UTSA-NLP at ChemoTimelines 2024: Evaluating Instruction-Tuned Language Models for Temporal Relation Extraction

Xingmeng Zhao and **Anthony Rios**
Department of Information Systems and Cyber Security
The University of Texas at San Antonio
{Xingmeng.Zhao, Anthony.Rios}@utsa.edu

Abstract

This paper presents our approach for the 2024 ChemoTimelines shared task. Specifically, we explored using Large Language Models (LLMs) for temporal relation extraction. We evaluate multiple model variations based on how the training data is used. For instance, we transform the task into a question-answering problem and use QA pairs to extract chemo-related events and their temporal relations. Next, we add all the documents to each question-answer pair as examples in our training dataset. Finally, we explore adding unlabeled data for continued pretraining. Each addition is done iteratively. Our results show that adding the document helps, but unlabeled data does not yield performance improvements, possibly because we used only 1% of the available data. Moreover, we find that instruction-tuned models still substantially underperform more traditional systems (e.g., EntityBERT).

1 Introduction

Extracting chemotherapy treatment timelines from clinical notes is crucial in Clinical Natural Language Processing (ClinicalNLP) for enhancing patient care and advancing cancer research (Cui et al., 2023). Researchers can construct detailed treatment timelines within Electronic Health Records (EHR) across various medical domains by identifying and extracting events related to chemotherapy treatments and their temporal information from medical documents. This work aims to develop an end-to-end system utilizing Large Language Models (LLMs) in a Question-Answer format for chemotherapy timeline extraction. Such a system will aid healthcare professionals in comprehending patient histories, thereby improving clinical text-mining efforts and assisting physicians in making more informed care decisions. Additionally, it will contribute to research in personalized cancer treatment development.

The main approach in clinical entity and relation extraction tasks heavily relies on pre-trained, domain-specific models like BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Gu et al., 2021), and EntityBERT (Lin et al., 2021). These models are trained on a broad range of biomedical corpora, like PubMed articles and clinical notes, to understand the complex language of the clinical domain, which is often succinct and laden with phrases, jargon, non-standard expressions, abbreviations, assumptions, and specialized knowledge. These models are then adapted or fine-tuned for specific tasks such as named entity recognition (NER), relation extraction (RE), and event extraction (EE), which often employ strategies like multi-task learning (MTL) and an all-in-one scheme to enhance performance across multiple tasks by leveraging shared knowledge and representations (Luo et al., 2023; Yadav et al., 2020). However, there are still challenges, such as a drop in performance when these models are used for out-of-domain tasks or very different sub-domains in terms of context and terminology, revealing their limitations in adaptability (Košprdić et al., 2023).

Recently, Large Language Models (LLMs) have shown remarkable potential in Natural Language Processing (NLP) tasks, including text generation, reasoning, text classification, summarization, and question answering, through their ability for zero-shot or few-shot learning (Xu et al., 2023). This capability allows them to adapt to new tasks quickly with minimal fine-tuning. This adaptability has resulted in their outstanding application performance, including NER and RE within the general domain. Models like CoT-ER (Ma et al., 2023), GPT-RE (Wan et al., 2023), and PromptNER (Ashok and Lipton, 2023) show that through few-shot learning or zero-shot learning, these generative LLMs can achieve performance levels competitive with the state-of-the-art methods in entity or relation extraction (Li et al., 2023; Brown et al., 2020; Wei et al.,

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 604–615
June 21, 2024 ©2024 Association for Computational Linguistics

2022; Liu et al., 2023). This achievement is primarily due to their capability to use task-specific and concept-level knowledge stored during pre-training, which is then effectively leveraged through prompting to generate relevant evidence for the tasks.

However, challenges arise when adapting LLMs from the general to the medical domain, primarily due to their lack of domain-specific knowledge and the difficulty in incorporating new, relevant factual information over time (Jiang et al., 2024; Brokman and Kavuluru, 2024; Li and Zhang, 2023). While LLMs have shown potential in biomedical natural language processing tasks through innovative in-context learning strategies, their application in specific tasks like NER and RE remains problematic. This is partly because current few-shot learning methods, which trained on large amounts of source data and fine-tuning on exemplars from the target domain, do not perform well in the medical context (Gutiérrez et al., 2022; Keloth et al., 2024; Ma et al., 2023). The discrepancy arises from significant differences in entity and relationship definitions between general and medical texts (Das et al., 2022). To address these challenges, researchers have explored various approaches, including the development of domain-specific generative LLMs like BioGPT (Luo et al., 2022), BioMedLM (Bolton et al., 2024), and BioBART (Yuan et al., 2022). These models are trained from scratch using medical corpora such as PubMed or are continually fine-tuned on medical data. Basically, fine-tuning is required for adequate performance on biomedical NLP tasks. These efforts represent steps toward bridging the gap in domain adaptation for LLMs. However, updating these models for the rapidly changing medical field is still non-trivial due to the risk of catastrophic forgetting during fine-tuning (Ren et al., 2024), highlighting the need for better training methods tailored to medical knowledge.

