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 Subtask 1 of the ChemoTimelines 2025 shared task on predicting temporal relations between preidentified 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 pretraining 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), pretrained 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 pretraining PubMedBERT on a clinical corpus. These models leveraged contexualized embeddings and domain adaptation techniques like fine-tuning and continual pretraining, and have shown improvement over conventional machine learning models.

Proceedings of the 7th Clinical Natural Language Processing Workshop, pages 22–29
October 30, 2025 ©2025 Association for Computational Linguistics

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.

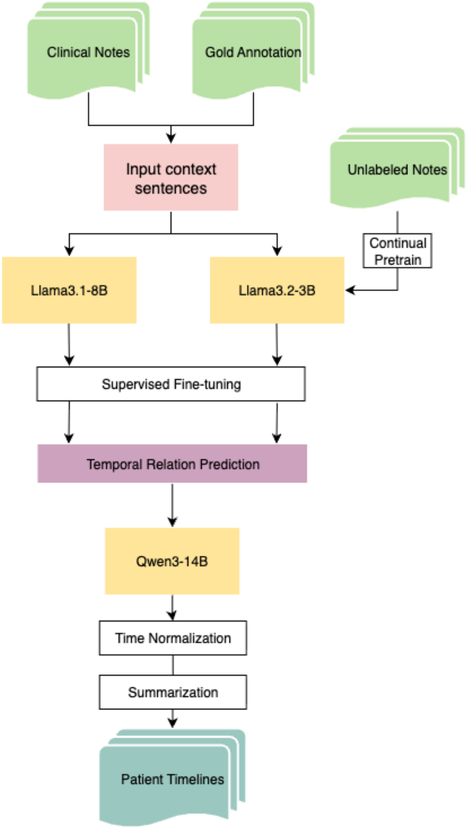

Figure 1: Overview of the submitted systems.

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.

3.2 Text Pre-processsing

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
Breast Cancer		
BEGINS-ON	131	27
CONTAINS	298	57
ENDS-ON	26	29
NO-REL (pre-threshold)	2320	710
NO-REL (post-threshold)	389	133
Melanoma		
BEGINS-ON	10	42
CONTAINS	37	157
ENDS-ON	1	2
NO-REL (pre-threshold)	293	1138
NO-REL (post-threshold)	35	192
Ovarian Cancer		
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, models 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 normalize 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 pretraining 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 <event>Adriamycin</event> and Cytoxan, with the last dose being on <time>7/27/13</time>."* This error did not appear in cases with other prepositions such as "through" or "in"; for example, *"Carboplatin, cytoxan, and <event>Avastin</event> x8 cycles through <time>May 2013</time>."* 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
Development set				
1. Llama3.1-8B + Qwen3-14B	0.83	**0.84**	**0.83**	0.82
2. Llama3.1-8B + Regex + Qwen3-14B	**0.84**	0.82	0.82	**0.87**
3. Llama3.2-3B + Regex + Qwen3-14B	0.80	0.83	**0.83**	0.73
Test set				
1. Llama3.1-8B + Qwen3-14B	**0.60**	0.56	**0.66**	0.57
2. Llama3.1-8B + Regex + Qwen3-14B	**0.60**	0.56	**0.66**	0.57
3. Llama3.2-3B + Regex + Qwen3-14B	0.59	**0.57**	0.60	**0.60**

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 pretraining 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 tasked 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 Wagholikar, 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 ADDTIONAL TEXT.”