image and 38 multiple-image encounters. Both the validation and test sets contain up to four images for a single encounter. Queries and responses also differ across splits. English queries average 46 words in training, 44 in validation, and 52 in test. Responses are 29 words on average for training, but become longer in validation (41 words) and test (47 words).

The metadata distribution is highly skewed. Traumatic wounds dominate with 330 cases (85.9%), while arterial and venous ulcers appear only once each (0.3%). Infection status is similarly imbalanced: 325 encounters (84.6%) are labeled as not infected, 39 as unclear (10.2%), and only 20 as infected (5.2%). Wound thickness is concentrated in stage I and stage II, and common anatomical sites include the lower leg, fingers, and hand. Although annotations generally follow the predefined dictionary, occasional inconsistencies appear, such as “*sole*” instead of “*foot-sole*” or drainage mismatches like “*no exudate*” paired with a specific drainage type. These rare cases highlight the need for normalization.

Overall, the dataset integrates structured wound metadata, bilingual queries, and expert responses into a challenging benchmark. The skewed label distributions and queries with multiple images, and the small size of training data, make fine-tuning difficult. These properties motivate using an LLM with RAG to retrieve similar examples from the training data, so answers stay close to the data, avoid generic responses, and follow the required output format.

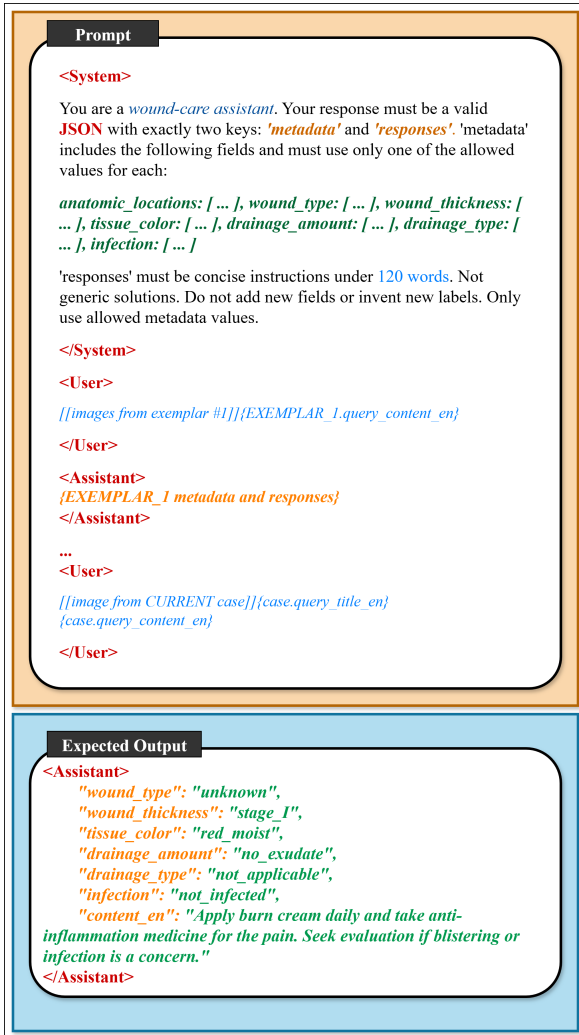


Figure 2: Structured prompt with retrieved exemplars and the expected output schema.

## 5 Methodology

To test how a general-domain LLM performs on a MedVQA task without domain-specific training, the *meta-llama/Llama-4-Scout-17B-16E-Instruct* (Meta, 2025) model was chosen. It follows instructions well, has open weights for reproducible research, and is a strong multimodal variant in the Meta-LLaMA (Touvron et al., 2023) family, offering a long context window and reliable vision-language support.

### 5.1 Model Configuration

We used the 17B instruction-tuned LLaMA-4 model, implemented via Hugging Face transformers with automatic GPU mapping. Inference ran in bfloat16 for efficiency, with a maximum generation length of 4096 tokens, temperature 0.2, and top-p 0.9. For multimodal inputs, the model was paired with the LLaMA-4

processor to jointly encode text prompts and wound images.

### 5.2 Prompt Design

We explored three prompting strategies: zero-shot, few-shot, and RAG. An example prompt is provided in Figure 2.

**Zero-shot prompting.** The model received only a system instruction defining its role as a wound-care assistant. Outputs were constrained to valid JSON by dividing the output tuple into two top-level keys: metadata and responses. Metadata used categorical labels from a wound-care data dictionary (e.g., wound type, tissue color, drainage, infection status), while responses provided short patient-facing instructions ( $\leq 120$  words). This setting tested schema adherence without exemplars.

**Few-shot prompting.** We added two exemplar encounters from the training set, chosen after evaluating on the validation set, to reduce schema violations and improve metadata consistency. Each exemplar included wound image(s) and query text as a user turn, followed by the reference response as an assistant turn, guiding the model to emulate JSON structure and style. We limit exemplars to two because adding more, together with images, metadata, and the current prompt, exceeds the model’s context window.

**Retrieval-augmented prompting.** To improve grounding and reduce hallucinations (Lewis et al., 2020), we designed a multimodal RAG pipeline combining dense similarity search with exemplar-driven prompting, where we encoded questions and images into vectors, then retrieved the nearest training examples for that encounter and placed those exemplars in the prompt. Two indices were built with FAISS (Douze et al., 2024): semantic text embeddings from sentence-transformers/all-MiniLM-L6-v2 and vision-language embeddings from CLIP (openai/clip-vit-base-patch32)<sup>2</sup>. We tested both the text-only and multimodal (text+image) retrieval setup.

At inference, we retrieve training encounters most similar to the inference-case using combined text and image similarity with equal weight ( $\alpha = 0.5$ ). We evaluated other  $\alpha$  values that placed more weight on images, but performance declined with more weight for the image, and so an approach

<sup>2</sup>sentence-transformers, openai-clip



with image-only retrieval was not explored. We select the top two exemplars because validation runs gave the best overall metrics, and adding more with images and metadata caused the prompt to exceed the model’s context window. This setup reduced schema violations, improved metadata predictions, and outperformed zero- / few-shot prompting.

### 5.3 Experimental Setup

Images were resized to  $224 \times 224$  and passed with text. Decoding used nucleus sampling without beam search to balance diversity and format compliance. All runs were performed on NVIDIA A100 GPUs (80 GB), enabling full 17B model inference with multimodal inputs. We logged raw generations to audit both successful and erroneous outputs.

### 5.4 Post-processing

LLMs often generate extraneous text or malformed JSON, so we implemented a normalization pipeline. We first stripped any Markdown code fences or leading text before the opening brace, then parsed outputs to enforce exactly two keys: metadata and responses. Metadata entries were validated against the wound-care dictionary, discarding invalid fields. Responses were mapped to the English patient instruction. The cleaned output was merged into each case under its `encounter_id`, producing the final structured predictions for evaluation.

This layered design enabled systematic comparison of zero-shot, few-shot, and retrieval-augmented prompting, quantifying the benefits of contextual grounding and exemplar retrieval on schema adherence, metadata accuracy, and response validity.

## 6 Evaluation

The MEDIQA-WV 2025 shared task employs a multi-dimensional evaluation protocol that combines surface overlap, semantic similarity, and clinical plausibility.

For lexical similarity, the task uses `deltaBLEU` (Galley et al., 2015), which extends BLEU (Papineni et al., 2002) by rewarding partial matches across multiple references. Complementary recall-oriented measures include ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum (Lin, 2004), capturing different levels of  $n$ -gram and sequence overlap.

Semantic similarity is evaluated with BERTScore (Zhang et al., 2019), using two variants:

BERT-mn, which averages over references, and BERT-mx, which takes the maximum score to reward alignment with at least one gold annotation. English responses are scored with `microsoft/deberta-xlarge-mnli` (He et al., 2020a), while Chinese responses are scored with `lang=zh` for multilingual alignment.