To address this, we explored instruction-tuning methods for large language models, focusing on an open-source language model. Traditional fine-tuning methods risk forgetting previous knowledge, so we adopted a novel training strategy, gradually extending training to include associated documents and unlabeled datasets. Initially, we instruction-tuned on Question-Answer (QA) pairs before integrating complete EHR documents. Then, we trained on QA pairs and documents simultaneously. Finally, we continue pre-training on the large unlabeled corpus. Jiang et al. (2024) demonstrates that this integration strategy ensures the retention of acquired knowledge. In the inference stage, our system directly generates output relations from input questions for subtask 1. For subtask 2, we first extract event entities and time expressions before predicting relationships between identified entities using different input questions. Our approach provides an end-to-end relation extraction system for extracting Chemotherapy Treatment Timelines. This system formulates the task as a text generation task, using clinical notes as input to generate relational triplets end-to-end, without requiring additional intermediate annotations, as seen in the REBEL method (Cabot and Navigli, 2021).

In summary, this paper makes the following contributions:

1. We introduce a novel approach that combines instruction-based fine-tuning with continuous knowledge acquisition to adapt pre-trained general LLMs to the medical domain, specifically targeting the extraction of chemotherapy treatment timelines.

2. We evaluate the performance of a smaller 7B model, OpenChat-3.5-7B (Wang et al., 2023b), on extracting chemotherapy treatment timelines for breast cancer, ovarian cancer, and melanoma datasets provided by the University of Pittsburgh/UMPC. Additionally, we conduct a detailed analysis of each training component to establish a robust framework for biomedical end-to-end relation extraction, with the potential to apply the same approach to other biomedical NLP tasks.

3. We conduct an error analysis to identify the strengths and weaknesses of our proposed approach, offering insights into areas for potential improvement.

2 Methods

In this section, we describe our instruction-tuned LLMs strategy. Figure 1 shows a high-level overview of our approach. We convert the information extraction task to the question-answer instruction format. Our strategy has three main components: 1) Instruction-tuning LLMs on task-specific QA pairs (i.e., Named Entity Recognition (NER) and Relation Classification (RE)); 2) Joint

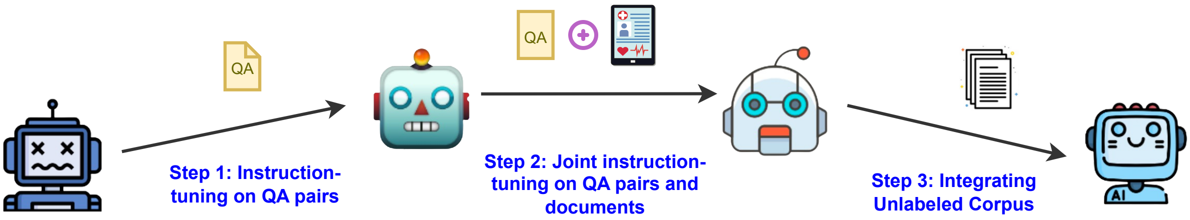

QA
Step 1: Instruction-tuning on QA pairs
QA
Step 2: Joint instruction-tuning on QA pairs and documents
Step 3: Integrating Unlabeled Corpus
AI

Figure 1: Overview of instruction-tuned LLMs Framework. First, we perform instruction-tuning on LLMs using task-specific QA pairs (e.g., NER and RE). Second, we conduct further instruction-tuning on QA pairs and associated documents to enhance its ability to progressively absorb knowledge from simpler to more complex data. Finally, we continue pre-training the model on an unlabeled corpus to refine its capabilities in the clinical domain further.

instruction-tuning on QA pairs and associated documents to enhance its ability to absorb knowledge progressively from simpler to more complex data; and 3) Continuing pre-training on unlabeled corpus, our intuition is first trained on QA pairs to understand knowledge access patterns, then progresses to training on a combination of QA and document data to align knowledge access through questions and knowledge encoding from documents, this will help absorb information from unlabeled data. We describe each component in the following subsections and how the three components are integrated into a unified training strategy.

