# Enhancing Temporal Reasoning in Clinical NLP: Challenges, Strategies, and Future Directions for Large Language Models

Zhaokun Wang

March 31, 2025

**Abstract**

Temporal reasoning is essential for understanding the chronological ordering of events in clinical narratives, which is crucial for constructing coherent patient timelines and improving healthcare outcomes. However, despite the remarkable capabilities of Large Language Models (LLMs) in various natural language processing tasks, the application of LLMs to robust temporal reasoning in clinical contexts remains challenging. This term paper for the Temporal Reasoning course provides a review of the current state of temporal reasoning in clinical NLP, focusing on the limitations of LLMs in handling temporal information. The work discusses the inherent challenges posed by the complexity of clinical language, implicit temporal cues, and the need for maintaining chronological coherence across long patient records. Additionally, the paper explores various strategies to enhance LLMs' temporal reasoning capabilities, including prompt engineering, fine-tuning on clinical temporal data, data augmentation, architectural advancements, and integration of external knowledge through knowledge graphs. The importance of developing specialized evaluation metrics and benchmarks tailored to the clinical domain is also highlighted. This paper underscores the need for further investigation and development to improve the application of LLMs in clinical NLP, ultimately aiming to enhance patient care through more accurate and reliable temporal reasoning.

## 1 Introduction

The ability to understand and reason about the temporal ordering of events is fundamental to human cognition, and it's especially important in healthcare. In the healthcare domain, knowing the sequence of events helps doctors to build clear patient timelines. These timelines are like a map that shows a patient's health history in order. They help doctors see how well treatments work, check how patients respond, and spot patterns in health results [1]. While the need for this time-based thinking in healthcare is clear, new technology like Large Language Models (LLMs) have brought

both opportunities and problems for doing this job automatically. LLMs are good at many language tasks, including ones in medicine [2]. They can handle large amounts of medical text, which makes them useful for things like summarizing information, guessing diagnoses, and even tricky tasks [3]. But even though LLMs are powerful, they still struggle with understanding time when it comes to robust temporal reasoning. For example, they have trouble keeping track of event order, figuring out time links that aren't clearly stated, and working with time details across different files. These problems remain even with advanced LLM technology, showing how hard this task is and why we need further investigation. Also, as LLMs are more integrated into healthcare workflows, fixing these time-related issues becomes urgent to make sure they are safe and helpful in real medical work [4]. This paper will look at how well LLMs handle time in medical language tasks, point out their weak spots, and suggest ways to improve. By studying what LLMs can do now, finding their limits, and thinking about how to make them better, this paper hopes to set up ideas for future studies to improve how LLMs work with time in medical information.

## 2 Background on Temporal Reasoning and Clinical NLP

Temporal reasoning is about figuring out time-related details, like the order of events, how long they last, and how they connect to each other [2]. In healthcare, this skill is very important. For example, knowing the order of treatments, when symptoms start and change, and the timing of medical steps helps doctors see if treatments work, guess what might happen to patients, and give better treatment [1]. Electronic Health Records (EHRs), which are the main source of medical data, have lots of time information. When understood well, this data can show a patient's health history in order [1]. But pulling out and working with this time information from medical notes is hard. Clinical notes are often written in unstructured format, not organized like official reports, and they don't always record time details clearly [5]. For instance, a doctor might write "patient felt better after starting medication," which hints at a time connection but doesn't say how long it took. This missing clear time info, plus tricky medical words and different writing styles, makes understanding time in medical texts a big challenge [6]. Over time, people have tried different ways to solve these problems. At first, they used rule-based systems that followed set patterns to find events and time details [1]. Later, as machine learning got popular, models trained on labeled medical texts became more common [1]. The 2012 Informatics for Integrating Biology and the Bedside (i2b2) challenge helped push this work forward by focusing on extracting time info from hospital discharge summaries [6]. This challenge asked people to find medical events, time phrases (like dates), and how they link together. Tools like TLEX (TimeLine EXtraction) were made to help build timelines from texts,

including medical ones, showing the need for automatic ways to do this. Also, systems like TimeText were built to handle time info in medical text, proving that this area keeps growing [7]. The shift from early rule-based tools to advanced machine learning and now to exploring Large Language Models (LLMs) shows a constant effort to make temporal reasoning more accurate and faster, even with the tough parts of medical language.

## 3 Large Language Models for Clinical NLP

Natural language processing has improved a lot with the arrival of Large Language Models (LLMs). These models are trained on huge amounts of text and code, and they can understand and write text that sounds like human [2]. In medicine, LLMs are really good at things like answering patient questions, summing up diagnoses, and writing discharge reports [3]. Some LLMs even do better than human experts in certain medical reasoning tasks, which shows they could change healthcare in big ways. One major way LLMs could help with time understanding in medical texts is by processing and understanding natural language well. This could make it easier to automatically extract and sort time details from messy clinical notes, something that usually takes a lot of human work [1]. Plus, LLMs might be able to figure out time connections that aren't clearly written and look at time info across many documents to create a fuller picture of a patient's health history.