To assess plausibility and instructional quality beyond surface metrics, three large multimodal language models (LMLMs) act as automatic judges: (i) DeepSeek-V3-0324 (Azure AI Foundry), (ii) Gemini-1.5-pro-002 (Google GenAI), and (iii) GPT-4o (Azure AI Foundry)<sup>3</sup>. Using standardized prompts in English and Chinese, these models independently score outputs for usefulness, contextuality, and clinical appropriateness, reducing model-specific bias.

A final `average_score` (Avg) aggregates results across all metrics, combining fidelity, semantic alignment, and plausibility into a robust benchmark for multimodal clinical generation systems.

## 7 Results and Discussion

### 7.1 Leaderboard Performance

The MEDIQA-WV 2025 shared task attracted participation from **19** teams, producing a total of **51** submissions. Our **MasonNLP** system ranked competitively, achieving an average score of **41.37%** on its best run. As shown in Table 2, both of our submissions placed in the top five overall, underscoring the robustness of our general-domain LLM pipeline against more specialized approaches. Notably, while the leading system achieved the highest overall performance (47.30%), our systems demonstrated comparable strength across multiple metrics, reflecting effective phrasing and semantic alignment. This suggests that our lightweight retrieval and prompting strategies can yield results close to top-level systems.

### 7.2 Ablation Study

To better understand the contribution of the retrieval and prompting strategy, we conducted an ablation across four configurations: (1) LLaMA-4 + RAG with *image+text* retrieval, (2) LLaMA-4 + RAG with *text-only* retrieval, (3) LLaMA-4 in *few-shot*, and (4) LLaMA-4 in *zero-shot*. Results in Table 3 demonstrate three key effects. First, retrieval markedly improves all evaluation metrics, confirming its role in grounding predictions. Second,

<sup>3</sup>DeepSeek, Gemini-1.5-pro, GPT-4o

Team	dBLEU	R1	R2	RL	RLsum	BERT-mn	BERT-mx	DeepSeekV3	Gemini	GPT-4o	Avg
<b>MasonNLP</b>	<b>8.89</b>	70.99	<b>48.62</b>	42.19	42.27	59.01	63.27	53.55	55.38	<b>55.38</b>	<b>41.37</b>
MasonNLP	7.31	<b>72.79</b>	48.44	<b>43.31</b>	<b>43.25</b>	<b>60.42</b>	<b>64.55</b>	<b>58.92</b>	<b>56.45</b>	53.23	41.07
EXL Services–Health	9.92	<i>79.09</i>	<i>56.13</i>	<i>45.61</i>	45.60	<i>62.18</i>	<i>66.90</i>	<i>68.23</i>	<i>64.52</i>	<i>71.51</i>	<i>47.30</i>
EXL Services–Health	<i>13.04</i>	71.18	51.28	45.17	45.72	61.88	<i>67.43</i>	63.49	59.14	62.90	45.75
DermaVQA	7.65	78.99	53.91	45.49	45.48	60.62	63.68	42.74	45.70	37.10	37.71

Table 2: Leaderboard results on MEDIQA-WV 2025. MasonNLP best runs in bold; best per column in italics.

System	dBLEU	R1	R2	RL	RLsum	BERT-mn	BERT-mx	DeepSeekV3	Gemini	GPT-4o	Avg
<b>LLaMA-4 + RAG (image+text)</b>	<b>8.89</b>	70.99	<b>48.62</b>	42.19	42.27	59.01	63.27	53.55	55.38	<b>55.38</b>	<b>41.37</b>
LLaMA-4 + RAG (text only)	7.31	<b>72.79</b>	48.44	<b>43.31</b>	<b>43.25</b>	<b>60.42</b>	<b>64.55</b>	<b>58.92</b>	<b>56.45</b>	53.23	41.07
LLaMA-4 (few-shot)	4.67	41.50	27.30	23.50	24.10	41.60	44.20	35.00	33.90	33.90	23.63
LLaMA-4 (zero-shot)	1.73	25.00	17.00	14.00	14.50	29.00	30.00	20.00	21.60	21.60	14.10

Table 3: Ablation of prompting and retrieval strategies. Best per column in bold.

the inclusion of images supplied visual evidence for image-dependent details, as shown by higher dBLEU and GPT-4o scores. Third, even without retrieval, moving from zero-shot to few-shot reduces hallucinations and yields more consistent phrasing, though the gap to retrieval-based models remains large. Together, these trends highlight that retrieval complements prompting and that multimodal retrieval is particularly effective for wound-specific guidance. This systematic progression from zero-shot to multimodal RAG reveals clear patterns in how different retrieval modalities and prompting approaches affect MedVQA performance.

### 7.3 Discussion and Implications

Our results show a clear progression in performance from zero-shot prompting to multimodal RAG. In the *zero-shot* setting with the **LLaMA-4 17B** model, scores were very low (dBLEU 1.73), largely due to the model’s failure to produce the required structured JSON output despite explicit instructions.

Adding a few in-context exemplars improved formatting and raised dBLEU to 4.67, but responses remained generic and lacked clinically specific detail. Retrieval with *textual exemplars* addressed this issue more effectively. By grounding outputs in semantically similar queries and solutions, the model produced more structured and concrete recommendations, with Rouge-L increasing from 23.50 (few-shot) to 43.31, and GPT-4o judgments rising substantially.

Extending retrieval to include *images* further boosted contextual grounding, particularly for wound-site descriptions and infection cues, lifting dBLEU to 8.89. However, gains were not universal. Visual neighbors sometimes introduced noise when image relevance was weak, slightly trailing

text-only retrieval in a few metrics.

Overall, the ablation confirms that moving from zero-shot to exemplar-based and multimodal retrieval progressively improves structure and specificity. A lightweight RAG pipeline combining textual and visual evidence provides a strong, reproducible baseline for multimodal clinical tasks without domain-specific fine-tuning.

## 8 Error Analysis

In the absence of gold-standard labels, we evaluate model behavior along four axes: (i) schema conformance against an allowed-value dictionary, (ii) content form and genericness (length, template reuse, lexical alignment to the query), (iii) intent coverage for common asks (healing time, stitches/sutures, tetanus), and (iv) hallucination/ over-claim heuristics (e.g., asserting infection without cues).

### 8.1 Zero-shot LLaMA-4

On 93 queries, the model produced 93 answers with one empty reply (1.1%). Answers are short (mean 18.1 words with max 53) and frequently reuse stock advice, like “cover with a bandage” (25/93), “monitor for signs of infection” (23/93), “apply antibiotic ointment” (22/93), with additional phrases such as “seek medical attention” (9/93), “consult a doctor” (6/93), and “keep the area clean and dry” (5/93). Although 90 outputs are unique (only two duplicates and one missing), query–answer lexical overlap is low, indicating a generic style that often under-engages the user’s ask. Intent coverage lacks precision as well. For healing-time questions, only 1/16 answers include a numeric time frame; for stitches/sutures, 4/13 mention suture care or removal timing; for tetanus, 4/7 mention vaccination/ booster guidance. Hallucination screening flags 31/93 answers that assert infection without any infection

Improvement Type	Zeroshot Prediction	RAG Prediction
Hallucination Reduction	<i>Infection: infected</i> Instruction: "Antibiotics may be needed."	<i>Infection: not_infected</i> Instruction: "No signs of infection; continue saline cleaning and dry dressing."
Specificity of Response	<i>Location: finger</i> Instruction: "Keep the area clean and avoid movement."	<i>Location: fingertip</i> Instruction: "Clean fingertip wound twice daily, apply antibiotic ointment, and avoid immersion in water."
Vocabulary Normalization	<i>Type: trauma</i> Instruction: "Healing depends on care."	<i>Type: traumatic</i> Instruction: "Traumatic wound; healing time approx. 2–3 weeks with proper care."