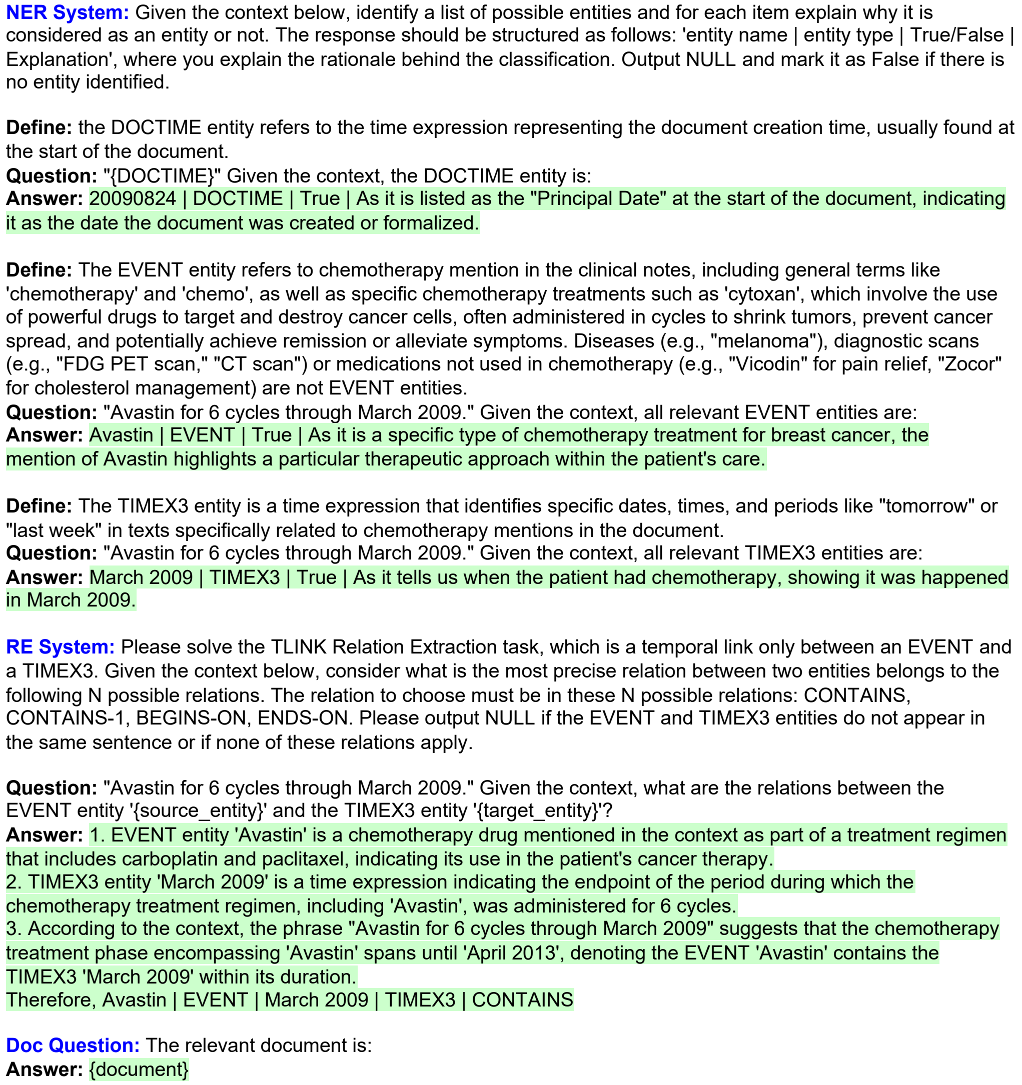
2.1 Step 1: Instruction-tuning on QA pairs

We fine-tuned the pretrained open-source LLMs (i.e., OpenChat-3.5-7B) for two clinical tasks: classifying TLINK temporal relations and recognizing named entities, including DocTimeRel, EVENTS, and TIMEX3. Subsequent sections detail the labeled datasets used for these instruction-tuning tasks, and Figure 2 illustrates the format used for task-specific question-answer pairs.

Relationship Classification QA Design. For the RE QA pairs, Let C represent the input context, and let $e_{\text{event}} \in C$ and $e_{\text{timex3}} \in C$ denote a chemotherapy event entity and a time expression entity, respectively. For a set of predefined relation classes R, the goal of relation extraction is to determine the relationship $y \in R$ between the entity pair $(e_{\text{event}}, e_{\text{timex3}})$ within C. If no predefined relation exists between them, the model predicts $y =$ NULL. Building on the prior work (Ma et al., 2023), we use a three-step reasoning framework combining concept-level entity knowledge and explicit evidence to design question-answer instructions. This approach aims to maximize the use of knowledge embedded in LLMs to support step-by-step reasoning. For the RE task, questions are formulated with instructions, definitions of potential relations, and the context. Answers are designed as a structured three-step reasoning process. First, we integrate concept-level knowledge about the event entity. Second, we apply a similar approach to the time expression entity. Third, to identify the most suitable relation label for the pair of entities within the context, we explicitly highlight relevant text spans as evidence and subsequently construct a coherent expression that combines the two entities and the relation label. An example using the relation label “CONTAINS” is shown in the last “Answer” in Figure 2. First, Avastin is described as a chemotherapy drug. Next, the TIMEX3 entity (March 2009) is described. Finally, some reasoning is described about a potential relation, and the relation is specified.

How does the model learn this reasoning framework? Inspired by Wan et al. (2023), we implement the OpenChat-3.5-7B model to generate logical reasoning in question-answer pairs. We employ few-shot learning to prompt the LLMs to generate a three-step reasoning process based on the question and corresponding given golden label. For example, we append the query “What are the three-step reasoning processes that lead to the relation between [entity1] and [entity2] being [relation] in the sentence [context]?” to the end of the question and corresponding a gold label. This prompt is then passed to the LLMs to generate the three-step reasoning. Specifically, we generate the reasons using an untrained OpenChat-3.5-7B without fine-tuning for all examples in our training dataset. These reasons are then used during our instruction tuning phase.

Named Entity Recognition QA Design. The NER QA instruction design is inspired by PromptNER (Ashok and Lipton, 2023), which shows the advantages of enhancing language models’ under-



NER System: Given the context below, identify a list of possible entities and for each item explain why it is considered as an entity or not. The response should be structured as follows: 'entity name | entity type | True/False | Explanation', where you explain the rationale behind the classification. Output NULL and mark it as False if there is no entity identified.