The ChemoTimelines 2024 task is a good example of how people are testing LLMs for building medical timelines [1]. This task was about making systems to extract chemotherapy timelines from EHRs. It included subtasks that challenged participants to both utilize provided gold standard annotations and to directly build timelines from raw clinical notes [8]. The goal was to find important events (like chemotherapy treatments), their time details, and how they connect, to help build better automatic tools for understanding cancer treatment paths [8]. The rise of LLMs and their use in tasks like ChemoTimelines 2024 shows that people are excited to use them for tough medical language problems, like temporal reasoning. While LLMs do well in other NLP jobs, working with time in complicated clinical data has special challenges. We need to look closely at what LLMs struggle with now to understand their limits better. The main thing we learned from the task overview is that smaller language models, trained specially for the job, did better than big LLMs that weren't trained for it. This means that for extracting chemotherapy timelines, tweaking a smaller model for this task might work better than just using a big model's general knowledge without changes. When LLMs didn't do well in Subtask 2, it might be because they messed up at different steps—like missing chemotherapy events, getting time details wrong, or not figuring out how events connect in time. Finding out where these mistakes come from is quite important for making better updates. In

the next parts, we'll look at the challeges that LLMs have with time reasoning in clinical NLP and enhancing strategies to make them better.

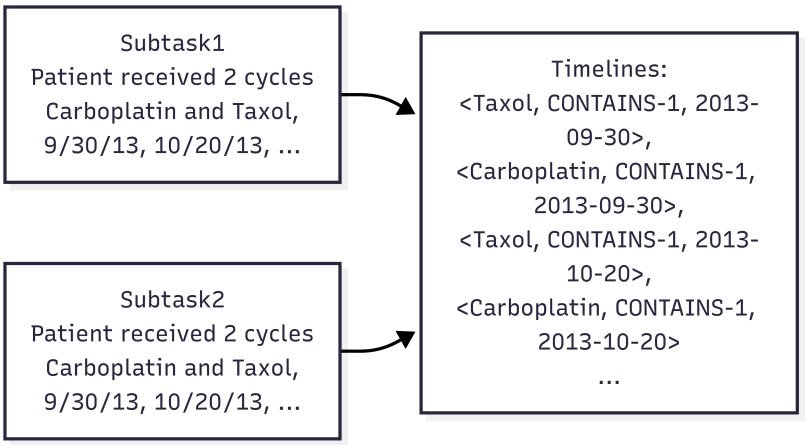


Figure 1: An overview of the chemotherapy share task 2024.

# 4 Challenges

## 4.1 Inherent Limitations of LLMs Affecting Temporal Understanding

Beyond the tough parts of clinical data, Large Language Models (LLMs) have their own built-in weaknesses that can make temporal reasoning tricky, even outside of medicine [2]. These issues often come from how LLMs are designed and trained. While they're awesome for many language tasks, they're not always the best fit for picking up the small but important details of time. One big problem is figuring out the right order of events, especially when time hints are quiet or not obvious [3]. Studies show that not many LLMs are great at temporal reasoning because it's naturally complicated [2]. Even the top models can fall way behind humans in tasks that need time understanding [5]. Some LLMs work by guessing the next word based on what came before, which might make them focus too much on the end of a story and miss how time connects things across a longer stretch [9]. Another hurdle is getting a good grip on how long events last and how often they happen [10]. Time hints in words can point to different parts of an event—like when it starts, stops, or how long it goes on [10]. Basic time tasks, like understanding duration and frequency, still trip up LLMs a lot [5]. For example, telling apart a one-time quick event from something that keeps happening over a long period needs a kind of time sense that LLMs don't always have down pat. Lastly, LLMs can struggle to make sense of and work with relative time phrases like "a week ago" or "two days after surgery," which pop up all the time in clinical notes [11]. Building special tools to clear up these relative time phrases in medical texts shows just how hard this job is [12]. Even the best systems out there

need work to handle these phrases well [11]. The tough part is pinning these loose time references to exact moments or lengths, which often needs background info that LLMs might not fully grasp. These built-in weaknesses hint that the way LLMs are made and trained might not be set up perfectly for tackling the twisty nature of time. Sure, they're fantastic at guessing the next word in a line, but that doesn't always mean they get the full picture of how time ties things together. The data used to train them might not have clear time markers, making it harder for LLMs to pick up and apply time ideas well.

