Temporal Reasoning in the Era of LLMs: A Survey

Anonymous ACL submission

Abstract

Temporal reasoning is a critical component of natural language understanding, yet it remains a challenging task due to the inherent ambiguity and implicit nature of temporal information in language. The rise of large language models (LLMs) has sparked interest in assessing their ability to reason about time. However, existing research adopts diverse methodologies, proposing different tasks, benchmarks, and evaluation strategies, making it difficult to form a cohesive view of the field. In this survey, we provide a comprehensive overview of recent work on temporal reasoning in the context of LLMs. We examine the range of tasks, benchmarks, and fine-tuning approaches, and compare these with pre-LLM temporal reasoning tasks. Our analysis reveals that current works, instead of building on previous findings in terms of temporal tasks and datasets, define their own tasks of temporal reasoning and create new datasets to solve them. Finally, we discuss how temporal reasoning evaluation can be advanced to better understand the temporal reasoning capabilities of language models.

1 Introduction

One of the essential aspects of natural language understanding (NLU) is being able to reason about time: to draw correct temporal conclusions from information expressed in language, and to successfully solve other language tasks related to time (Vashishtha et al., 2020). In particular, it refers to the ability of a natural language processing (NLP) system to understand, interpret, and reason about time-related information within a text in order to answer questions or make inferences about events. Temporal reasoning capabilities are essential for understanding narratives, answering time-sensitive questions, and performing commonsense inference. Many real-world applications require temporal understanding, for example, summarization, storytelling, timeline construction, such as in the medical field, where constructing a timeline from a patient's medical records can assist in AI models for healthcare (Sun et al., 2013).

When provided with a text, a model should be able to detect temporal cues and reason about aspects such as temporal relations between events, their duration and frequency, in order to perform a time-related task (which could be formulated as classification or question answering (QA)). However, temporal information is often expressed in text implicitly or in an ambiguous form. Moreover, temporal narratives can be described with complex structures, where the events are not mentioned in chronological order. These issues make the reasoning challenging as it requires extra context or knowledge in order to be able to correctly interpret the temporal information (Leeuwenberg and Moens, 2019). In early works on temporal NLU, this has led to datasets with low inter-annotator agreement, a large amount of "vague" relations, and different annotation schemes proposed, which causes inconsistencies in existing works (Table 1).

While LLMs are state-of-the-art in many tasks (Bubeck et al., 2023), prior work has shown that they struggle with complex and abstract reasoning (Tan et al., 2023; Jain et al., 2023). Moreover, the inconsistencies in temporal reasoning that existed in previous works still hold in the LLM era, with the lack of a unified benchmark to evaluate the temporal reasoning capabilities of LLMs and without a consistent definition of temporal reasoning and its tasks. Current evaluation is largely still limited on accuracy metrics on a few simplified tasks, and the full reasoning capabilities of LLMs remain underexplored (Huang and Chang, 2023). For LLMs, some works have explored their abilities in different settings, such as zero-shot, few-shot, and fine-tuning (Yuan et al., 2023; Kougia et al., 2024; Chan et al., 2024; Zhou et al., 2020; Xiong et al., 2024). The results show that this type of tasks and specifically temporal relation prediction still pose a challenge to LLMs.

Based on the aforementioned challenges, we survey previous attempts to answering the following research question: how good are LLMs in reasoning over time-related concepts in order to solve temporal tasks? We examine two important aspects in order to also give valuable insights for future research on this task. First, the current datasets and benchmarks are presented. Second, the existing approaches for temporal reasoning are discussed.

To this end, we categorize temporal reasoning tasks, and analyze fundamental works in this field. We study the performance of LLMs on temporal reasoning and discuss current evaluation practices. We compare previous deep learning approaches with recent LLM benchmarks and discuss the challenges and future directions towards achieving and evaluating temporal reasoning. To the best of our knowledge, this is the first survey covering all temporal reasoning datasets and benchmarks before and after the emergence of LLMs, giving an overview of their internal relations and addressed tasks.

2 Background

Temporal reasoning preliminaries. Temporal reasoning requires understanding the temporal relations between time elements, such as order, duration, simultaneity, or frequency, and reasoning over them in order to make inferences, predictions, or conclusions based on temporal constraints. For example, "Event A happened at 2 PM. Event B happened at 3 PM. Which came first?" requires reasoning over event order, and "If the train left at 2 PM and the trip takes 3 hours, when does it arrive?" requires reasoning over duration.

However, not all time-related questions require reasoning—some queries can be answered with simple fact retrieval ("What day is New Year's Eve?") or pattern recognition ("The baby slept for 8 hours. How long did she sleep?").