Define: the DOCTIME entity refers to the time expression representing the document creation time, usually found at the start of the document.
Question: "{DOCTIME}" Given the context, the DOCTIME entity is:
Answer: 20090824 | DOCTIME | True | As it is listed as the "Principal Date" at the start of the document, indicating it as the date the document was created or formalized.

Define: The EVENT entity refers to chemotherapy mention in the clinical notes, including general terms like 'chemotherapy' and 'chemo', as well as specific chemotherapy treatments such as 'cytoxan', which involve the use of powerful drugs to target and destroy cancer cells, often administered in cycles to shrink tumors, prevent cancer spread, and potentially achieve remission or alleviate symptoms. Diseases (e.g., "melanoma"), diagnostic scans (e.g., "FDG PET scan," "CT scan") or medications not used in chemotherapy (e.g., "Vicodin" for pain relief, "Zocor" for cholesterol management) are not EVENT entities.
Question: "Avastin for 6 cycles through March 2009." Given the context, all relevant EVENT entities are:
Answer: Avastin | EVENT | True | As it is a specific type of chemotherapy treatment for breast cancer, the mention of Avastin highlights a particular therapeutic approach within the patient's care.

Define: The TIMEX3 entity is a time expression that identifies specific dates, times, and periods like "tomorrow" or "last week" in texts specifically related to chemotherapy mentions in the document.
Question: "Avastin for 6 cycles through March 2009." Given the context, all relevant TIMEX3 entities are:
Answer: March 2009 | TIMEX3 | True | As it tells us when the patient had chemotherapy, showing it was happened in March 2009.

RE System: Please solve the TLINK Relation Extraction task, which is a temporal link only between an EVENT and a TIMEX3. Given the context below, consider what is the most precise relation between two entities belongs to the following N possible relations. The relation to choose must be in these N possible relations: CONTAINS, CONTAINS-1, BEGINS-ON, ENDS-ON. Please output NULL if the EVENT and TIMEX3 entities do not appear in the same sentence or if none of these relations apply.

Question: "Avastin for 6 cycles through March 2009." Given the context, what are the relations between the EVENT entity '{source_entity}' and the TIMEX3 entity '{target_entity}'?
Answer: 1. EVENT entity 'Avastin' is a chemotherapy drug mentioned in the context as part of a treatment regimen that includes carboplatin and paclitaxel, indicating its use in the patient's cancer therapy.
2. TIMEX3 entity 'March 2009' is a time expression indicating the endpoint of the period during which the chemotherapy treatment regimen, including 'Avastin', was administered for 6 cycles.
3. According to the context, the phrase "Avastin for 6 cycles through March 2009" suggests that the chemotherapy treatment phase encompassing 'Avastin' spans until 'April 2013', denoting the EVENT 'Avastin' contains the TIMEX3 'March 2009' within its duration.
Therefore, Avastin | EVENT | March 2009 | TIMEX3 | CONTAINS

Doc Question: The relevant document is:
Answer: {document}

Figure 2: An instruction example for clinical document and task-specific QA pairs. Both subtask 1 and subtask 2 use the same training dataset and process. However, subtask 1 focuses on identifying temporal relations by generating specific relation pairs through tailored questions during inference. In contrast, subtask 2 first identifies chemo-related entities with distinct instructions before determining their temporal relationships. Tokens used for computing losses are highlighted in green.

standing of textual logical structures. This understanding is used to improve NER tasks by employing Chain-of-Thought Prompting, guiding the model through a step-by-step reasoning process that leads to entity identification. This technique boosts entity recognition accuracy and offers a versatile framework adaptable to various entity types by adjusting definitions and explanations within the prompting template (Ashok and Lipton, 2023; Wang et al., 2023a). Therefore, in our NER QA instruction design, each question includes instructions and definitions of entities, with answers detailing the chosen entities in the format of "entity name | entity type | True/False | Explanation," where the Explanation includes the rationale behind the NER type classification. Inspired by Ashok and Lipton (2023), this method employs Chain-of-Thought Prompting to refine our model's understanding of textual logic, enhancing NER tasks by guiding step-by-step reasoning. We've crafted a structured output template for the LLMs that identifies and classifies entities. This structure has the potential to enhance accuracy through outcome supervision using reinforcement learning (Gao et al., 2024). Additionally, the True/False component marks noun phrases that are relevant entities we want to extract (True) or irrelevant (False). In our experiments, we learn to generate relevant entities because we are fine-tuning, hence we only use True. However, we kept the option for False in future work by adding incorrect entities.

This format displays the model's decision-making process, making it adaptable across different NER types by simply modifying definitions. Similarly, we use the non-finetuned OpenChat-3.5-7B model, employing few-shot learning with manually created demonstrations to generate explanations for all examples in the training data. In general, our NER QA instruction includes three distinct categories of entities: EVENTS, which refer specifically to mentions of chemotherapy treatments; DocTimeRel, which represents the temporal relationship between an event and the time the document was created; and Temporal Expressions (TIMEX3), which are precise references to times linked to chemotherapy treatments. These entities are illustrated in Figure 2, which shows "Avastin" and "March 2009" as example extractions.