## 4.2 Complexity of Clinical Temporal Information

### 4.2.1 Handling Diverse Temporal Expressions

Clinical text comes with a big mix of temporal expressions, like exact dates and times (like "January 1, 2023, 10:00 AM"), relative phrases (like "a week ago" or "two days after surgery"), and fuzzy ones that aren't clear (like "some time ago" or "recently") [11]. Figuring out and fixing these different expressions is a key step for tougher temporal reasoning jobs, such as pulling out timelines [12]. Temporal Expression Recognition and Normalization (TERN) lays the base for this work, and how well it does really matters to keep mistakes from piling up [12]. The ISO-TimeML system is often used to tag these expressions, sorting them into groups like DATE, TIME, DURATION, and FREQUENCY [12]. While simple terms like "today" are usually handled pretty well by current tools, the many ways to say similar time ideas—like "a while," "some time," or "a moment"—make things a lot harder [13]. Plus, some temporal phrases need a clear starting point to make sense, since they don't give enough info on their own [13]. Studies show that Large Language Models (LLMs) struggle with temporal reasoning, especially in Temporal Relation Extraction (TempRE), even when they're used without any training tweaks. They often do worse than models specially adjusted for the job [14]. This points to how tricky temporal expressions in clinical text can be, creating a real roadblock for even the best LLMs if they haven't been trained or adapted for it. Relative temporal expressions, which show up a lot in clinical notes and are super important for putting events in order on a timeline, need their types sorted out correctly before they can be fixed up properly [12].

### 4.2.2 Problems with Making Time Phrases Clear

Turning time phrases into a standard form, like ISO 8601, is really important for doing time tasks later [12]. Large Language Models (LLMs) are good at understanding and writing text, but making all kinds of time phrases fit one clear standard is still a big problem. Older systems that use rules or basic deep learning don't work well for different areas or languages [13]. Also, new ways of

using LLMs for this are just starting and might not be strong enough for medical use yet [13]. Some LLM methods need extra steps afterward to connect time phrases properly, which shows they don't naturally do the tricky changes and understanding needed to get time phrases right [13].

## 4.3 Limitations of LLMs in Capturing Temporal Dynamics

### 4.3.1 Struggles with Keeping Time Order Straight in Long Clinical Documents

Electronic Health Records (EHRs) pile up a ton of info over a patient's life, sometimes stretching across years. Large Language Models (LLMs) can hit a wall when trying to keep a clear time order across these long stories [3]. Studies looking at zero-shot LLMs summing up hefty clinical texts show they're good at spotting key time events, but they often mess up the order in the summary, especially with really drawn-out records [3]. Research backs this up—LLMs keep running into trouble with time flow in long clinical narratives, even when they get big context windows to work with [3]. What's interesting is that structured data, like tables in EHRs, doesn't always get used well by LLMs, hinting they might not be great at tapping into obvious time clues [3]. The way some LLMs are built—with limits on how much they can focus on at once or a "lost-in-the-middle" issue where they forget stuff in the center of long chunks—can trip them up, making it hard to sort out and reason about the time sequence in lengthy clinical files accurately [3].

### 4.3.2 Challenges in Figuring Out the Right Order and Flow of Medical Events

Clinical narratives often roll out a string of medical events, and nailing down their time order and how they move forward is super important, even if the text doesn't spell it out [11]. Automatically picking up the time links between these events is a must for things like pulling out info or summing up texts later on [7]. Clinical Temporal Relation Extraction (CTRE) steps in big here, piecing together the event order in clinical documents to give a clear view of a patient's medical past [11]. But studies show LLMs don't do so hot in zero-shot temporal relation extraction compared to models trained just for this, pointing to a real struggle in guessing these links without extra training [14]. Getting the time order right often means understanding cause-and-effect, typical disease paths, and standard treatment steps—stuff that might not be fully packed into the LLM's training data. So, LLMs can stumble over hidden time connections that need deeper thinking and medical know-how.

### 4.3.3 Difficulties in Catching Hidden Time Clues and Links

Clinical text likes to drop time info in sneaky ways—through verb tenses, how the story's told, or shared medical know-how. LLMs might find it tough to spot and use these quiet hints for temporal

reasoning [10]. Human language is full of these time signals, and pulling them all together into a solid time picture is a huge part of Temporal Information Extraction (TIE) [10]. On top of that, knowing the usual order and length of events can be key to figuring out what the time stuff really means [10]. Digging into mistakes shows that even top-notch LLMs like GPT-4 trip over the finer points and hidden time clues [5]. This hints that while LLMs rock at handling time stuff that's said straight out, they're not as sharp at piecing together time links from subtle word hints or medical background that's not right there in the text.