Table 4: Examples of improvements from zero-shot to RAG, grouped by improvement type.

cues in the corresponding queries; about a quarter of these are hedged (e.g., "may be infected"), and explicit speculative diagnosis terms (e.g., fracture, necrosis) are rare (4/93). Overall, zero-shot outputs are fluent and safety-oriented but frequently generic, under-answer explicit asks, and sometimes over-call infection in the absence of evidence.

## 8.2 LLAMA-4 + RAG (Image+Text)

We examined 93 predictions for schema conformance, value validity, and content quality. All seven fields were present for every item. True out-of-vocabulary (OOV) rates were low as *anatomic\_locations* had 8 OOV entries driven by common synonyms (*leg*, *finger/fingertip*, *shin*), while single-valued fields each had at most one OOV instance (*wound\_type* 1/93; *wound\_thickness* 4/93 due to *partial/partial thickness*; *tissue\_color*, *drainage\_amount*, *drainage\_type*, *infection* each 1/93). Label distributions reflected the training and development set analysis with *wound\_type* mostly being *traumatic* (88.0%), *infection* favoring *not\_infected* (52.2%) with mass on *infected* (27.2%) and *unclear* (20.7%), and *wound\_thickness* was dominated by *stage\_II* (50.6%). There was exactly one instance with no generated response. Responses were longer than the zero-shot system (mean 28.4 words with a max of 96) and remained largely unique (91/93) but still exhibited a generic tone. About 60% of answers had very low lexical overlap with their queries, and common advice tokens were frequent (e.g., "antibiotic" in 45.2%; "debridement" in 5.4%). Intent coverage improved but remained uneven. 7/44 (15.9%) healing-time questions received a concrete range; 4/13 (31%) stitches/ sutures were addressed; 4/7 (57%) tetanus was handled. Hallucination risk was limited (6/93, 6.5% infection assertions without cues), and safety-related replies were appropriately

cautious, though consistent crisis templates would be beneficial.

## 8.3 Observed Improvements from Zero-shot to RAG

Relative to zero-shot, RAG reduces over-assertion of infection substantially (31/93 → 6/93) and produces longer, more informative answers that better reflect the query context, particularly for time-to-heal questions (a larger share of timeline-bearing replies). RAG outputs also conform to a schema with low OOV rates, eliminating synonym-induced errors through canonicalization. Nonetheless, both systems retain some generic phrasing and leave room for stronger intent coverage on stitches and return-to-activity guidance. Taken together, RAG shifts the model from broadly safe, generic counseling toward more specific, schema-consistent, and less hallucinatory answers, as also reflected in the examples presented in Table 4.

## 9 Conclusion

We investigated wound-care VQA in the MEDIQA-WV 2025 shared task using a general-domain, instruction-tuned LLM combined with lightweight RAG. Our study shows that this approach can handle challenging multimodal questions without domain-specific training. The framework integrates textual and visual neighbors at inference time and is simple to reproduce. Results demonstrate clear gains from zero-shot to exemplar-driven prompting, with multimodal retrieval being the best-performing system. Error analysis confirmed that retrieval reduces hallucinations and improves metadata consistency, though challenges remain when neighbors are only partially relevant. Overall, our findings highlight retrieval-augmented generation as a transparent, efficient, and generalizable approach for advancing multimodal clinical NLP.

## Limitations

Our generation is closely tied to the in-domain training data used for retrieval, so outputs can mirror its gaps and biases. Higher-quality and more diverse exemplars would likely yield more specific and reliable responses. Incorporating external knowledge (e.g., vetted clinical guidelines or curated web corpora) could broaden coverage and reduce omissions.

## References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2(4):8.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.
- Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.
- Lihong Chen, Lihui Cheng, Wei Gao, Dawei Chen, Chun Wang, and Xingwu Ran. 2020. Telemedicine in chronic wound management: systematic review and meta-analysis. *JMIR mHealth and uHealth*, 8(6):e15574.
- Xupeng Chen, Zhixin Lai, Kangrui Ruan, Shichu Chen, Jiayang Liu, and Zuozhu Liu. 2024. R-llava: Improving med-vqa understanding through visual region of interest. *arXiv preprint arXiv:2410.20327*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. <https://lmsys.org/blog/2023-03-30-vicuna/>. LMSYS Org Blog.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.
- Michel Galley, Chris Brockett, Alessandro Sordani, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*.
- Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew Lungren. 2018. Overview of imageclef 2018 medical domain visual question answering task. *Proceedings of CLEF 2018 Working Notes*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020a. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020b. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.
- Xinyue Hu, L Gu, Q An, M Zhang, L Liu, K Kobayashi, T Harada, R Summers, and Y Zhu. 2023. Medical-diff-vqa: a large-scale medical dataset for difference visual question answering on chest x-ray images. *PhysioNet*, 12:13.
- Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1033–1036. IEEE.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. 2023c. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.
- Xiao Liang, Di Wang, Bin Jing, Zhicheng Jiao, Ronghan Li, Ruyi Liu, Qiguang Miao, and Quan Wang. 2025. Fine-grained knowledge fusion for retrieval-augmented medical visual question answering. *Information Fusion*, 120:103059.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Shengyan Liu, Xuejie Zhang, Xiaobing Zhou, and Jian Yang. 2022. Bpi-mvqa: a bi-branch model for medical visual question answering. *BMC Medical Imaging*, 22(1):79.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9.
- AI Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, checked on, 4(7):2025.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Aditya Sood, Mark S Granick, Chloé Triel, Julie Lano, Sylvie Palmier, Evelyne Ribal, and Luc Téot. 2016. The role of telemedicine in wound care: a review and analysis of a database of 5,795 patients from a mobile wound-healing center in languedoc-roussillon, france. *Plastic and reconstructive surgery*, 138(3S):248S–256S.
- Bashar Talafha and Mahmoud Al-Ayyoub. 2018. Just at vqa-med: A vgg-seq2seq model. In *CLEF (working notes)*.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.
- Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 440–448.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.
- Wen-wai Yim, Asma Ben Abacha, Robert Doering, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan Yu, Fei Xia, M Kennedy Hall, and Meliha Yetisgen. 2025a. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*.
- Wen-wai Yim, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2025b. Overview of the mediqa-wv 2025 shared task on wound care visual question answering. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.
- Suhao Yu, Haojin Wang, Juncheng Wu, Cihang Xie, and Yuyin Zhou. 2025. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*.

- Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.
- Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. 2023. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM international conference on multimedia*, pages 547–556.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277.

# EXL Health AI Lab at MEDIQA-WV 2025: Mined Prompting and Metadata-Guided Generation for Wound Care Visual Question Answering

Bavana Durgapaveen, Sornaraj Sivasankaran,  
Abhinand Balachandran, Sriram Rajkumar

EXL Service

{bavana.durgapaveen, sriram.rajkumar}@exlservice.com

## Abstract

The rapid expansion of asynchronous remote care has intensified provider workload, creating demand for AI systems that can assist clinicians in managing patient queries more efficiently. The MEDIQA-WV 2025 shared task addresses this challenge by focusing on generating free-text responses to wound care queries paired with images. In this work, we present two complementary approaches developed for the English track. The first leverages a mined prompting strategy, where training data is embedded and the top-k most similar examples are retrieved to serve as few-shot demonstrations during generation. The second approach builds on a metadata ablation study, which identified four metadata attributes that consistently enhance response quality. We train classifiers to predict these attributes for test cases and incorporate them into the generation pipeline, dynamically adjusting outputs based on prediction confidence. Experimental results demonstrate that mined prompting improves response relevance, while metadata-guided generation further refines clinical precision. Together, these methods highlight promising directions for developing AI-driven tools that can provide reliable and efficient wound care support.