2.2 Step 2: Joint instruction-tuning on QA pairs and documents

In this training phase, the instruction combines QA pairs with their relevant documents. Intuitively, QA pairs are typically simple, unlike documents, which are usually more complex and dense, containing numerous factual details not available in a single (or few) sentence. Therefore, Jiang et al. (2024) suggests that it is effective to deliberately expose LLMs to QA data before continued pre-training on documents so that the process of encoding knowledge from complex documents considers how this knowledge is accessed through questions. During this phase, LLMs improve at digesting detailed content from documents, building on the QA pairs they've already learned. The training starts with QA pairs to grasp basic knowledge access patterns and then adds documents to enhance question-based knowledge access and document understanding. The instruction is created based on each document; we position all the NER QA pairs, followed by the RE QA pairs. Finally, the document itself is formatted as a QA pair, with the question identifying the document and the answer being the document's content, as illustrated in Figure 2. Jiang et al. (2024) found that placing the documents after the QA pairs leads to better performance than placing them before. We also experimented with positioning the document before and after the QA pairs and tested on the melanoma development set. The results showed that placing the document after the QA pairs yielded better performance. Therefore, we put the document after the QA pairs in our following experiments.

2.3 Step 3: Integrating Unlabeled Corpus

In this training phase, we aim to improve how the fine-tuned OpenChat-3.5-7B model handles clinical documents, which are often complex and full of medical terminology. Instead of using instruction-tuning alone, we continued "pre-training" the model on unlabeled documents (i.e., training on unlabeled data after instruction-tuning).[1] This potentially helps the model learn a specialized vocabulary for the clinical domain, capturing important terms such as diseases, symptoms, medications, and medical procedures in their original context (Lin et al., 2021). This approach is crucial for enhancing the model's performance on tasks specific to the clinical field. Based on Jiang et al. (2024), there's a concern that directly continuing pre-training on a vast, unlabeled clinical corpus might lead to the model forgetting previously acquired knowledge. However, by initially training on QA pairs to grasp knowledge access patterns and then moving on to a blend of QA and document data, we can strengthen the model's ability to assimilate document knowledge. This method helps mitigate the issue of catastrophic forgetting by aligning how the model accesses knowledge through questions with how it encodes knowledge from documents (Ouyang et al., 2022; Jiang et al., 2024). Technically, we employed Byte-Pair Encoding (BPE) (Gage, 1994) to break down the text into small context windows, considering the OpenChat-3.5-7B model's 8192 token maximum context limit, setting our windows to 7800 tokens for efficiency. We prepared the training data by joining these pieces with an end-of-sequence (eos) token and then splitting the extended text into sections. This structured training method is designed to make the model more effective at analyzing and interpreting medical documents.

3 Experiments

In this section, we provide a brief overview of the dataset, discuss the evaluation metrics, discuss our results on the validation dataset, and briefly mention the final model performance in the competition on the test set.

[1] Because of lack of time and limited GPU resources, we were not able to use the entire unlabeled dataset and only learned on less than 1% of the unlabeled data.

3.1 Dataset

In this shared task, we use both unlabeled and labeled EHRs, including radiology reports, pathology notes, clinical notes, oncology notes, discharge summaries, and progress reports, from the University of Pittsburgh/UPMC to construct the end-to-end system for Extracting Chemotherapy Treatment Timelines. For the unlabeled data, this included EHR notes from approximately 62,000 patients with breast and ovarian cancer and 16,000 patients with melanoma. For the labeled data, we have gold annotations for 310 patients' histories, focusing on EVENTs, TIMEX3s entities, and temporal relations (TLINKs) between an EVENT and a TIMEX3. The training set includes EHRs for ovarian (26 patients), breast (33 patients), and melanoma (10 patients), while the development set comprises records from ovarian (8 patients), breast (16 patients), and melanoma (3 patients). Additionally, for ethical reasons and to protect patient privacy, the data has been de-identified (Jiarui Yao, 2024).

An EVENT refers to any relevant chemotherapy treatment on the clinical timeline, each with a temporal relation to the document creation time (DocTimeRel), categorized as BEFORE, BEFORE-OVERLAP, OVERLAP, or AFTER. Temporal expressions (TIMEX3) denote discrete references to time, normalizations to a unified format (e.g., "YYYY-MM-DD") using TimeNorm (Laparra et al., 2018; Xu et al., 2019). Additionally, temporal relations (TLINKs) link an EVENT and TIMEX3, including categories such as CONTAINS, CONTAINS-1, BEFORE, BEGINS-ON, and ENDS-ON, where CONTAINS-1 is the inverse of CONTAINS, meaning the Target CONTAINS the Source (Styler IV et al., 2014).

For training, we created positive NER QA pairs from all gold standard examples, even though there were no relations between EVENT and TIMEX3. For RE QA pairs, we randomly selected three pairs of chemo events and time expressions with no temporal relation, where the answer would be NULL.

3.2 Hyperparameters

In our experiments, we trained models on 2 Nvidia A6000 GPUs using DeepSpeed Zero stage 2 (Rasley et al., 2020), HuggingFace Accelerate (Gugger et al., 2022), and FlashAttention2 (Dao, 2023) for a maximum of 10 epochs and using Melanoma dev set to select best epoch for all three stage training. Following Jiang et al. (2024), we employed the AdamW optimizer (Loshchilov and Hutter, 2018) with specific parameters (β_1 = 0.9, β_2 = 0.95, weight decay = 0.1) and set a maximum context length of 1024.