Types of temporal reasoning. Based on the task (e.g., temporal question), different reasoning strategies can be required: 1. *Temporal commonsense reasoning*, 2. *Logic-based temporal reasoning*, 3. *Discourse temporal reasoning*, and 4. *Arithmetic temporal reasoning*.

An important aspect that is often needed for temporal tasks is **temporal commonsense reasoning**. As in general reasoning,[1] this refers to the ability of leveraging everyday knowledge and assumptions humans use to understand and navigate the world such as daily routines and time norms. For example, "The event is at 5 PM. Is this in the afternoon?" involves temporal commonsense reasoning as it depends on conventions about time notation (AM vs. PM) and knowledge of how we divide the day. Time is often expressed in natural language in implicit and vague ways, so world knowledge and commonsense reasoning are often required to solve real-world temporal tasks, sometimes combined with other types of temporal reasoning (Zhou et al., 2019). **Logic-based temporal reasoning** applies temporal logic to solve time-related tasks. For example, the transitivity property can be used to answer questions like "If event A happened before event B, and event B happened before event C, when did event A occur in relation to event C?". This question corresponds to the transitivity rule: A *before* B and B *before* C $\Rightarrow$ A *before* C, and is used to infer the relation of pair (A, C). Another property of temporal logic is symmetry, e.g., A *before* B $\Rightarrow$ B *after* A. The rules that result from these properties can be used to enforce and evaluate the temporal consistency of systems, where a system that follows these rules is considered consistent. Temporal logic is a formal system for representing and reasoning about time using logical operators and strict syntax. Hence, temporal rules are explicit in logic; however, in natural language, they are only implicitly followed, and time in general is expressed in a non-formal way. In these cases, **discourse temporal reasoning** can be applied, which involves understanding narrative flow and reference resolution by interpreting temporal cues and grammatical patterns among others.[2] Finally, **arithmetic temporal reasoning** involves arithmetic estimations of time or calculations between temporal elements.

Datasets for benchmarking temporal reasoning capabilities test those strategies to different degrees, depending on the dataset composition and problem formulation. The term *temporal reasoning* is defined inconsistently among different works, and can mean different types of temporal tasks. For example, *temporal reasoning* refers to temporal relation extraction in Feng et al. (2023), time-sensitive QA

[1] For a detailed explanation of the different types of general reasoning we refer the reader to the survey of Huang and Chang (2023).

[2] A study of temporal cues and how they are expressed in language can be found in the survey by Leeuwenberg and Moens (2019).

in Tan et al. (2023) and 38 different temporal subtasks including duration, arithmetic etc. in Wang and Zhao (2024).

Early approaches to temporal reasoning. One of the first and most influential works on creating a temporal reasoning framework is that by Allen (1983), known as Allen's interval algebra. First, he considered events as time intervals and defined all the possible ordering combinations between two events. A relation was assigned to each combination (13 in total), and then, based on temporal logic, transitivity rules were formed for this relation set. Building on Allen's interval algebra, Pustejovsky et al. (2003a) created TimeML, an annotation scheme designed to annotate events, temporal expressions, and the temporal relations (called TLINKs) between them. TimeML was used to annotate the TimeBank corpus (Pustejovsky et al., 2003b) and the following datasets (Pustejovsky et al., 2010; Cassidy et al., 2014; Styler IV et al., 2014; Ning et al., 2018b; Naik et al., 2019).[3]

Building on these foundations, subsequent works focused on Temporal Information Extraction (TIE), which included the extraction of temporal events and expressions, and the creation of timelines by assigning temporal relations. A lot of these works applied temporal reasoning to different stages of their proposed approach in order to achieve correct and consistent relations. For example, temporal logic was employed for creating the temporal graph closure during the data annotation stage to obtain more relations with less annotation effort (Pustejovsky et al., 2003b; Styler IV et al., 2014; Sun et al., 2013; Naik et al., 2019). Other works have employed temporal logic on classifier predictions, either during training or at inference time, to enhance performance (Tang et al., 2013; Chambers et al., 2014; Ning et al., 2017, 2018a; Wang et al., 2022). Moreover, other approaches have incorporated linguistic or causality-based rules based on how temporal cues are expressed in language and on the fact that temporal and causal relations are known to interact with each other (Chambers et al., 2014; Ning et al., 2018a).

Current approaches to temporal reasoning. Since 2019, there has been significant interest in the field with new annotation efforts (Ning et al., 2020; Zhou et al., 2021; Alsayyahi and Batista-Navarro, 2023; Qin et al., 2021; Tan et al., 2024; Lal et al., 2024) and methods (Feng et al., 2023; Fang et al., 2024; Wei et al., 2023; Su et al., 2024). With the emergence of LLMs, the research interest has shifted to QA benchmarks covering a broader spectrum of temporal phenomena (e.g., event ordering, duration, frequency, etc.) and to evaluating their temporal reasoning capabilities. In the following sections, we provide a detailed overview of these works and describe the findings of our study.