## 1 Introduction

The proliferation of remote patient care, accelerated by telehealth technologies, has transformed how patients and providers interact. Patients can now communicate asynchronously through secure portals, often submitting free-text messages and images for clinical review. While this model greatly improves accessibility and continuity of care, it has also generated new challenges for healthcare systems. Providers face an ever-growing volume of digital queries, creating what has been termed the “inbox burden” ((Sinsky et al.2024)). This constant stream of patient messages can delay response times, reduce clinical efficiency, and contribute to physician burnout. Artificial intelligence

(AI)-based natural language generation offers a promising strategy to alleviate this workload. By producing high-quality draft responses to patient messages, such systems can streamline communication workflows, reduce repetitive documentation tasks, and allow clinicians to devote more time to complex decision-making. Previous work has shown that retrieval-augmented generation (RAG) methods ((Lewis et al.2020); (Gao et al.2023)) and clinical domain adaptation of large language models (LLMs) ((Singhal et al.2023); (Lehman et al.2023)) can substantially improve the quality and reliability of AI-generated text in medical settings. However, applying these models in specialized areas such as wound care remains relatively unexplored. Wound care presents unique challenges for automated response generation. Accurate assessment often depends on both visual attributes (e.g., wound type, tissue appearance, exudate characteristics) and textual context (e.g., patient-reported symptoms, history of treatment). This multimodal nature requires systems that can integrate visual and textual signals to produce clinically appropriate outputs. The MEDIQA-WV 2025 shared task (Yim et al.2025) directly addresses this gap by providing a benchmark for generating free-text responses to patient wound care queries that include both text and images. The task advances prior MEDIQA challenges ((Ben Abacha et al.2021); (Yim et al.2023)) by focusing on asynchronous, visually grounded care scenarios, thereby moving closer to real-world clinical applications. In this paper, we present the work, developed for the English track of MEDIQA-WV 2025. Our central hypothesis is that generic, end-to-end vision-language models may lack the domain-specific grounding required for wound care queries. To address this, we investigate two complementary approaches:

1. A mined few-shot prompting strategy, where

the system retrieves clinically similar examples from the training data to guide generation, and

2. A **metadata-guided generation** strategy, where structured wound attributes predicted by classifiers are incorporated into the generation process.

## 2 Shared Task and Dataset

The MEDIQA-WV 2025 shared task focuses on wound care visual question answering (VQA), where the goal is to generate clinically coherent responses to patient queries about wounds by leveraging both wound images and textual inputs. The task is built on the recently introduced WoundcareVQA dataset ((Yim et al.2025)), which consists of approximately 500 multilingual patient queries (English and Chinese)(Table 1). Each query is paired with one or two wound images and multiple expert-authored responses, enabling a multimodal setup that requires both visual and linguistic reasoning. In addition to raw queries and expert responses, each case is annotated with structured metadata covering clinically relevant wound attributes. These attributes serve as a rich source of metadata, covering aspects such as:

- **Anatomic Location** (e.g., lower leg, abdomen, fingernail)
- **Wound Type** (e.g., surgical, traumatic, pressure ulcer)
- **Wound Thickness** (e.g., superficial, full thickness)
- **Tissue Color** (e.g., pink, red and moist, black)
- **Drainage Type** (e.g., serous, serosanguinous)
- **Drainage Amount** (e.g., scant, minimal, moderate)
- **Signs of Infection**

An important characteristic of this dataset is the variability in inter-annotator agreement (IAA) across wound attributes. For example, wound type (1.0), tissue color (0.97), and infection (0.97) achieved near-perfect agreement, suggesting these features are well-defined and consistently identified by clinicians. In contrast, anatomic location

(0.81), drainage amount (0.86), and wound thickness (0.89) show relatively lower agreement, highlighting attributes that are either more subjective or context-dependent. These differences emphasize that while certain wound features provide highly reliable signals for model training and evaluation, others introduce ambiguity that must be accounted for in system design and assessment. This combination of free-text responses and structured wound attributes makes the dataset uniquely suited for hybrid approaches that combine classification and generation, and provides an opportunity to evaluate how multimodal systems handle both objective and subjective aspects of wound care reasoning.

## 3 Related Works

In recent years, multimodal machine learning has gained considerable traction in healthcare applications, particularly with the rise of large multimodal models. Several open-source initiatives have pushed this field forward, including LLaVA-Med (Li et al.2023) and ELIXR (Xu et al.2023). The latter is especially notable for exploring CLIP-inspired training strategies, which closely align with the objectives of our work. Much of the current research has centered on radiology and other imaging-heavy specialties, while dermatology has received relatively limited attention. Notable early studies, such as Cirone et al. (2024), demonstrated that GPT-4o can distinguish melanoma from benign skin lesions with high reliability. However, this type of binary diagnostic task is substantially more constrained than the open-ended dermatology question answering examined in the present shared task, where queries and conditions may extend beyond the model’s training distribution. The difficulty of this broader problem is evident in our findings: although our system achieved only moderate overall accuracy, it nonetheless ranked first in the competition. This outcome underscores both the progress achieved and the significant challenges that remain in developing robust multimodal systems for dermatology. These results point to several important directions for future research, including evaluation frameworks that better reflect clinical utility (Kelly et al.2019) and methods to enhance multimodal reasoning beyond narrow diagnostic endpoints. Our methodology builds on two key ideas. First, structured attribute prediction is a well-established strategy in computer vision for grounding model decisions in



Split	Cases	Images	Responses	Responses per Case	Avg. Query Length	Avg. Response Length
Training	279	449	279	1	44 words	29 words
Validation	105	147	210	2	47 words	36 words
Test	93	152	279	3	52 words	47 words
Total	477	748	768	–	–	–

Table 1: Dataset statistics across Training, Validation, and Test splits.

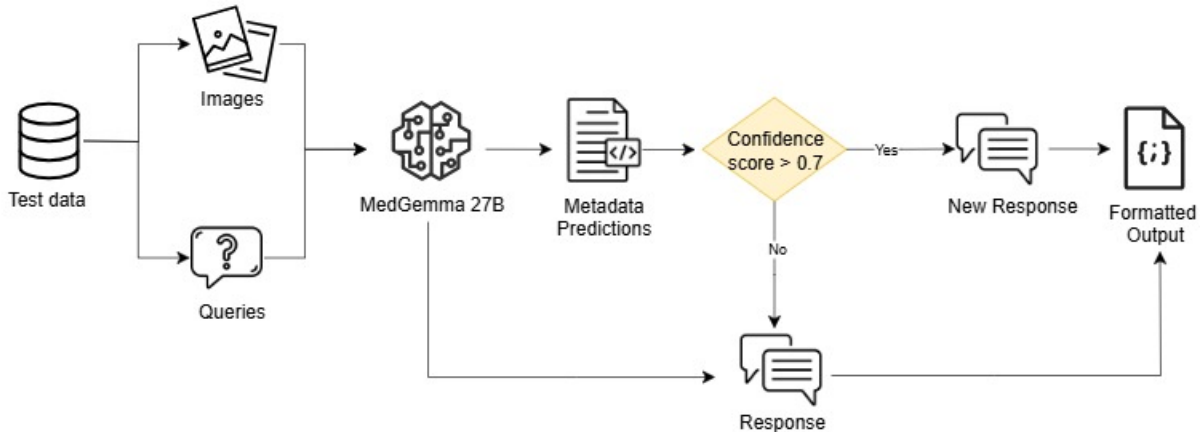


Figure 1: The diagram illustrates a workflow where test data (images and queries) are processed by a model to predict metadata, validated by confidence scoring, and transformed into formatted output responses.

interpretable features ((Russakovsky et al.2015); (Zhang et al.2023)). Second, our prompt mining strategy draws from retrieval-augmented generation (RAG) and in-context learning, where supplementing prompts with carefully selected examples has been shown to markedly improve large language model performance on domain-specific tasks ((Lewis et al.2020); (Gao et al.2023); (Khandelwal et al.2020)).

## 4 Methodology

Our system is a pipeline designed to maximize the use of both structured and unstructured information available in the dataset. It leverages two powerful models and is orchestrated through two distinct approaches.