## 4.4 Knowledge and Reasoning Problems in Time Context

### 4.4.1 Not Enough Medical Knowledge to Understand Time Information

To do good time reasoning in healthcare, models need special medical knowledge about diseases, treatments, and how things usually happen in clinics. Regular large language models (LLMs) might not have this knowledge, which makes it hard for them to understand time details correctly [15]. Figuring out what time stuff means often depends on knowing the usual order and length of medical events, which needs medical know-how [10]. Studies say adding medical knowledge to time reasoning tools is a big step for the future [16]. To make LLMs trustworthy in medicine, they need clear medical facts they can use [15]. Without enough medical knowledge, LLMs might get the time importance of events wrong or miss important guesses about the whole patient timeline. For example, knowing how long a medicine usually works or how a disease normally grows is key to putting events in the right time order. A regular LLM might not know this unless it's trained on lots of medical texts or connected to a big medical knowledge collection.

### 4.4.2 Struggling to Tell What Time Information Is Good or Old

Medical knowledge and clinic rules change over time as new studies come out and people learn more. LLMs might find it hard to tell the difference between new and old information when they do time reasoning, which can lead to mistakes [15]. Most LLMs are trained on data up to a certain time and can't keep up with the latest medical discoveries [15]. This means they can't easily spot what's the best way to do things now versus old ideas that don't work anymore [15]. To use LLMs in medicine, they'll need regular updates with new research as it comes out [15]. Also, future LLMs should be able to find time-important facts and figure out how medical knowledge changes over time [15]. This problem can mess up time reasoning if it's based on old medical info, which might hurt the safety and quality of tools that help doctors decide things using these models.

### 4.4.3 Needing Stronger Knowledge Thinking for Tricky Time Situations

Time reasoning in healthcare often deals with complicated situations with lots of events and conditions mixed together. LLMs might need better thinking skills based on knowledge to handle these tough spots well [15]. Figuring out new time connections from what's already there often means using logic rules and medical facts [10]. Some time reasoning tools use special knowledge parts to deal with these guesses [7]. Studies show we need more work to add strong knowledge-based thinking to LLMs so they can do tricky, trustworthy, and clear reasoning in medicine based on facts [15]. Even with their progress, current LLMs aren't as good as people at time reasoning, especially for harder things like understanding time stories and cause-and-effect [5]. While LLMs are great at spotting patterns and making text, they might not have the deep thinking skills to work through complicated time links and get the right answers in tough medical cases. For example, with patients who have many long-term illnesses and different treatments, figuring out how these events connect over time and affect each other needs more than just finding time words—it takes logic rules and lots of medical knowledge to guess the tricky time links.

## 4.5 The Need for Specialized Evaluation Metrics and Benchmarks

One big hurdle in pushing Large Language Models (LLMs) to get better at clinical temporal reasoning is that we don't have enough testing tools made just for this field [17]. The usual benchmarks for temporal reasoning might not catch the special twists and needs of clinical work [18]. Sure, they can check how well LLMs handle time order, how long things last, or how often they happen in everyday situations. But they often skip the key types of temporal reasoning that matter in medicine—like tracking disease growth, treatment timelines, or how symptoms tie to diagnoses over time [18]. New benchmarks like TRAM (Temporal Reasoning for large lAnguage Model benchmark) and ToT (Test of Time) show people are working hard to build better ways to test LLMs' temporal reasoning skills [5]. Still, these tools might need more tweaking to really tackle the finer points of the clinical world. When it comes to evaluation metrics for clinical temporal reasoning, the focus should land on results that actually help doctors [17]. Think about things like how correct the patient timelines are, whether the time links tied to diagnoses and treatments hold up, or if LLMs can nail answers to time-based medical questions [17]. Looking at systems like TimeText, which checked how well it built time connections and answered time questions, gives a solid example of testing that matters in medicine [7]. Bringing in benchmarks like TIMER-Bench, built to test temporal reasoning across long-term patient records, takes us a step closer to filling this gap [17]. Putting together trustworthy gold standard annotations for time details in clinical text is no easy task—it's complicated and

takes a lot of time [11]. It needs doctors' know-how and a careful look at tricky, unclear cases. The slow progress since the i2b2 Clinical Temporal Relations Challenge shows just how hard it is to set these standards [11]. Building annotated clinical corpora, like the one in [19], is a huge job that really drives home how much effort goes into making resources to train and test temporal reasoning systems in medicine. Coming up with specialized benchmarks and metrics that fit the unique traits and real-world importance of temporal reasoning in healthcare is super important. It's the key to figuring out what LLMs can really do and pointing the way for future research in this big-deal area.