3 Datasets and Benchmarks

Figure 1 depicts a structured overview of the temporal datasets, organized by reasoning focus and task, in line with the structure of this section. In Table 1, we summarize key statistics and details for each dataset.

3.1 Pre-LLM Temporal Datasets

First temporal relation annotation approaches. TimeBank (Pustejovsky et al., 2003b), which resulted from the TimeML guidelines (see Section 2), is the first dataset systematically annotated with temporal relation annotations in natural language text (i.e., news articles). However, TimeBank contains 13 fine-grained relations, which are difficult to annotate (Cassidy et al., 2014; Ning et al., 2018a) and have sparse annotations because the annotators were instructed to label only relations critical to the document's understanding, leaving much of the document unlabeled (Cassidy et al., 2014). This motivated subsequent datasets to create more coarse-grained relation sets by merging some of the original relations (Table 1) (Sun et al., 2013; Cassidy et al., 2014; Ning et al., 2018b; Naik et al., 2019). The TempEval shared tasks (2007–2013) (Verhagen et al., 2007, 2010; UzZaman et al., 2013) focused on identifying temporal relations between events and time expressions using datasets based on TimeBank. Also, the 2012 i2b2 challenge (Sun et al., 2013) addressed temporal relation extraction in Electronic Health Records (EHRs) to support patient timeline construction.

Temporal relation datasets based on TimeBank. In order to address the annotation sparsity, TB-Dense dataset (Cassidy et al., 2014) re-annotates 36 documents from TimeBank, adding denser temporal links between event pairs within one or neighbouring sentences. Building on the same documents, MATRES (Ning et al., 2018b) introduces a new annotation scheme that focuses only on event start points and applies multi-axis modeling to im-

[3] Section 3 presents a detailed overview of temporal relation datasets.

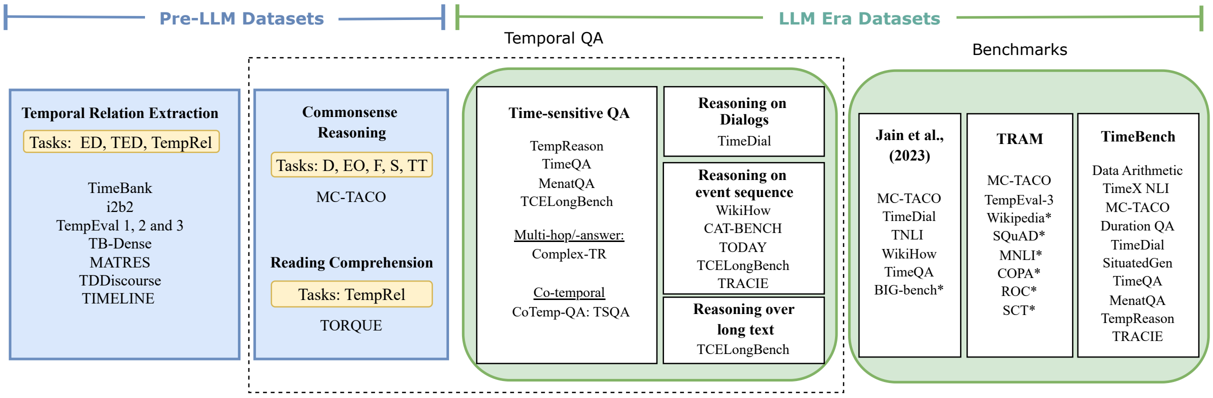

Figure 1: **Overview of temporal reasoning datasets**. The ones with "*" are not initially designed for temporal reasoning tasks. Temporal tasks: event detection (ED), temporal expression detection (TED), and temporal relation extraction (TempRE), ED: Event Duration, EO: Event Ordering, F: Frequency, S: Stationarity, TT: Typical Time.

Dataset	# Documents	# Relations	Frequency
TimeBank (Pustejovsky et al., 2003b)	300	13	-
i2b2 (Sun et al., 2013)	310	3	-
TempEval (UzZaman et al., 2013)	20	13	1
TB-Dense (Cassidy et al., 2014)	36	5	2
TempEvalQA (Llorens et al., 2015)	28	11	1
CaTeRs (Mostafazadeh et al., 2016)	320	4	1
MATRES (Ning et al., 2018b)	36	4	5
TDDiscourse (Naik et al., 2019)	36	5	1
WikiHow (Zhang et al., 2020)	112,505	0	2
TIMELINE (Alsayyahi and Batista-Navarro, 2023)	48	4	1
MC-TACO (Zhou et al., 2019)	13,225	0	5
TORQUE (Ning et al., 2020)	3,200	2	1
TRACIE (Zhou et al., 2021)	5,400	2	2
TIME-DIAL (Qin et al., 2021)	1,100	0	2
TimeQA (Chen et al., 2021)	41,200	4	4
SituatedQA (Zhang and Choi, 2021)	12,200	0	1
TempLAMA (Dhingra et al., 2022)	50,000	0	1
TempReason (Tan et al., 2023)	52,800	2	2
TODAY (Feng et al., 2023)	2,241	2	1
MenatQA (Wei et al., 2023)	2,853	0	1
Complex-TR (Tan et al., 2024)	10,800	0	1
CAT-BENCH (Lal et al., 2024)	4,260	2	1
TCELongBench (Zhang et al., 2024b)	88,821	0	1
CoTemp-QA (Su et al., 2024)	4,748	4	1