For instruction tuning on question-answer pairs, we used a batch size of 128 and learning rates of 3e-5 for direct pairs and 5e-6 when documents were associated while continuing pre-training on unlabeled datasets at a batch size of 36 and a learning rate of 3e-5. We use spaCy's "en_ner_bc5cdr_md" model for sentence boundary detection and text segmentation. Moreover, we adopted Low-Rank Adaptation (LoRA) fine-tuning (Hu et al., 2021) with a rank of 256, LoRA alpha of 512, and LoRA dropout of 0.05, targeting modules ["q_proj", "o_proj", "k_proj", "v_proj", "gate_proj", "up_proj", "down_proj", "fc_in", "fc_out","wte"], to optimize specific target modules within pre-trained language models (LLMs), effectively reducing the number of parameters needed for training without altering the original model weights. This approach was facilitated by using the "trl" library from HuggingFace (von Werra et al., 2020), enhancing our model's performance and efficiency.

When training on QA pairs, we compute the average negative log-likelihood loss by focusing only on the tokens within the answer. This approach is inspired by Lin et al. (2024), which suggests that not all tokens are equally important in language model training. We can enhance the model's efficiency and performance by selectively focusing on tokens that align with the desired distribution. For QA + Doc training, we treat the phrase "The relevant document is" as a question and apply next-token prediction loss to the document's tokens, treating them as an expanded answer. This is because the document provides a rich context that informs the model's understanding, enabling it to learn from contextually relevant tokens, as shown in Figure 2.

In the inference stage, we experimented with different settings with temperatures from 0.1 to 0.9, top p values from 0.1 to 0.6, and top k options of 10, 20, and 30. After experimenting, we found that the best settings were a temperature of 0.2, a top p value of 0.5, and a top k of 20.

[1] https://github.com/clulab/timenorm

3.3 Evaluation Metrics

The final output of our system employs the following approach to summarize event triples into patient-level timelines: We begin by using gold-standard DOCTIME annotations for subtask 1. In subtask 2, we predict DOCTIME by analyzing the first sentence of each document, discarding any document that lacks a DOCTIME prediction. Next, we normalize all temporal expressions to a standard format using the TimeNorm package (Laparra et al., 2018; Xu et al., 2019), with DOCTIME as the anchor time. We then de-duplicate timeline entries where chemotherapy events, time expressions, and their relations are identical. Using the timeline summarization system described by Jiarui Yao (2024), we prioritize specific temporal labels from a predefined hierarchy (e.g., BEGINS-ON/ENDS-ON → CONTAINS) for chemotherapy events and only include generic terms like "chemotherapy" in the timeline if there is no mention of a specific drug like "cytoxan" on the same day with the same label.

Performance in this shared task is measured by comparing generated patient-level timelines against gold-standard timelines. Specifically, we evaluate the accuracy of identified tuples containing chemotherapy events, their temporal relations, and time expressions ("chemo EVENT", "temporal_relation", "TIMEX3") compared to the correct timelines. The F1 score is calculated for each patient and then averaged across all patients to yield the macro F1 score. This evaluation employs a relaxed criterion, acknowledging certain temporal relations, specifically "contains-1" with "begins-on" and "contains-1" with "ends-on", as equivalent (Jiarui Yao, 2024).

3.4 Results

In the inference stage, for subtask 1, we directly fed questions to the model to generate output relations. For subtask 2, the model processes each sentence first to extract the chemo event entity. Inspired by Cui et al. (2023), we adopt a sentence window approach to extract associated time expressions. If the target treatment entity is within the target sentence, the model selects k sentence before and after the target sentence to gather contextual information. Due to constraints in time and computing resources, we initially set the window size to zero. If an event entity is detected, we extract the time expression by reprocessing the sentence through the model. Furthermore, to enhance accuracy for subtask 2, we implemented rule-based postprocessing. This approach uses regular expressions to identify and remove inaccurate named entity recognition (NER) predictions for EVENTS and TIMEX3, specifically targeting the pattern associated with chemo entities.

Table 1 shows the official results on the dev set for subtask 1. Our best performance is achieved when instruction tuning with QA and associated documents, leading to a slight accuracy improvement across all disease types, with an overall average score of .68. This indicates the benefit of integrating document context into our training regimen. However, we observed a slight decrease in performance for all three disease types when we continued pretraining on the unlabeled dataset. This decline may be attributed to the limited usage of training data, as we only utilized 1% of the unlabeled data. This did not fully explore the potential of continuous training capabilities, possibly explaining the observed performance dip. Further exploration and more extensive use of the unlabeled data might be necessary to fully optimize the model's performance.

Table 2 shows the official results on the dev set for subtask 2. The model shows variable performance across cancer types, struggling notably with ovarian cancer (.17) and achieving a total average precision of .47. This suggests that subtask 2's entity extraction and relation task is more challenging, especially in complex cancer data.

Table 3 shows the official results on the test set for subtask1. Our method ranks in the mid-tier compared to other teams, with a total average precision of .69. This indicates our approach's competitiveness but also highlights a gap to top-performing models and the baseline.