# 5 Enhancing Strategies

## 5.1 Introduction

Given the vital role of temporal reasoning in healthcare, enhancing LLMs in this domain is crucial. This part explores strategies such as prompt engineering, fine-tuning, data augmentation, model architecture improvements, and knowledge graph integration to address existing limitations. Accurate medical timeline interpretation is essential for patient history analysis and predictive modeling, making advancements in LLMs a key step toward safe clinical deployment. Challenges span multiple processing levels, from parsing text to inferring temporal relationships, necessitating a multifaceted approach to improvement.

## 5.2 Leveraging Prompt Engineering for Improved Temporal Understanding

Prompt engineering—the knack of shaping and tweaking inputs for Large Language Models (LLMs)—is super important for steering these models to tackle specific jobs, like temporal reasoning. By designing prompts that clearly spell out the task and toss in helpful background info, researchers and users can really boost how spot-on and useful the LLM's answers are when it comes to sorting out time stuff. One popular trick in prompt engineering is Chain-of-Thought (CoT) prompting, which nudges LLMs to walk through their thinking one step at a time. This means building prompts with words like "thought," "action," and "observation" to guide the model's brainpower. Take event sequencing, for example—a prompt might look like this:

1. Thought: What's the order of these events?
2. Action: Pick out the main events.
3. Observation: What can we figure out from them?

This setup not only makes the thinking process clearer but also helps the model churn out answers that hang together well—like listing events in time order or puzzling out cause-and-effect over time. Still, some studies hint that while CoT works great for general thinking, it might not always hit the mark for the twisty details of temporal reasoning, suggesting we need sharper prompt ideas [2]. For temporal reasoning with tables, there's a method called C.L.E.A.R. (Comprehend, Locate, Examine, Analyze, Resolve). This step-by-step approach aims to beef up how LLMs handle time links in tabular data. It starts with getting the question's background using field know-how, then finding the right info in the table, checking it out, digging into the time connections, and finally answering based on that. Another cool idea made just for temporal reasoning is Narrative-of-Thought (NoT) [2]. NoT takes a bunch of events, turns them into a Python class, and then asks a smaller language model to whip up a story that's rooted in time [2]. That story then acts as a roadmap for creating a time graph showing how events link up [2]. The big aim of NoT is to tap into LLMs' knack for making and understanding text to build stories, which helps with the trickier job of drawing time graphs—especially when there aren't clear timestamps. NoT's success in closing the gap between big and small LLMs shows that guiding reasoning with a middle-step story can really help models that struggle with tough time puzzles [2]. On top of these tricks, there are some handy tips for prompt engineering in healthcare to get the most out of LLMs for temporal reasoning and other medical tasks [20]. These include being super clear in the prompt, throwing in lots of related details, trying out different prompt styles, stating the main goal upfront, and tweaking prompts based on what the model spits out [20]. For instance, instead of a vague "Tell me about a patient's history," a better prompt might say, "Sum up this patient's key medical events in time order—diagnoses, treatments, procedures, with rough dates." These detailed, context-packed prompts tend to pull out sharper, more useful answers from LLMs in healthcare [20]. Coming up with special prompting methods like NoT shows that researchers are starting to see that one-size-fits-all prompts might not cut it for the unique headaches of temporal reasoning in clinical data. NoT's win with its story-middle-step hints that folks are zeroing in on what this task needs and leaning toward more custom-fit approaches. Plus, how well prompt engineering works ties right into how the prompt lines up with the LLM's inner gears and knowledge stash. With smartly crafted prompts, we can steer the model's focus to the time bits of the input and nudge its reasoning to churn out more accurate, dependable results.

### 5.3 Fine-tuning Large Language Models on Clinical Temporal Data

Fine-tuning is a crucial process in adapting pre-trained LLMs for specific downstream tasks, such as clinical temporal reasoning [21]. This involves taking a model that has already been trained on a massive general-purpose dataset and further training it on a smaller, task-specific dataset [21]. By

adjusting the model's weights based on this new data, fine-tuning allows the LLM to better understand the nuances of the specific domain and improve its performance on the targeted task.

Various fine-tuning strategies can be employed to adapt LLMs for clinical temporal reasoning. Standard fine-tuning involves updating all the parameters of the pre-trained model on the task-specific data [22]. However, with the increasing size of LLMs, this approach can be computationally expensive and require significant resources. Parameter-Efficient Fine-tuning (PEFT) techniques have gained popularity as they allow for efficient adaptation of these large models with a much smaller number of trainable parameters [22]. Examples of PEFT methods include Low-Rank Adaptation (LoRA) and prompt tuning [22]. Studies have shown that hard-prompting with unfrozen LLMs can achieve state-of-the-art results in clinical temporal relation extraction [22]. Furthermore, fine-tuning models using temporal instruction-response pairs, such as with the TIMER-Instruct methodology, has been shown to improve performance on reasoning over EHRs [17]. Ensemble-based fine-tuning strategies, which combine the predictions of multiple fine-tuned models, have also demonstrated the potential to further enhance performance in temporal relation extraction tasks [23].