Table 1: **Datasets with temporal tasks**. We show the dataset sizes (# Documents), annotated relations, tasks included—event detection (ED), temporal expression detection (TED), and temporal relation extraction (TempRE)—and average number of words per document.

prove inter-annotator agreement by excluding temporally incomparable events. Extending TB-Dense further, TDDiscourse (Naik et al., 2019) adds annotations for long-distance temporal relations across sentences.

Temporal datasets beyond the TimeBank corpus. In 2019, Zhou et al. (2019) introduced MC-TACO, a multiple choice QA dataset targeting five commonsense temporal aspects (i.e., duration, temporal ordering, typical time, frequency, and stationarity), using single-sentence contexts from the MultiRC dataset (Khashabi et al., 2018). TORQUE (Ning et al., 2020) features human-written questions about temporal relations between events that are mostly implicitly mentioned in short news contexts, enabling deeper temporal reasoning evaluation. In a recent temporal relation annotation effort, Alsayyahi and Batista-Navarro (2023) published TIMELINE, which introduces a multi-axis annotation scheme where annotators answer questions about event pairs, and an algorithm infers temporal relations automatically.

3.2 Temporal Datasets in the LLM Era

Temporal QA datasets. Unlike the pre-LLM temporal annotation efforts that mainly focused on relation prediction, recent LLM-era datasets cover a broader range of temporal aspects and tasks, including event frequency and duration prediction, temporal NLI, time-sensitive QA, dialogue-based temporal reasoning, and multi-hop inference. The TRACIE dataset (Zhou et al., 2021) focuses on implicit events, hence testing models' abilities to interpret timelines in narratives, often requiring commonsense reasoning. TIMEDIAL (Qin et al., 2021) is a multiple-choice cloze QA dataset that tests temporal understanding in multi-turn dialogues by asking models to fill in missing time expressions. Tan et al. (2023) published a QA dataset called TempReason extracted from WikiData in order to serve as a benchmark for temporal reasoning. They defined three temporal reasoning levels and constructed the questions based on them: 1. time-time relations, 2. time-event relations, 3. event-event relations. Complex-TR (Tan et al., 2024), inspired by TempReason, is a multi-hop, multi-answer temporal QA dataset designed to probe complex reasoning over co-occurrence and event sequences. CAT-BENCH (Lal et al., 2024) is a benchmark focusing on step ordering. It evaluates whether a particular step in a plan must occur before or after another, emphasizing the understanding of causal and temporal dependencies. TCELongBench (Zhang et al., 2024b) focuses on temporal, long context evaluation of QA pairs, tailored to three distinct tasks: 1)TLB-detail QA, which tests LLMs' ability to find evidence across numerous articles; 2)TLB-order QA, focusing on understanding temporal sequences; and 3)TLB-forecast QA, challenging LLMs to predict future events based on past information. CoTemp-QA (Su et al., 2024) studies another aspect, which they call co-temporal reasoning. It assesses LLMs' capabilities to comprehend and reason about events that occur concurrently or have overlapping durations. The dataset is constructed from real-world temporal facts, including biographical data of notable individuals.

A different direction of research has employed techniques like modifications or contradictions of the context in order to evaluate the performance of the models when these changes are introduced. (Feng et al., 2023) proposed a new task and dataset called TODAY in which human annotators are asked to write a sentence that, if added at the beginning of the context, can change the current relation between two events. The annotators also write an explanation of how that change will occur. Wei et al. (2023) introduced MenatQA, which is based on TimeQA (Chen et al., 2021), but they added changes to the original context or question of each instance to make the dataset more complex. The changes have three categories: 1. Scope: The time range in a question is shifted so that it is not the same as the range mentioned in the context, 2. Order: the events in a context are shuffled so that they are not mentioned chronologically in the context, and 3. Counterfactual: a temporal hypothesis that contradicts the context is added.