Table 4 shows the official results on the test set for subtask 2. We face significant challenges, with a total average precision of .22, considerably lower than the baseline. This underscores the complexity of subtask 2 and the need for method improvement.

Overall, our method employs generative LLMs, which, despite their innovative approach, encounter difficulties when competing against traditional state-of-the-art (SOTA) BERT methods in specific tasks like NER and RE. The broad capabilities of generative models aimed at creating new content may not directly translate to the high specificity required for these tasks in the medical domain. This discrepancy is evident in our performance on dev

	Breast			Melanoma			Ovarian			Total Average
	Type A	Type B	Average	Type A	Type B	Average	Type A	Type B	Average	
train QA	.81	.50	.66	.80	.70	.75	.57	.57	.57	.66
+ train QA + DOC	.82	.51	.67	.83	.74	.78	.58	.58	.58	.68
+ train on unlabeled corpus	.77	.39	.58	.80	.70	.75	.56	.56	.56	.63

Table 1: Official results on the dev set for subtask 1.

	Breast			Melanoma			Ovarian			Total Average
	Type A	Type B	Average	Type A	Type B	Average	Type A	Type B	Average	
train QA + Doc	.78	.41	.59	.71	.57	.64	.17	.17	.17	.47

Table 2: Official results on the dev set for subtask 2.

Team	Breast	Melanoma	Ovarian	Total Average
LAILab 1	.96	.87	.88	.90
Wonder 2	.90	.84	.77	.84
NLPeers 1	.72	.81	.75	.77
BioCom 1	.88	.61	.72	.74
Lexicans 1	.68	.83	.61	.71
UTSA-NLP 1 (Ours)	.70	.68	.69	.69
EmoryClincalRXMiners 1	.44	.47	.34	.40
Baseline	.93	.87	.88	.89

Table 3: Official results on the test set for subtask 1.

Model	Breast	Melanoma	Ovarian	Total Average
LAILab 2	.62	.74	.74	.70
KCLab 1	.68	.49	.45	.54
Wonder 3	.63	.39	.55	.53
NYULangone	.19	.32	.18	.23
UTSA-NLP (Ours)	.25	.21	.18	.22
Baseline	.59	.43	.71	.58

Table 4: Official results on the test set for subtask 2.

and test sets, especially for subtask 2, where our approach trails behind the baseline model built based on EntityBERT (Lin et al., 2021). This outcome suggests that leveraging the strengths of generative models for such specific tasks requires a strategic reevaluation of our model's application or methodology.

3.5 Error Analysis

Our error analysis shows that the model is prone to generating false positive relation triples. This issue appears to be rooted in insufficient NULL relation examples during training, leading to the model's poor performance in recognizing the absence of a relationship between EVENT and TIMEX3 entities.

"Gemcitabine used in August 2010 and cisplatin used from March 2012."

For instance, in the above sentence:[2] "Gemcitabine used in August 2010 and cisplatin used from March 2012." In this case, two chemotherapy treatment events are linked with specific time expressions. Our approach to relation extraction involves testing every possible combination of EVENT and TIMEX3 entities, such as Gemcitabine with August 2010, Gemcitabine with March 2012, cisplatin with August 2010, and cisplatin with March 2012. Notably, the combinations of Gemcitabine with March 2012 and cisplatin with August 2010 do not have a temporal relation. Nevertheless, our model erroneously predicts a relation for these pairs. This flaw is primarily due to the difficulty in generating high-quality negative examples for creating QA pairs, which is essential for accurately predicting a NULL relationship.

In subtask 2, we also need to identify EVENT entities accurately. However, generative language models (LLMs) struggle with this, often misidentifying unrelated entities as EVENTS. These errors include categorizing diseases (like "melanoma" or "Parkinson"), diagnostic scans ("FDG PET scan," "CT scan"), diagnostic codes ("PD13-007285PD"), people ("Person2"), and non-chemotherapy medications ("Vicodin," "Zocor") as EVENT entities, despite instructions to exclude them. To address these inaccuracies, we use regular expressions to filter and refine our EVENT entity identification, based on a list of valid chemotherapy events extracted from the training and development sets. This use of regular expressions as a post-processing step ensures the exclusion of these inaccurately named entities.

[2] All examples have been modified and do not directly match the training data to ensure data privacy.

“Patient underwent diagnostic CT scans in June 2012 .”

For example, when analyzing the sentence “Patient underwent diagnostic CT scans in June 2012,” our model incorrectly classifies “diagnostic CT scans” as a chemotherapy EVENT. Although the model explains that “diagnostic CT scans | EVENT | True | As it is crucial for diagnosing the disease and planning chemotherapy,” meaning CT scans are important for diagnosis, not chemotherapy events, the model still wrongly labels them as EVENT entities. This leads to many false positives in identifying entities.