### 5.4 The Role of Data Augmentation in Enhancing Robustness

Data augmentation is a handy bunch of tricks that beef up the size and mix of training datasets in machine learning [24]. By tweaking what's already there in different ways, it boosts how well models work and helps them handle new stuff better, cutting down on the chance they'll just memorize the training data [24]. When it comes to clinical temporal reasoning with Large Language Models (LLMs), data augmentation can step up big-time to tackle the messiness and variety baked into clinical text. One way to do this is by cooking up fake data using LLMs themselves. With their knack for picking things up from just a few examples, LLMs can be nudged to whip up a pile of synthetic samples—especially in tricky or private areas where real labeled data is hard to come by or costs a lot [24]. For example, in situations with little data, you can give an LLM one example and its label, then tell it to make more like it with the same label for sorting tasks [24]. While this fake data idea sounds promising for filling gaps, there's a catch—some worry it might mess up the model or weaken it if the new stuff isn't varied enough or sneaks in biases [17]. Making sure this synthetic data is top-notch and diverse is key to really toughening up the model. Beyond making new data, you can also shake up what's already there with data reformation tricks to add more flavors on a smaller scale [24]. Classic moves from Natural Language Processing (NLP), like swapping words with synonyms, flipping text into another language and back for a twist, or mixing up words randomly, work great for clinical text in temporal reasoning jobs [25]. There are also rule-based tweaks—where you set up rules to build new examples—and neural tricks, using deep neural

networks trained on other tasks to spice up the data [25]. Given how touchy clinical data can be, privacy-smart data augmentation is super important [26]. One neat trick uses scrubbed patient data as a starting point to guide an LLM in bulking up trial data [26]. This balances the perks of tapping LLMs for more data with the must-do job of keeping patient info safe and secret [26]. Instead of dumping raw patient data straight into the LLM, this method takes a safer middle step, showing how to lift model performance while sticking to ethical and legal rules. The push for privacy-smart data augmentation in the clinical world shines a light on the ethical stuff we can't ignore when handling sensitive patient details. Coming up with ways to sharpen LLMs for tasks like temporal reasoning without risking patient privacy is a big deal for moving these tools forward responsibly in healthcare. Digging into all kinds of data augmentation ideas—from old-school NLP moves to fancier LLM-made fake data—shows how hard people are working to craft models that stand strong and adapt well for clinical temporal reasoning. The wild variety and tangled nature of clinical language mean we need a bunch of different augmentation approaches to train models on a wide mix of examples, helping them stay tough and accurate in real medical settings.

### 5.5 Improving Model Design for Time-Based Data

Researchers are changing how large language models (LLMs) are built to make them better at working with time-related clinical data [17]. These changes try to fix the problems that come with handling information that happens in a sequence, especially in healthcare where patient records cover a long time. One big idea is to mix different types of data, like pictures from medical tests and data that changes over time, into LLMs to better understand a patient's health[27]. Normal LLMs are mostly trained on text and don't use other kinds of data that doctors see in real life[28]. Newer LLMs, called multimodal models, can use text, pictures (like X-rays or MRIs), sounds (like heartbeats), and time-based data (like ECGs or ongoing health readings)[29]. This mix can make time reasoning better by adding extra details, like showing how a disease changes in pictures or tracking body changes over time[29]. Another idea is using Graph Neural Networks (GNNs) to connect patient information and time changes inside LLMs[29]. GNNs can share details between similar patients and show how visits over time are linked. This creates a richer mix of data that can be added to the LLM's middle steps[29]. It helps the LLM use both text and other data to make better guesses by clearly showing patient connections and how their health changes over time [29]. People are also looking at new model designs besides the usual transformer model to handle long sequences faster. Mamba is a new type of LLM that uses selective state space models to fix some transformer problems, especially with very long data sets. By using the Structured State Space (S4) model, Mamba can work with long-term patterns, deal with uneven data, and stay quick during training and testing. This could be

really helpful for handling the big time-based data in patient health records. The basic transformer design, which most LLMs use, relies on self-attention to connect words in a sentence[28]. It works well for many language tasks, but it struggles with very long sequences and tricky time patterns[17]. The attention part gets slower as the data gets longer, which is a problem for full medical histories. This issue pushes researchers to try new designs, like mixing in different data types, using GNNs, and building models like Mamba, to make LLMs better at understanding time in healthcare data, especially when it's long and complicated.