Temporal reasoning LLM benchmarks. Several recent benchmarks have been introduced to assess the temporal reasoning capabilities of LLMs using existing datasets including a wide range of temporal aspects (see Fig. 1). Jain et al. (2023) first evaluated eight LLMs across six datasets, while Wang and Zhao (2024) expanded this effort into a larger multiple-choice QA benchmark with 8 datasets covering 38 temporal subtasks. Most recently, TIMEBENCH merged 10 datasets into 16 fine-grained temporal subtasks to enable a more comprehensive evaluation framework (Chu et al., 2024).

4 Temporal Reasoning Methods

4.1 Temporal training

Pretraining with Temporal Span Masking. A widely used approach for improving temporal QA performance is Temporal Span Masking (TSM), which builds upon the Salient Span Masking (SSM) method, which involves reconstructing masked named entities as a language model pretraining objective (Guu et al., 2020). TSM extends this idea by masking temporal expressions—such as specific dates, time durations, or recurring temporal phrases—instead of or alongside named entities (Tan et al., 2023; Qin et al., 2021; Cole et al., 2023). This adaptation aims to enhance the model's understanding of temporal information during pretraining. Qin et al. (2021) and Cole et al. (2023) trained base models using TSM and/or SSM, experimenting with different configurations of masked spans and model variants. Going a step further, Tan et al. (2023) introduced a method called Time-Sensitive Reinforcement Learning (TSRL). After initial TSM and SSM pretraining, they construct negative answer sets for each question—answers that are structurally correct in terms of subject and relation but pertain to incorrect time periods—and use a reward function that penalizes selections from these negative sets. Positive rewards are assigned when model predictions match the correct, temporally aligned answers, reinforcing temporal sensitivity in the decision-making process.

Fine-tuning strategies. Several methods have been proposed to fine-tune pretrained models for improved temporal reasoning. Feng et al. (2023)

Dataset	Model					Human	
	Ref.	Type	Setting	Acc	F1	Acc	F1
MC-TACO	Chu et al. (2024)	GPT-4	Few-Shot	-	88.3	-	87.1
MATRES	Roccabruna et al. (2024)	Llama2 13B	Fine-Tuning	84.3	-	88.0	-
WikiHow	Jain et al. (2023)	Llama 7B, GPT-3.5	Few-Shot	55.0	-	98.0	-
TRACIE	Chu et al. (2024)	Llama2 70B	Few-Shot + CoT	67.0	-	82.5	-
TimeDial	Chu et al. (2024)	GPT-4	Few-Shot	94.6	-	97.8	-
TempReason	Chu et al. (2024)	GPT-4	Few-Shot + CoT	92.4	-	97.1	-
TimeQA	Chu et al. (2024)	GPT-4	Few-Shot	-	73.7	-	93.3

Table 2: For each of the seven most commonly used datasets (i.e., with frequency more than one based on Table 1), we show the best LLM performance across all the papers it was used in and the human performance.

propose a joint learning framework using TODAY, a dataset described in Section 3, which is annotated with distributional shifts rather than absolute temporal labels. The task is framed as textual entailment, where the premise includes the additional sentence, the context, and the question, and the hypothesis includes two events and the relation between them. Two entailment instances are generated per example: one for “before” and one for ‘after”. The model is trained using cross-entropy loss on hard labels from datasets such as MATRES or TRACIE, and marginal ranking loss on the relative annotations in TODAY, enabling it to perform well across both absolute and relative reasoning cases.

To address the long-context challenge in open-domain QA, Tan et al. (2024) introduce a combined data augmentation and context refinement strategy. First, they generate pseudo-instruction tuning data by shifting the time ranges in original instances and prompting ChatGPT (OpenAI, 2024) to hallucinate fictional entities to be added in the questions. Then, to reduce input length while preserving relevance, they apply cosine similarity over sentence embeddings to select the paragraphs from the context that are the most relevant to each question.

Following a different fine-tuning direction, Yang et al. (2024) incorporate contrastive and reinforcement learning techniques. They introduce time-aware embeddings derived from temporal expressions in the input and use a Granular Contrastive Reinforcement Learning objective. This approach evaluates model responses based on semantic and temporal vector proximity to correct and incorrect answers, offering a reward signal that encourages temporally robust predictions beyond string-level matching. All of the aforementioned methods employed T5 (Raffel et al., 2020) as the base model.

General-purpose temporal reasoning LLMs. Expanding beyond task-specific models, other works aim to equip large language models with general temporal reasoning abilities. Xiong et al. (2024) propose TG-LLM, a fine-tuning framework that improves temporal understanding through translation of textual input into latent temporal graph representations. As part of this effort, they introduce a synthetic dataset, TGQA, specifically designed to support training for temporal graph construction. Similarly, Su et al. (2024) develop TIMO, a temporal reasoning framework trained with TRAM (Wang and Zhao, 2024), which categorizes tasks into mathematical time and pure time reasoning. TIMO employs a self-critic optimization approach in which a reward model based on a formal mathematical evaluator scores generated responses, enabling reinforcement learning to guide the model toward higher-quality, temporally sound answers.