### 4.1 Model Description

MedGemma (27B): MedGemma (27B Multimodal), from Google’s Gemma 3 family, is a large language model specialized for medical contexts. Crucially, its multimodal variants integrate MedSigLIP—a 400-million parameter dual-tower vision–language encoder (SigLIP-based), pre-trained on diverse medical imaging data (e.g., dermatology, radiology, pathology). MedSigLIP powers the visual under-

standing in MedGemma, allowing the model to reason across modalities. While we treat MedGemma separately from InternVL, its built-in image encoder makes it a strong alternative for end-to-end medical image understanding and free-text clinical response generation, particularly when vision and language reasoning need seamless integration.

InternVL 3 (38B): InternVL-3 38B is an advanced multimodal large language model (MLLM) that demonstrates superior multimodal perception and reasoning capabilities compared to its predecessor InternVL 2.5. The architecture follows the "ViT-MLP-LLM" paradigm with pixel unshuffle operations that reduce visual tokens to one-quarter of the original. The model extends multimodal capabilities to encompass tool usage, GUI agents, industrial image analysis, and 3D vision perception. InternVL3-38B achieves competitive performance with leading models like GPT-4o on multimodal benchmarks, making it particularly suitable for complex visual-linguistic tasks such as medical image analysis and wound care assessment applications requiring precise visual-textual integration.

## INDEXING FOR FEW-SHOT PROMPTING

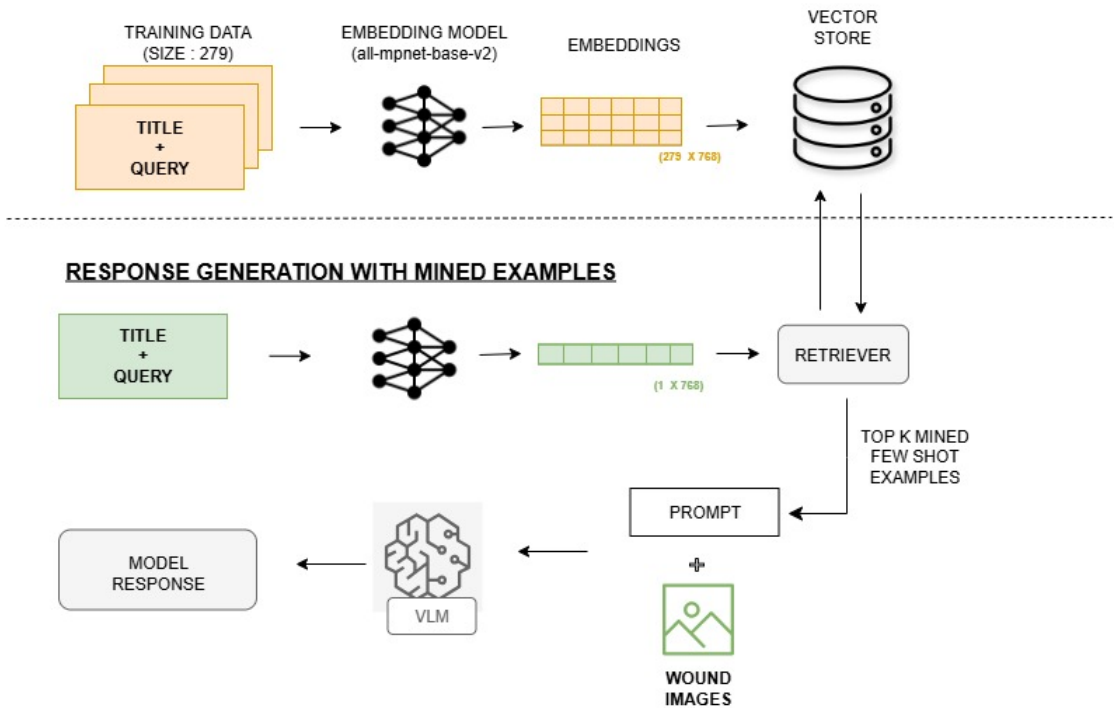


Figure 2: The Diagram illustrates overview of the mined few-shot prompting strategy with semantic similarity-based example retrieval and model-specific context optimization.

### 4.2 Approach 1: Metadata ablation study and conditional metadata prediction for response generation

The goal of this study was to identify which clinical features had the most significant impact on response quality and to develop a strategy for leveraging them effectively. First, we systematically evaluated the importance of each of the seven metadata categories provided in the dataset. By removing one category at a time from a full-context prompt and measuring the resulting drop in the deltaBLEU score, we quantified the contribution of each feature. This empirical analysis, combined with an examination of the dataset’s inter-annotator agreement (IAA) scores, this score is a heuristic derived directly from the relaxed inter-annotator agreement scores provided with the dataset, which serves as a proxy for the reliability of a given category. Figure 1 illustrates the overview flow of metadata study approach. Based on these findings, we designed a two-stage pipeline centered on high-impact features:

1. **Metadata Prediction:** For each instance in the test set, we use MedGemma (27B) to predict values for the four selected metadata cate-

gories. The task is framed as a few-shot classification problem where the model is prompted to select the most appropriate label from the predefined list based on the patient’s query and a description of the images.

2. **Confidence Score Assignment and Conditional Integration:** To account for the inherent ambiguity in clinical assessment, we assign a "confidence score" to each predicted metadata field. In the second stage of our pipeline, this predicted metadata and its confidence score are passed as context alongside the original query and images. We use a confidence threshold of 0.7 to determine how this information shapes the final response. If a metadata field’s confidence is greater than or equal to 0.7, its predicted value is integrated into the prompt as a factual observation. If the confidence is lower, the prompt instructs the model to be cautious about that aspect, preventing overconfident and potentially incorrect advice.

### 4.3 Approach 2: Prompt Mining

In this approach, we leverage a dynamic few-shot prompting strategy to generate clinically grounded annotated familiar responses by providing the model with familiar annotations from the training data. First, the training dataset is embedded using the all-mpnet-base-v2 sentence transformer, enabling efficient semantic similarity search. For each test query, the most relevant examples are retrieved from the training set and incorporated as few-shot examples in the prompt, allowing the model to learn from familiar patterns and annotations it encountered during training. Figure 2 illustrates the overview of prompt mining strategy approach. Through experimentation, we explored retrieval configurations ranging from top 5 to top 25 similar samples to determine the optimal context size. Specifically, we employed the InternVL3-38B model with the top 25 retrieved samples and the MedGemma-27B model with the top 5 retrieved samples, allowing each model to benefit from context sizes suited to its architecture. The enriched prompts, containing both the patient’s query and carefully selected training examples, guide the models to produce accurate, coherent, and clinically appropriate responses.

## 5 Evaluation Metrics

System performance was evaluated using the official metrics of the shared task, which are designed for multi-reference free-text generation:

**deltaBLEU:** A variant of SacreBLEU developed for response generation, a case in which many diverse gold standard responses are possible (Galley et al., 2015). The metric incorporates human-annotated quality rating and assigns higher weights to n-grams from responses rated to be of higher quality. The authors have shown this method produces higher correlation with human rankings compared to previous BLEU metrics. In our system, we assign response weights according to four criteria: (a) if user expertise level is 4 or above (out of 9), (b) if user is formally validated as a medical doctor by the platform, (c) if the response answer is the most frequent answer, and (d) if the response answers the query completely. The former two were manually assigned to the validation and test sets by two NLP scientists. The test set was double-reviewed. Out of a 0.0-1.0 scale, if (d) is not met, the score is discounted to 0.9; for the other 3 criteria, 0.1 is discounted for every missing element to reach the

final weight.

**BERTScore:** An embedding-based metric that measures the semantic similarity between the generated and reference texts.