4 Related Work

Continual Knowledge Acquisition. In continual knowledge acquisition, several studies have investigated the ability of language models (LMs) to retain and update knowledge over time. Hu et al. (2023) and Ovadia et al. (2023) explore the effectiveness of different pre-training approaches using smaller LMs like BART (Lewis et al., 2020) and EntityBERT (Lin et al., 2021). Zhu and Li (2023); Jiang et al. (2024); Keloth et al. (2024) delve into fine-tuning LMs on QA pairs related to individuals, with a focus on mixed training settings combining biographies and QA pairs. These studies are a foundation for exploring strategies to incorporate QA data before continued pre-training. Additionally, researchers have sought to adapt LMs to specialized domains, such as medicine, with Li and Zhang (2023); Hu et al. (2024); Zhang et al. (2023) proposing various strategies. However, a common challenge in continual knowledge acquisition is the potential for inaccuracies or difficulties in clinical NLP tasks. Models like BioGPT (Luo et al., 2022), BioMedLM (Bolton et al., 2024), and BioBART (Yuan et al., 2022) address these concerns by continuing training specifically within the medical domain.

Instruction Fine-tuning. Recently, instruction tuning, also known as supervised fine-tuning, has gained prominence for its ability to draw out knowledge from Large Language Models (LLMs) using high-quality annotated data or data from proprietary models (Wei et al., 2021; Zhou et al., 2024; Brokman and Kavuluru, 2024; Zhou et al., 2023). This process enhances LLMs’ capacity to address user inquiries and improves their factual accuracy, a focal point of our research. Additionally, the zero-shot and few-shot in-context learning capabilities of LLMs, which operate with minimal or no training data, present a significant advantage for efficient learning. These approaches, further discussed by Wei et al. (2021) and highlighted in the works of Wang et al. (2024) and Sanh et al. (2021), underscore the potential of instruction tuning in refining LLMs’ factuality and responsiveness.

5 Limitation

Due to the constrained timeline and limited resources of the shared task, our exploration was restricted to basic setups. We did not create negative examples for NER QA pairs and only used a limited set of negative examples for RE QA pairs by randomly selecting three unrelated pairs of chemotherapy events and time expressions. Additionally, our limited use of just 1% of the unlabeled dataset resulted in decreased performance across all three disease types, suggesting that we didn’t fully exploit the continuous training capabilities.

Furthermore, our experiments only considered entities within the same sentence, overlooking cases where entities span multiple sentences in the ChemoTimelines dataset. This oversight could significantly impact model performance evaluation. NER and RE tasks are sensitive to prompt design, and our initial single prompt strategy may not have been optimal. More comprehensive training and experiments, including ablation tests, will be necessary to evaluate and enhance our system’s performance and efficiency thoroughly.

6 Conclusion and Future Work

This paper presents our end-to-end system for extracting Chemotherapy Treatment Timelines from the Clinical NLP ChemoTimelines share the task. We explored various instruction tuning strategies for open-source generative LLMs, providing a starting point for developing NER and RE models in the medical domain. Our future work will explore the implementation of outcome supervision and process-based reward mechanisms in reinforcement learning training to address the issue of false positive predictions (Gao et al., 2024).

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

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.

Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Aviv Brokman and Ramakanth Kavuluru. 2024. How important is domain specificity in language models and instruction finetuning for biomedical relation extraction? *arXiv e-prints*, pages arXiv–2402.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

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.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J Passonneau, and Rui Zhang. 2022. Container: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Jun Gao, Huan Zhao, Wei Wang, Changlong Yu, and Ruifeng Xu. 2024. Eventrl: Enhancing event extraction with outcome supervision for large language models. *arXiv preprint arXiv:2402.11430*.

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 (HEALTH)*, 3(1):1–23.

Sylvain Gugger, L Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, M Sun, and B Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable.

Bernal Jiménez Gutiérrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nathan Hu, Eric Mitchell, Christopher D Manning, and Chelsea Finn. 2023. Meta-learning online adaptation of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4418–4432.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen-tau Yih, and Srinivasan Iyer. 2024. Instruction-tuned language models are better knowledge learners. *arXiv preprint arXiv:2402.12847*.

WonJin Yoon Eli Goldner Guergana Savova Jiarui Yao, Harry Hochheiser. 2024. Overview of the 2024 shared task on chemotherapy treatment timeline extraction.

Vipina K Keloth, Yan Hu, Qianqian Xie, Xueqing Peng, Yan Wang, Andrew Zheng, Melih Selek, Kalpana Raja, Chih Hsuan Wei, Qiao Jin, et al. 2024. Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*, page btae163.

Miloš Košprdić, Nikola Prodanović, Adela Ljajić, Bojana Bašaragin, and Nikola Milošević. 2023. A transformer-based method for zero and few-shot biomedical named entity recognition. *arXiv preprint arXiv:2305.04928*.

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.

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.

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

Guozheng Li, Peng Wang, and Wenjun Ke. 2023. Revisiting large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892.

Mingchen Li and Rui Zhang. 2023. How far is language model from 100% few-shot named entity recognition in medical domain. *arXiv preprint arXiv:2307.00186*.

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.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. 2024. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. 2023. Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics*, 39(5):btad310.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Xilai Ma, Jing Li, and Min Zhang. 2023. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Weijieying Ren, Xinlong Li, Lei Wang, Tianxiang Zhao, and Wei Qin. 2024. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. *arXiv preprint arXiv:2402.18865*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2021. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

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

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023b. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2024. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.

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.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *arXiv preprint arXiv:2005.11184*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*.

Zeyuan Allen Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

This is a section in the appendix.