4.2 Prompting

Prompting plays a central role in temporal QA tasks, affecting how models interpret questions and retrieve or reason over temporal information. Here, we present an overview of existing works and the various dimensions of prompt design that are explored, including question formulation, context format, prompting settings, and specialized strategies tailored for temporal reasoning.

Prompt formulations. The formulation of the QA task guides the structure and content of the prompt. Common QA types include open-book and closed-book QA (Tan et al., 2023), open-domain QA (Tan et al., 2024), cloze-style QA (Qin et al., 2021), multiple-choice formats (Qin et al., 2021; Wang and Zhao, 2024; Fang et al., 2024), yes/no questions (Lal et al., 2024; Kougia et al., 2024), and free-form answers (Zhang et al., 2024a; Chu et al., 2024; Su et al., 2024). Many studies adopt multiple formulations to evaluate performance across settings (Tan et al., 2023; Qiu et al., 2023; Yuan et al., 2023).

Prompt settings. Based on the number of examples and reasoning style, prompting can be categorized into zero-shot, few-shot, and Chain-of-Thought (CoT) prompting. Most studies explore all three, while some focus only on zero-shot setups (Kougia et al., 2024; Wei et al., 2023). For few-shot, approaches vary from one-shot (Tan et al., 2024), to C-shot (Chan et al., 2024; Roccabruna et al., 2024),[4] and 5-shot (Wang and Zhao, 2024). CoT prompting has also been integrated into several systems (Wang and Zhao, 2024; Lal et al., 2024; Chu et al., 2024; Su et al., 2024; Qiu et al., 2023) to improve temporal reasoning capabilities showing promising performance.

Prompt context. The context provided in prompts varies significantly. Most works include some form of context, such as free-text passages (Jain et al., 2023; Wang and Zhao, 2024; Yuan et al., 2023; Wei et al., 2023; Chan et al., 2024; Qiu et al., 2023; Chu et al., 2024; Su et al., 2024; Fang et al., 2024), while others use structured formats like code (Zhang et al., 2024a). Notably, Lal et al. (2024) avoid providing explicit context altogether. This variation in context design influences how models reason temporally across different tasks.

Other prompting strategies for temporal reasoning. Beyond the standard prompting approaches, some studies have introduced additional strategies designed for temporal reasoning. Jain et al. (2023) use code prompts that structure input as code-like syntax to guide LLMs. Yuan et al. (2023) propose event-ranking prompts that require ordering events relative to a reference. Zhang et al. (2024a) introduce the Narrative-of-Thought (NoT) method, which first elicits a temporally grounded narrative before generating answers using a Temporal Graph Prompt. This end-to-end CoT-based approach enhances the model's use of temporal structure in reasoning.

5 Findings

How do LLMs perform on temporal reasoning tasks? All the works we have studied for this survey report that LLMs struggle with temporal reasoning tasks and perform worse than smaller supervised models, e.g., BERT, and humans. As shown in Table 2, for all the datasets except for MC-TACO,[5] The LLM performance can be up to 43% lower than the human one. Moreover, while LLMs can handle simpler tasks like assessing event duration or frequency (Jain et al., 2023), they frequently fail at more complex tasks such as predicting correct event sequences, especially when faced with conflicting knowledge, counterfactuals, or multi-hop reasoning (Fang et al., 2024; Feng et al., 2023; Wei et al., 2023). This indicates a gap in deeper temporal understanding and inference.

Can temporal fine-tuning approaches improve performance on temporal tasks? Results of the pre-training and fine-tuning approaches show that they always yield improvements over the base model, and sometimes even outperform larger LLMs like Flan-T5-Large and GPT-3.5 in zero-shot settings (Tan et al., 2023; Yang et al., 2024) or GPT-3.5 and GPT-4 in one-shot settings (Tan et al., 2024). Moreover, fine-tuning allows models to generalize better across different datasets, showing robustness in complex tasks such as multi-hop reasoning and co-temporal inference (Feng et al., 2023; Su et al., 2024; Zhang et al., 2024b; Tan et al., 2024). Models like TIMO (Su et al., 2024) and TODAY-trained variants (Feng et al., 2023) significantly outperform GPT-4 on specific relation extraction benchmarks, despite having fewer parameters, illustrating the power of targeted training. Fine-tuned models also mitigate some common LLM failures, such as inconsistencies in temporal symmetry and bias towards contemporary dates (Yuan et al., 2023; Qiu et al., 2023).