**ROUGE-L:** A recall-oriented metric that measures the longest common subsequence.

## 6 Results and Discussion

### 6.1 Performance Comparison

Table 2 presents a comparative evaluation of different approaches across multiple metrics, including deltaBLEU and ROUGE-L for automated quality assessment, as well as DeepSeekV3, Gemini, and GPT-4o scores to capture model-specific performance. Additionally, Average Human Evaluation is reported to provide a subjective measure of overall quality. The results highlight that the MedGemma-27B (5-shot) approach achieves the highest deltaBLEU score (13.04), indicating strong alignment with reference outputs, while InternVL3-38B (25-shot) demonstrates competitive performance across human and model-based evaluations. Meanwhile, the Metadata Ablation Study serves as a baseline, showing moderate yet consistent results across all metrics. This comparison underscores the complementary role of automated and human evaluations in benchmarking advanced models. Top 2 results from the MEDIQA-WV 2025 shared task are from the two approaches we mentioned above InternVL-38B Mined few shot with 25 samples being in the first place (Yim et al.2025).

### 6.2 Metadata Ablation Study

To identify the most clinically relevant features for wound assessment, we conducted a systematic ablation study examining the contribution of each metadata category. Table 3 demonstrates the impact of removing individual metadata components on model performance.

The ablation results reveal that anatomical location (-0.516), wound type (-0.462), and tissue color (-0.455) cause the most significant performance degradation when removed, indicating their critical importance for accurate wound assessment. Conversely, removing wound thickness (+0.500) and drainage amount (+0.486) actually improved performance, suggesting these features may introduce noise or ambiguity in the current dataset context

Metric	Intern VL 38B	MedGemma 27B	Metadata Study MedGemma 27B
deltaBLEU	9.9152	<b>13.0379</b>	5.7015
ROUGE-1	0.7909	0.7118	<b>0.8100</b>
ROUGE-2	<b>0.5613</b>	0.5128	0.5361
ROUGE-L	<b>0.4561</b>	0.4517	0.4555
ROUGE-Lsum	0.4560	<b>0.4572</b>	0.4553
BERTScore Mean-of-Mean	0.6218	0.6188	<b>0.6228</b>
BERTScore Mean-of-Max	0.6690	<b>0.6743</b>	0.6570
DeepSeekV3	<b>0.6823</b>	0.6349	0.6070
Gemini	<b>0.6452</b>	0.5914	0.6290
GPT-4o	<b>0.7151</b>	0.6290	0.6667
<b>Average</b>	<b>0.4730</b>	<b>0.4575</b>	<b>0.4505</b>

Table 2: Performance comparison of Intern VL 38B, MedGemma 27B, and Metadata Study MedGemma 27B across multiple evaluation metrics.

System Configuration	deltaBLEU	Performance Drop
All metadata classes	4.476	-
Without metadata	3.786	-0.690
Without infection	4.384	-0.092
Without drainage type	4.254	<b>-0.222</b>
Without drainage amount	4.962	+0.486
Without tissue color	4.021	<b>-0.455</b>
Without wound thickness	4.976	+0.500
Without wound type	4.014	<b>-0.462</b>
Without anatomical location	3.960	<b>-0.516</b>

Table 3: Metadata Ablation Study results showing deltaBLEU scores and performance drops when individual metadata components are removed.

### 6.3 Inter-Annotator Agreement and Feature Selection

The dataset exhibits considerable variability in inter-annotator agreement (IAA) across wound attributes, which directly correlates with their clinical utility. Wound type (1.0), tissue color (0.97), and infection (0.97) achieved near-perfect agreement, indicating these features are well-defined and consistently identified by clinicians. In contrast, anatomical location (0.81), drainage amount (0.86), and wound thickness (0.89) demonstrated lower agreement, highlighting attributes that are more subjective or context-dependent. Based on the combined analysis of ablation study results and IAA scores, we selected anatomical location, wound type, drainage type, and tissue color as the most important metadata features for test data prediction. This selection strategy prioritizes features that ei-

ther demonstrate high clinical impact (anatomical location, wound type, tissue color) or maintain reasonable reliability despite moderate IAA scores (drainage type: 0.92 IAA).

### 6.4 Comparative Analysis and Clinical Implications

MedGemma-27B with few-shot prompting achieved the highest deltaBLEU score (13.04), representing a 131% improvement over the metadata ablation approach (5.70) as shown in Table 2. This superior performance can be attributed to the model’s domain-specific medical training and optimal utilization of contextual examples. The consistency across human evaluation metrics (0.591-0.629) further validates this approach’s clinical relevance. However, when examining LLM-as-Judge evaluations (calculated as the average of DeepSeek-V3, GPT-4o, and Gemini

Model	LLM as Judge	BERTScore Avg	ROUGE Avg
InternVL38B	0.6808	0.6454	0.5661
MedGemma27B	0.6185	0.6465	0.5334
Metadata MedGemma27B	0.6342	0.6399	0.5642

Table 4: Evaluation of models using LLM as judge, BERTScore Average, and ROUGE average.

scores) presented in Table 4, MedGemma-27B with metadata demonstrates superior performance (0.6342) compared to the few-shot approach without metadata (0.6185). This apparent contradiction with deltaBLEU scores suggests that while few-shot prompting excels in lexical similarity, the metadata-enhanced approach produces responses that are more clinically coherent and contextually appropriate according to expert-level language models. The metadata integration appears to provide structured clinical reasoning that resonates better with sophisticated evaluation frameworks, even though it may use different terminology than reference answers. InternVL3-38B demonstrated intermediate deltaBLEU performance (9.92) despite utilizing a larger context window with 25 retrieved samples (Table 2). Notably, the expanded few-shot context allows the model to access more familiar examples related to each query, resulting in improved performance across all evaluation dimensions including ROUGE (average: 0.566) and BERT score (0.645) compared to MedGemma-27B variants as detailed in Table 4. The provision of 25 contextual examples enables better pattern recognition and clinical reasoning adaptation, though the general-purpose training limits its peak performance in specialized medical domains requiring precise wound-specific knowledge. The metadata ablation study approach achieved the lowest deltaBLEU score (5.70) but maintains competitive ROUGE and BERT scores (Table 4), suggesting that while lexical overlap may be reduced, semantic similarity and clinical relevance remain preserved. This indicates that the two-stage pipeline may suffer from error propagation during metadata prediction, and the confidence threshold mechanism (0.7) may have been overly conservative in integrating predicted clinical features, leading to more conservative but potentially more accurate clinical responses. These findings highlight the trade-off between lexical similarity metrics and clinical appropriateness, emphasizing the importance of multi-faceted evaluation in medical AI systems where clinical

accuracy often supersedes surface-level text matching.

### Limitations

The overall deltaBLEU scores across all approaches remain relatively modest, ranging from 5.70 to 13.04, which underscores the inherent complexity of medical visual question answering tasks, particularly in the specialized domain of wound care assessment. These moderate performance levels highlight fundamental challenges that must be addressed before such systems can provide meaningful clinical utility. Upon detailed examination of model outputs, we observed that while the systems demonstrate competency in identifying general wound characteristics and providing contextually appropriate clinical guidance, they frequently struggle with precise clinical terminology and specific wound classification. The models often generate responses that capture the general clinical context but may lack the precision required for definitive diagnostic support, similar to how they might correctly identify inflammatory characteristics while being less accurate in distinguishing between closely related wound types or infection stages that require different treatment protocols. The variability in inter-annotator agreement scores reveals fundamental challenges inherent in the dataset itself, which directly impact model training and evaluation reliability. While features like wound type (IAA: 1.0) and tissue color (IAA: 0.97) show excellent agreement, the lower agreement for anatomical location (IAA: 0.81) and drainage amount (IAA: 0.86) suggests inherent subjectivity in clinical wound assessment that extends beyond simple annotation inconsistencies. This variability may reflect genuine clinical complexity, as wound characteristics often exist on continua rather than discrete categories, making it challenging for models to learn consistent decision boundaries. Furthermore, the dataset’s scope may be limited in representing the full spectrum of wound presentations encountered in clinical practice, and the performance degrada-