### 5.6 Integrating External Knowledge through Knowledge Graphs

Bringing in outside medical know-how from Knowledge Graphs (KGs) opens up a cool way to boost how well Large Language Models (LLMs) handle clinical temporal reasoning [30]. KGs lay out medical ideas and their connections in a clear, organized way, adding a nice layer to the less obvious stuff LLMs pick up from tons of messy text [30]. Tying the LLMs' reasoning to this neat knowledge can help cut down on wrong info—like made-up stuff (hallucinations)—and fill in gaps where they might lack detailed medical smarts, especially in tricky healthcare areas [30]. Temporal Knowledge Graphs (TKGs) are a special kind of KG that toss in time details, showing facts with timestamps attached [31]. These graphs can track how medical knowledge and patient stories shift over time, making them super handy for sharpening up temporal reasoning in LLMs [31]. Mixing TKGs with LLMs could seriously lift their game—helping them get the order and timing of medical events, figure out time links, and guess what's next based on time patterns [31]. There's a bunch of ways folks are trying to blend this KG knowledge into LLMs, called knowledge fusion. One go-to trick is Retrieval Augmented Generation (RAG), where the LLM grabs useful bits from an outside knowledge stash—like a KG—and mixes it into what it's working with, making its answers richer and better guided [27]. Setups like KARE (Knowledge-Augmented Retrieval with LLM Reasoning) and medIKAL (Integrating Knowledge Graphs as Assistants of LLMs) show how pairing KG info with LLM thinking can lead to sharper healthcare guesses—like better diagnoses and clear reasoning steps [30]. These systems often pull together medical KGs from all sorts of places, like science papers and doctor guidelines, then use them to beef up patient data while the LLM reasons it out [30]. The Unified Medical Language System (UMLS) is a huge medical knowledge bank that's perfect for building these medical KGs [32]. With its giant list of medical terms and their links, the UMLS helps craft KGs that grab a wide chunk of medical know-how useful for clinical temporal reasoning and other healthcare jobs [32]. By tapping into tools like the UMLS, researchers can whip up strong KGs that team up nicely with LLMs to bump up how well they handle and think through time info in clinical texts. The rising buzz around mixing KGs with LLMs shows people get that, yeah, LLMs are awesome at

juggling words, but they can do even better with a dose of organized medical facts to make their answers more accurate, steady, and easy to follow in clinical work. Hooking their reasoning to the clear info in KGs might help LLMs dodge some of their built-in weak spots, delivering more solid and smart results for big healthcare tasks. The ongoing push to figure out the best ways to fuse this knowledge is all about finding the sweet spot between structured and messy info, aiming to build stronger, more dependable AI setups for clinical temporal reasoning.

## 6 Conclusion

This paper provide a review of the current state of temporal reasoning in clinical NLP, highlighting the limitations of LLMs in handling temporal information. We have discussed the challenges posed by the complexity of clinical language, implicit temporal cues, and the need for maintaining chronological coherence across long patient records. Additionally, we have explored various strategies to enhance LLMs' temporal reasoning capabilities, including prompt engineering, fine-tuning on clinical temporal data, data augmentation, architectural advancements, and integration of external knowledge through knowledge graphs.

Despite these efforts, significant challenges remain. LLMs still struggle with maintaining consistency in temporal relationships, inferring implicit temporal connections, and effectively processing temporal information across multiple documents. The increasing integration of LLMs into healthcare workflows amplifies the importance of addressing these temporal reasoning deficiencies to ensure the safe and reliable use of these powerful tools in clinical practice.

Future research should focus on developing more specialized evaluation benchmarks and metrics tailored to the clinical domain, addressing the identified gaps in temporal consistency and implicit information handling, and exploring novel strategies to enhance the temporal reasoning abilities of LLMs. By improving the application of LLMs in clinical NLP, we can ultimately enhance patient care through more accurate and reliable temporal reasoning.

# References


[1] Yukun Tan, Merve Dede, and Ken Chen. 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, 2024.

[2] Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. Narrative-of-thought: Improving temporal reasoning of large language models via recounted narratives. *arXiv preprint arXiv:2410.05558*, 2024.

[3] Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. Zero-shot large language models for long clinical text summarization with temporal reasoning. *arXiv preprint arXiv:2501.18724*, 2025.

[4] Junhyuk Seo, Dasol Choi, Taerim Kim, Won Chul Cha, Minha Kim, Haanju Yoo, Namkee Oh, YongJin Yi, Kye Hwa Lee, and Edward Choi. Evaluation framework of large language models in medical documentation: Development and usability study. *Journal of Medical Internet Research*, 26:e58329, 2024.

[5] Yuqing Wang and Yun Zhao. Tram: Benchmarking temporal reasoning for large language models. *arXiv preprint arXiv:2310.00835*, 2023.