Which LLMs and of what size have shown better performance? Among the models surveyed, newer commercial LLMs such as GPT-4 consistently achieve the strongest overall performance in temporal reasoning tasks, particularly when evaluated under few-shot and CoT prompting settings (Wang and Zhao, 2024; Zhao and Rios, 2024; Tan et al., 2024; Chu et al., 2024). However, the ability to systematically compare commercial models remains limited, as they are often evaluated only on reduced subsets of datasets due to cost and API constraints (Wang and Zhao, 2024; Tan et al., 2023). Additionally, while increasing model size generally correlates with better temporal reasoning ability,

[4] C refers to the number of classes in the task, e.g., temporal relations.

[5] As the authors of MC-TACO mention, human performance is low because commonsense can vary between individuals, so a single person's answer might not always match the gold label.

several studies report diminishing returns beyond a certain scale (Qiu et al., 2023; Xiong et al., 2024; Zhang et al., 2024a; Feng et al., 2023). This suggests that targeted finetuning and training strategies can be more impactful than simply scaling up model size.

Which prompting strategy has the best performance? As indicated in Table 2, few-shot learning—especially when combined with CoT reasoning—produces significantly better results than zero-shot prompting, which often yields the weakest performance (Qiu et al., 2023; Tan et al., 2024; Chan et al., 2024). This gap underscores the importance of example-based guidance for complex temporal tasks.

6 Discussion and future directions

Gaps and fragmentation in temporal reasoning research. Despite growing interest and a surge of new benchmarks, many studies do not build effectively on prior findings. Key insights—such as LLMs' inconsistent predictions (Kougia et al., 2024), difficulties with counterfactuals and conditions (Feng et al., 2023; Wei et al., 2023), and inability to produce human-like reasoning (Lal et al., 2024) remain underexplored. Another major challenge is the lack of a unified definition of temporal reasoning, leading to fragmented efforts where each work emphasizes novelty over continuity. As seen in Table 1, most recent datasets have been used only once, highlighting limited integration and reuse.

There is also a disconnect between newly developed QA datasets and pre-LLM temporal relation datasets. Except for MC-TACO datasets with more complex temporal relations are rarely used in LLM evaluations, with preference given to simpler ordering datasets (only “before” and “after” relations). Bridging this gap by incorporating richer temporal relation datasets could significantly deepen our understanding of LLMs' reasoning abilities and encourage more cohesive progress in the field.

Evaluating temporal reasoning capabilities As discussed in Section 2, the boundary between retrieval and reasoning is sometimes unclear, especially for templated or simple questions. For instance, "When do people usually eat breakfast?" could be answered either through memorized knowledge or temporal commonsense reasoning. Similarly, "Does 3:00 PM come before 5:00 PM?" might rely on arithmetic reasoning or pattern matching. In contrast, more complex questions are more likely to require genuine reasoning. For example, computing a future date ("What day is 100 days after March 3, 2025?"), or reasoning about contradictions ("He spoke during a meeting that ended at 3 PM but arrived afterward—can that be true?").

Tan et al. (2023) reported that LLM performance deteriorates from year to month prediction. Furthermore, many works have reported lower LLM performance for complex temporal tasks (Tan et al., 2023; Lal et al., 2024), tasks including long dependencies (Chan et al., 2024) and counterfactuals (Fang et al., 2024) concluding that our understanding of LLMs' temporal reasoning capabilities may be misleading (Feng et al., 2023). As a result, there is a clear need for more rigorous and challenging benchmarks to accurately assess LLM performance on temporal reasoning.

7 Conclusion

With the advanced capabilities of LLMs, interest in their temporal understanding has increased. Yet, temporal reasoning is understood in different ways, and often narrowly evaluated on simplified tasks, such as question answering with limited context. Sometimes, current evaluations are even a step backwards from pre-LLM evaluations (that, e.g., considered larger contexts and more complex temporal relations). Our survey discusses core aspects of temporal reasoning, including a taxonomy of associated tasks, and a comprehensive overview of existing approaches to solve these tasks, as well as datasets for benchmarking their performance. We find that LLMs continue to struggle with temporal reasoning, especially on tasks involving complex inferences such as co-temporal relations, often under-performing compared to smaller, fine-tuned models, and do not reach human-level competency. We survey previous work on different strategies for improving the temporal capabilities of LLMs. We find that prompting strategies alone remain insufficient compared to fine-tuning and task-specific pretraining, which in turn consistently improve performance and generalization. Our findings underscore the need for clearer task definitions, unified benchmarks, and diagnostic evaluations.

Limitations

Due to the rapid evolution of LLMs and the frequency of new benchmark releases, some very recent models or datasets may not be fully covered.

Our analysis focuses primarily on published and publicly available resources up to early 2025, and does not include proprietary data or unpublished evaluations.