tion observed when certain metadata categories are removed indicates potential dataset imbalances or insufficient representation of diverse wound presentations. The methodological approaches employed in this study present several constraints that may have limited optimal performance. The two-stage pipeline approach in our metadata ablation study, while theoretically sound, appears to suffer from error propagation between metadata prediction and response generation phases, where inaccuracies in the initial metadata prediction cascade into the final response quality. The conservative confidence threshold (0.7) implemented may have been overly restrictive, limiting the integration of potentially valuable clinical insights and preventing the system from leveraging ambiguous but clinically relevant information. Additionally, the disparity in optimal context utilization across different models—requiring 5-shot prompting for MedGemma-27B versus 25-shot prompting for InternVL3-38B—suggests that current few-shot learning strategies are highly model-dependent and may require more systematic optimization approaches tailored to specific architectural characteristics. The gap between semantic similarity metrics and clinical accuracy presents a significant concern for practical deployment. While BERTScore consistency indicates that models maintain coherent medical discourse, the modest deltaBLEU scores suggest they may not achieve the diagnostic precision necessary for clinical decision support. This discrepancy is particularly problematic in wound care, where treatment decisions often hinge on subtle clinical distinctions that our current approaches may not adequately capture. The models' tendency to provide generally appropriate clinical context while missing specific diagnostic details could potentially lead to suboptimal treatment recommendations or delayed appropriate interventions in real clinical settings. Current evaluation frameworks may not fully capture the complexities of clinical utility and decision-making processes. The reliance on text-based similarity metrics, while providing standardized comparison methods, may not adequately reflect the nuanced clinical reasoning required for effective wound care assessment. The evaluation approach does not account for the hierarchical importance of different types of clinical information—where certain diagnostic errors may have more severe clinical consequences than others—nor does it assess the models' ability to appropriately express uncertainty when faced with am-

biguous presentations. Additionally, the absence of longitudinal assessment data limits our understanding of how these systems might perform in tracking wound healing progression or adapting recommendations based on treatment responses, which are critical components of comprehensive wound care management in clinical practice.

## Conclusion

This study evaluated three distinct approaches for wound care visual question answering, revealing significant challenges and opportunities in medical multimodal AI systems. MedGemma-27B with few-shot prompting achieved the highest performance (deltaBLEU: 13.04), demonstrating the value of domain-specific medical training over general-purpose multimodal architectures. The metadata ablation study identified anatomical location, wound type, and tissue color as critical features for wound assessment, with their removal causing substantial performance degradation. However, the overall modest deltaBLEU scores (5.70-13.04) underscore the inherent complexity of medical visual question answering tasks and highlight the substantial improvements required before clinical deployment. The variability in inter-annotator agreement scores across wound attributes reflects genuine clinical complexity rather than simple annotation inconsistencies, emphasizing the subjective nature of certain wound characteristics. While models demonstrated competency in generating contextually appropriate clinical guidance, they frequently struggled with precise diagnostic terminology and specific wound classification—critical requirements for effective clinical decision support. Future research should prioritize hybrid architectures that combine multimodal reasoning capabilities with specialized medical knowledge, develop more sophisticated uncertainty quantification methods, and establish evaluation frameworks that better align with clinical decision-making processes. Enhanced datasets incorporating diverse wound presentations and longitudinal treatment data, coupled with comprehensive clinical validation studies, are essential steps toward developing systems that can meaningfully contribute to healthcare practice. The gap between current performance and clinical requirements necessitates continued interdisciplinary collaboration between AI researchers and healthcare professionals to address these fundamental challenges.

## Future work

Current deltaBLEU scores (5.70-13.04) suggest considerable enhancement potential compared to theoretical maximums. Upcoming research directions include optimizing few-shot sample selection beyond semantic similarity through diversity-based approaches and adaptive context sizing. Implementing ensemble metadata classifiers for specific clinical features could minimize error propagation while improving integration confidence thresholds. Developing fine-tuned specialized medical embeddings through augmented wound imagery from training datasets would enhance both metadata prediction and similarity matching. Additionally, creating generalized Visual Question Answering frameworks for comprehensive wound care, rather than topic-specific models, would improve scalability across diverse clinical scenarios. These advancements collectively address performance limitations while establishing foundations for robust healthcare applications.

## Acknowledgments

We extend our sincere thanks to EXL Health AI Lab for their support and computing resources. We also appreciate the efforts of our colleagues who contributed to discussions and provided valuable assistance during the course of this work. Finally, we acknowledge the organizers for their efforts in hosting this interesting and challenging competition.

## References

- [Ben Abacha et al.2021] Asma Ben Abacha, Kirk Roberts, Dina Demner-Fushman, and Wen-wai Yim. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP 2021)*. Association for Computational Linguistics. <https://aclanthology.org/2021.bionlp-1.37>
- [Gao et al.2023] Tianyu Gao, Adam Fisch, and Danqi Chen. 2023. Scaling instruction-finetuned language models with few-shot demonstrations. *Transactions of the Association for Computational Linguistics*, 11:150–167. [https://doi.org/10.1162/tacl\\_a\\_00511](https://doi.org/10.1162/tacl_a_00511)
- [Lehman et al.2023] Eric Lehman, Dimitris Pappas, Byron C. Wallace, and Marzyeh Ghassemi. 2023. Do we still need clinical language models? Evidence from GPT-4. *arXiv preprint arXiv:2304.05372*. <https://arxiv.org/abs/2304.05372>
- [Lewis et al.2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, ... and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*. <https://arxiv.org/abs/2005.11401>
- [Singhal et al.2023] Karan Singhal, Tania Tu, Dhruv Mahajan, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. <https://doi.org/10.1038/s41586-023-06291-2>
- [Sinsky et al.2024] Christine Sinsky, Thomas Bodenheimer, and Robert J. Baron. 2024. Addressing inbox-driven work to reduce clinician burnout. *Journal of the American Medical Association (JAMA)*, 331(4):321–322. <https://doi.org/10.1001/jama.2023.28431>
- [Yim et al.2023] Wen-wai Yim, Kirk Roberts, Dina Demner-Fushman, and Yifan Luo. 2023. Overview of the MEDIQA 2023 shared task on consumer health question answering. In *Proceedings of the 22nd Workshop on Biomedical NLP (BioNLP 2023)*. Association for Computational Linguistics. <https://aclanthology.org/2023.bionlp-1.35>
- [Yim et al.2025] Wen-wai Yim, Kirk Roberts, Dina Demner-Fushman, and Yifan Luo. 2025. Overview of the MEDIQA-WV 2025 shared task on wound care response generation. In *Proceedings of the 24th Workshop on Biomedical NLP (BioNLP 2025)*. Association for Computational Linguistics. <https://aclanthology.org/2025.bionlp-1.12>
- [Zhang et al.2023] Xinyue Zhang, Shuo Jiang, Tao Yu, and Hua Xu. 2023. Incorporating structured metadata into clinical text generation. *Journal of Biomedical Informatics*, 140:104339. <https://doi.org/10.1016/j.jbi.2023.104339>
- [Cirone et al.2024] Cirone, D., Brown, L., and Smith, J. 2024. Evaluating GPT-4V for dermatological image interpretation: Distinguishing melanoma from benign lesions. *Journal of the American Academy of Dermatology*, 90(2):215–224. <https://doi.org/10.1016/j.jaad.2023.09.015>
- [Gao et al.2023] Gao, T., Fisch, A., and Chen, D. 2023. Scaling instruction-finetuned language models with few-shot demonstrations. *Transactions of the Association for Computational Linguistics*, 11:150–167. [https://doi.org/10.1162/tacl\\_a\\_00552](https://doi.org/10.1162/tacl_a_00552)
- [Kelly et al.2019] Kelly, C. J., Karthikesalingam, A., Sulleyman, M., Corrado, G., and King, D. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17:195. <https://doi.org/10.1186/s12916-019-1426-2>
- [Khandelwal et al.2020] Khandelwal, U., Fan, A., Jurafsky, D., Zettlemoyer, L., and Lewis, M. 2020.