[6] Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association*, 20(5):828–835, 2013.

[7] Li Zhou, Simon Parsons, and George Hripcsak. The evaluation of a temporal reasoning system in processing clinical discharge summaries. *Journal of the American Medical Informatics Association*, 15(1): 99–106, 2008.

[8] Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. Overview of the 2024 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 557–569, 2024.

[9] Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. Will llms replace the encoder-only models in temporal relation classification? *arXiv preprint arXiv:2410.10476*, 2024.

[10] Artuur Leeuwenberg and Marie-Francine Moens. A survey on temporal reasoning for temporal information extraction from text. *Journal of Artificial Intelligence Research*, 66:341–380, 2019.

[11] Amy L Olex and Bridget T McInnes. 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, 2021.

[12] Amy L Olex and Bridget T McInnes. Temporal disambiguation of relative temporal expressions in clinical texts. *Frontiers in Research Metrics and Analytics*, 7:1001266, 2022.

[13] Alejandro Sánchez de Castro, Lourdes Araujo, and Juan Martinez-Romo. Generative llms for multilingual temporal expression normalization. In *ECAI 2024*, pages 3789–3796. IOS Press, 2024.

[14] Vasiliki Kougia, Anastasiia Sedova, Andreas Stephan, Klim Zaporojets, and Benjamin Roth. Analysing zero-shot temporal relation extraction on clinical notes using temporal consistency. *arXiv preprint arXiv:2406.11486*, 2024.

[15] Yifan Peng, Justin F Rousseau, Edward H Shortliffe, and Chunhua Weng. Ai-generated text may have a role in evidence-based medicine. *Nature medicine*, 29(7):1593–1594, 2023.

[16] Carlo Combi and Yuval Shahar. Temporal reasoning and temporal data maintenance in medicine: issues and challenges. *Computers in biology and medicine*, 27(5):353–368, 1997.

[17] Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam Shah. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *arXiv preprint arXiv:2503.04176*, 2025.

[18] Elham Asgari, Nina Montana-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *medRxiv*, pages 2024–09, 2024.

[19] Weiyi Sun. *Time will tell: Temporal reasoning in clinical narratives and beyond*. State University of New York at Albany, 2014.

[20] Bertalan Meskó. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*, 25:e50638, 2023.

[21] Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurélie Névéol, Xavier Tannier, and Christian Lovis. Prompt engineering paradigms for medical applications: Scoping review. *Journal of Medical Internet Research*, 26:e60501, 2024.

[22] Jianping He, Laila Rasmy, Haifang Li, Jianfu Li, Zenan Sun, Evan Yu, Degui Zhi, and Cui Tao. Prompting large language models for clinical temporal relation extraction. *arXiv preprint arXiv:2412.04512*, 2024.

[23] Lijing Wang, Timothy Miller, Steven Bethard, and Guergana Savova. Ensemble-based fine-tuning strategy for temporal relation extraction from the clinical narrative. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 103–108, 2022.

[24] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Luu Anh Tuan, and Shafiq Joty. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1679–1705, 2024.

[25] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101, 2021.

[26] Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Large language models for healthcare data augmentation: An example on patient-trial matching. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1324, 2024.

[27] Tuan Nguyen, Thanh Huynh, Minh Hieu Phan, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Carer-clinical reasoning-enhanced representation for temporal health risk prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10392–10407, 2024.

[28] Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505, 2024.

[29] Inyoung Choi, Sukwon Yun, Jiayi Xin, Jie Peng, Tianlong Chen, and Qi Long. Multimodal graph-llm: Leveraging graph-enhanced llms for multimodal healthcare predictions.

[30] Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *arXiv preprint arXiv:2410.04585*, 2024.

[31] Marco Postiglione, Daniel Bean, Zeljko Kraljevic, Richard JB Dobson, and Vincenzo Moscato. Predicting future disorders via temporal knowledge graphs and medical ontologies. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[32] Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670, 2025.

## Appendix: Use of AI-Based Tools

This appendix documents the use of artificial intelligence (AI)-based tools in the preparation of this academic work.

### List of Steps Involving AI-Based Tools

- **DeepSeek**: I consulted DeepSeek models to learn more formal organization of an abstract and introduction chapter. The suggested frameworks were adapted and rewritten entirely in my own words.
- **QuillBot**: QuillBot was used sparingly to rephrase sentences for improved readability and flow. All suggestions were manually reviewed and edited to ensure alignment with my original intent and academic style.
- **DeepL and Youdao Translation**: DeepL and Youdao Translation assisted in translating a small number of technical terms and short phrases from Chinese to English to clarify meaning during drafting. These translations were verified and incorporated into my own text.