The diversity in task formulations, evaluation protocols, and dataset structures across studies makes direct performance comparisons challenging. Although we attempt to unify definitions and identify commonalities, differences in experimental setups may limit the generalization of some conclusions.

Finally, while we analyze model performance and prompting strategies in depth, we do not conduct new empirical experiments. Future work could complement this survey with large-scale empirical evaluations under standardized conditions to more precisely assess temporal reasoning capabilities across models.

References

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Sarah Alsayyahi and Riza Batista-Navarro. 2023. Timeline: Exhaustive annotation of temporal relations supporting the automatic ordering of events in news articles. *arXiv preprint arXiv:2310.17802*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228, Bangkok, Thailand.

Jeremy R. Cole, Aditi Chaudhary, Bhuwan Dhingra, and Partha Talukdar. 2023. Salient span masking for temporal understanding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3052–3060, Dubrovnik, Croatia.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Tianqing Fang, Zhaowei Wang, Wenxuan Zhou, Hongming Zhang, Yangqiu Song, and Muhao Chen. 2024. Getting sick after seeing a doctor? diagnosing and mitigating knowledge conflicts in event temporal reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3846–3868, Mexico City, Mexico.

Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin, and Dan Roth. 2023. Generic temporal reasoning with differential analysis and explanation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12013–12029, Toronto, Canada.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774, Singapore.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings* *of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

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

Yash Kumar Lal, Vanya Cohen, Nathanael Chambers, Niranjan Balasubramanian, and Ray Mooney. 2024. CaT-bench: Benchmarking language model understanding of causal and temporal dependencies in plans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19336–19354, Miami, Florida, USA.

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

Hector Llorens, Estela Saquete, Borja Navarro, and Estela Saquete. 2015. Semeval-2015 task 5: Qa tempeval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 45–54, Denver, Colorado. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. Tddiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. A multi-axis annotation scheme for event temporal relations. *arXiv preprint arXiv:1804.07828*.

OpenAI. 2024. Chatgpt. Large language model (May 19 version). Available at `https://chat.openai.com/`. Response to prompt: "create citation for chatgpt.".

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The timebank corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster, UK.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2023. Are large language models temporally grounded? *arXiv preprint arXiv:2311.08398*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. 2024. Will LLMs replace the encoder-only models in temporal relation classification? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20402–20415, Miami, Florida, USA.

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

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and

Min Zhang. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13014–13033, Bangkok, Thailand.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2024. Towards robust temporal reasoning of large language models via a multi-hop QA dataset and pseudo-instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6272–6286, Bangkok, Thailand.

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

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden.

Liang Wang, Peifeng Li, and Sheng Xu. 2022. DCT-Centered Temporal Relation Extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2087–2097, Gyeongju, Republic of Korea.

Yuqing Wang and Yun Zhao. 2024. Tram: Benchmarking temporal reasoning for large language models. *arXiv preprint:2310.00835*.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. MenatQA: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1434–1447, Singapore.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.

Wanqi Yang, Yanda Li, Meng Fang, and Ling Chen. 2024. Enhancing temporal sensitivity and reasoning for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14495–14508, Miami, Florida, USA.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

Michael JQ Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*.

Xinliang Frederick Zhang, Nicholas Beauchamp, and Lu Wang. 2024a. Narrative-of-thought: Improving temporal reasoning of large language models via recounted narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16507–16530, Miami, Florida, USA.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024b. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1588–1606, Bangkok, Thailand.

Xingmeng Zhao and Anthony Rios. 2024. UTSA-NLP at ChemoTimelines 2024: Evaluating instruction-tuned language models for temporal relation extraction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 604–615, Mexico City, Mexico.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. *arXiv preprint arXiv:2005.04304*.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online.

A Appendix

A.1 Paper Selection

In order to select the papers mentioned in this survey, we first include papers introducing datasets with temporal relations and papers that discuss temporal logic/ annotation schemes. Then, we select papers that perform temporal reasoning from 2019 onward since we focus on LLMs and the last temporal reasoning survey was published in 2019 (Leeuwenberg and Moens, 2019). Hence, we searched the existing literature with the keywords: *temporal relation, temporal corpus, temporal annotation, temporal logic, temporal reasoning, time reasoning, temporal understanding, temporal language model, time language model, temporal survey, temporal review, time survey, time review, temporal ordering, temporal information extraction*. The keyword search was performed on the titles of the papers. The initial search included 304 papers, from which we filtered out papers containing the words *temporal knowledge graph* and *video* as we focus on papers that work on textual input. The resulting 257 papers were then manually checked regarding their relevance to the scope of this survey (temporal reasoning and experiments with LLMs). After this, the relevant papers we found were 28. We also studied 10 more papers that introduced datasets used or mentioned in the related work of the original set of 28, and are relevant to our survey, but did not come up during the search.