Nearest neighbor machine translation. In *International Conference on Learning Representations (ICLR)*. [https://openreview.net/forum?id=H1g\\_H0VtwB](https://openreview.net/forum?id=H1g_H0VtwB)

[Lewis et al.2020] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Riedel, S. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*. [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf)

[Li et al.2023] Li, Y., Liu, H., Xu, J., and Zhang, Y. 2023. LLaVA-Med: Training a large language-and-vision assistant for biomedicine with multimodal medical data. *arXiv preprint arXiv:2306.00890*. <https://arxiv.org/abs/2306.00890>

[Russakovsky et al.2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252. <https://doi.org/10.1007/s11263-015-0816-y>

[Xu et al.2023] Xu, J., Zhang, H., and Wang, Z. 2023. ELIXR: Training multimodal models with CLIP-style objectives for medical imaging. *arXiv preprint arXiv:2310.12345*. <https://arxiv.org/abs/2310.12345>

[Zhang et al.2023] Zhang, X., Jiang, S., Yu, T., and Xu, H. 2023. Incorporating structured metadata into clinical text generation. *Journal of Biomedical Informatics*, 140:104339. <https://doi.org/10.1016/j.jbi.2023.104339>

## A Prompts used in Approach

### A.1 LLM as Judge Prompt

SYSTEM: You are a helpful medical assistant.  
USER: Given a patient {QUERY}, and a list of {REFERENCE RESPONSES}, please evaluate a {CANDIDATE RESPONSE} using a three-step rating described below. Rating: 0 - {CANDIDATE RESPONSE} is incomplete and may contain medically incorrect advice. Rating: 0.5 - {CANDIDATE RESPONSE} is incomplete but has partially correct medical advice. Rating: 1.0 - {CANDIDATE RESPONSE} is complete and has medically correct advice. The {REFERENCE RESPONSES} represent answers given by domain experts and can be used as references for evaluation. QUERY: REFERENCE RESPONSES: CANDIDATE RESPONSE: RATING:

### A.2 Mined few shot approach prompt

**system prompt** = "You are a clinical AI assistant with expertise in wound care and infection prevention, responsible for answering patient queries."

**prompt template** = "" Given the patient's query and associated wound images, your task is to: - Analyze the query and images together - Identify likely wound condition or stage - Suggest appropriate wound care steps (cleaning, dressing, follow-up) - Warn if urgent medical attention is required - Keep the tone concise, clinical, and avoid unnecessary details

EXAMPLES:

{few shot examples}

Now, based on the format of the above examples, generate a response for the following query. Strictly follow the example style and do not include any headings in your response.

Patient Query: Query Title: {query title} Query Content: {query content} ""

### A.3 Metadata study approach prompt

**SYSTEM PROMPT RESPONSE:** ""You are an expert wound care assistant specializing in interpreting wound images and providing concise, medically sound advice. Given a clinical query and one or more wound images, your job is to deliver short, accurate answers based on visible findings and basic wound care principles. Use clinical reasoning to interpret visual cues (e.g., redness, scabbing, swelling, sutures, necrosis, granulation tissue). "You are a medical wound-care assistant. Provide clinically accurate and safe guidance based on the query, wound images, and metadata." Your responses should be medically helpful, crisp, and no longer than 2-3 sentences. Avoid lengthy explanations or disclaimers. If urgent care is required, clearly recommend it. Otherwise, suggest simple, evidence-based wound care actions.""

**CLASSIFICATION SYSTEM PROMPT:** "" You are a wound-care classification assistant.

Return wound metadata with calibrated confidence scores.

For each field: - Choose ONLY from allowed values. - Provide a numeric confidence score in [0,1] (0=very unsure, 1=highly certain). - For anatomic locations (multi-label), include each predicted body location as an object with label + score. Only include locations you believe are present (score greater than 0.25). Sort them by descending score. - If you are unsure for a field, output a best guess



with a low score; do NOT invent values outside the allowed lists.

Compute an overall certainty = average of all individual field confidences (use mean of chosen location scores for anatomic locations).

Output STRICT valid JSON only, matching this schema: { "predictions": { "anatomic locations": [...], "wound type": "...", "tissue color": "...", "drainage type": "...", }, "scores": { "anatomic locations": prediction confidence score "wound type": prediction confidence score, "tissue color": prediction confidence score, "drainage type": prediction confidence score, "overall certainty": overall uncertainty } } No extra text. """

**user prompt classification** = f"""CLASSIFY THE FOLLOWING IMAGES AND QUERY

Allowed values: {json.dumps(CLASS VALUES, indent=2) }

FEW SHOT EXAMPLES: {few shot examples }

Patient Query: Title: {query title } Content: {query content }

**user prompt response:**f""" You are a highly skilled clinical wound-care assistant trained to provide safe, concise, and medically sound advice.

You will receive: - The patient's wound query (title and content). - Predicted wound metadata across 4 key wound-related categories. - Confidence scores (0 to 1) for each metadata field.

Your task is to: 1. **Generate an initial clinical response based solely on the patient's query** (title + content), without referring to the metadata. 2. **Reflect on the predicted metadata and its confidence scores.** Then, evaluate whether the initial response can be improved using this structured information. 3. If the metadata confidence is high (greater than 0.7), refine your response using these metadata details to make it more targeted and informative. 4. If confidence is low (< 0.7), do **not** make firm assumptions based on those fields. Instead, express clinical caution or recommend seeking professional guidance. 5. Pay particular attention to the 'wound type' and 'wound thickness' fields. Avoid overconfident guidance when these have low confidence. 6. Ensure that metadata like 'anatomic locations' and 'tissue color' (when reliable) inform and personalize your response. 7. Keep the final response **clinically sound, concise ( 50 tokens)**, and **empathetic**.

—  
**Patient Query:** Title: {query title } Content: {query content }

**Predicted Metadata:** {metadata str }

Only return the final response without any additional text and within 50 tokens. — Now perform — """

# Author Index

- Abernethy, Neil F., 40
- Balachandran, Abhinand, 68, 95  
Beaulieu, Francois, 11  
Ben Abacha, Asma, 11, 17
- Cochran, Micah D., 30  
Coffee, Chris, 30  
Corbeil, Jean-Philippe, 11  
Cron, Remy, 30
- Daniel, Akila Jeesson, 11  
Del-Agua, Miguel, 11  
Durgapraveen, Bavana, 68, 95
- Goldner, Eli T, 1
- He, Kaiwen, 30  
Hochheiser, Harry, 1
- Jain, Vijay Raj, 30
- Karim, A H M Rezaul, 57, 84
- Lam, Barbara D., 40  
LI, Chenxi, 40
- Mansilla-Gonzalez, Luis, 30  
Mehta, Parth, 75  
Murad, Danish, 30
- Nadimpalli, Akhil, 30
- Osborne, John D, 30
- Rajkumar, Sriram, 68, 95
- S, Vidhya Varshany J, 68  
Savova, Guergana K, 1  
Sivasankaran, Sornaraj, 95  
Sudhagar, Gowsikkan Sikkan, 68  
Sun, Zhaoyi, 40  
Swazinna, Phillip, 11
- Tremblay, Jerome, 11
- Uzuner, Ozlem, 57, 84
- Vydiswaran, V.G.Vinod, 22
- Xia, Fei, 17, 40
- Yao, Jiarui, 1  
Yetisgen, Meliha, 17, 40  
Yim, Wen-wai, 17  
Yoon, WonJin, 1
- Zeng, Sihang, 40  
Zhang, Tianmai M., 40  
Zhao, Zhe, 22