It's High Time: A Survey of Temporal Information Retrieval and Question Answering

Bhawna Piryani,[♠] Abdelrahman Abdallah,[♠] Jamshid Mozafari,[♠] Avishek Anand,[♡] Adam Jatowt[♠]
[♠]University of Innsbruck
{bhawna.piryani, abdelrahman.abdallah,jamshid.mozafari, adam.jatowt}@uibk.ac.at
[♡]TU Delft
avishek.anand@tudelft.nl

Abstract

Time plays a critical role in how information is generated, retrieved, and interpreted. In this survey, we provide a comprehensive overview of *Temporal Information Retrieval* and *Temporal Question Answering*, two research areas aimed at handling and understanding time-sensitive information. As the amount of time-stamped content from sources like news articles, web archives, and knowledge bases increases, systems must address challenges such as detecting temporal intent, normalizing time expressions, ordering events, and reasoning over evolving or ambiguous facts. These challenges are critical across many dynamic and time-sensitive domains, from news and encyclopedias to science, history, and social media. We review both traditional approaches and modern neural methods, including those that use transformer models and Large Language Models (LLMs). We also review recent advances in temporal language modeling, multi-hop reasoning, and retrieval-augmented generation (RAG), alongside benchmark datasets and evaluation strategies that test temporal robustness, recency awareness, and generalization.

1 Introduction

From analyzing centuries-old texts, understanding historical events, to answering questions about emerging developments, time shapes how we seek and interpret information. As digital content continues to grow exponentially across time-stamped sources like news archives, social media, and knowledge bases, the ability to process and reason over temporal information has become essential (Alonso et al., 2007). Temporal IR, which searches time-stamped documents, and Temporal QA, which answers time-sensitive queries, together address these needs. Both, collectively referred as Temporal IR/QA, aim to incorporate time-awareness to adapt results to specific periods and resolve time-sensitive queries (Campos et al., 2014).

Temporal IR/QA faces distinct challenges that set it apart from standard IR/QA settings. These include identifying temporal intent in queries, interpreting expressions such as "post-World War II" or "in 1998," and modeling relationships between events and their timelines (Berberich et al., 2010). Queries may target past, present, or future events, and require systems to identify relevant time frames, order events, and resolve implicit temporal cues. Overcoming these obstacles demands methods that extend beyond traditional keyword-based search and basic retrieval techniques.

For Example, in Figure 1, *Q1: "At what age did Obama win the Nobel Peace Prize?"* requires constructing a chronology of events by identifying and grounding two temporal anchors, Obama's birth year (1961) and the year he received the Nobel Peace Prize (2009). The model must then establish a temporal relationship and apply reasoning to compute the answer: *48 years old. Q2: "What does President Obama's climate policy tell us about how the U.S. viewed climate change during his late years of service?"* demands understanding of the query's intended time and contextual temporal grounding. The model must recognize relative temporal expressions such as *"today, next week"* and associate them with a reference time such as the document's publication date. It also needs to retrieve or reason over documents written during the relevant policy timeframe, reconstructing the contemporaneous narrative.

Research in Temporal IR/QA has evolved significantly, building on early foundations to address increasingly complex temporal challenges. Initial efforts relied on rule-based systems (Harabagiu and Bejan, 2005) and statistical models (Berberich et al., 2010) that used document timestamps and hand-crafted rules to interpret time-related information (Li and Croft, 2003). While these methods established key principles, they struggled to scale or handle diverse temporal contexts. The rise of

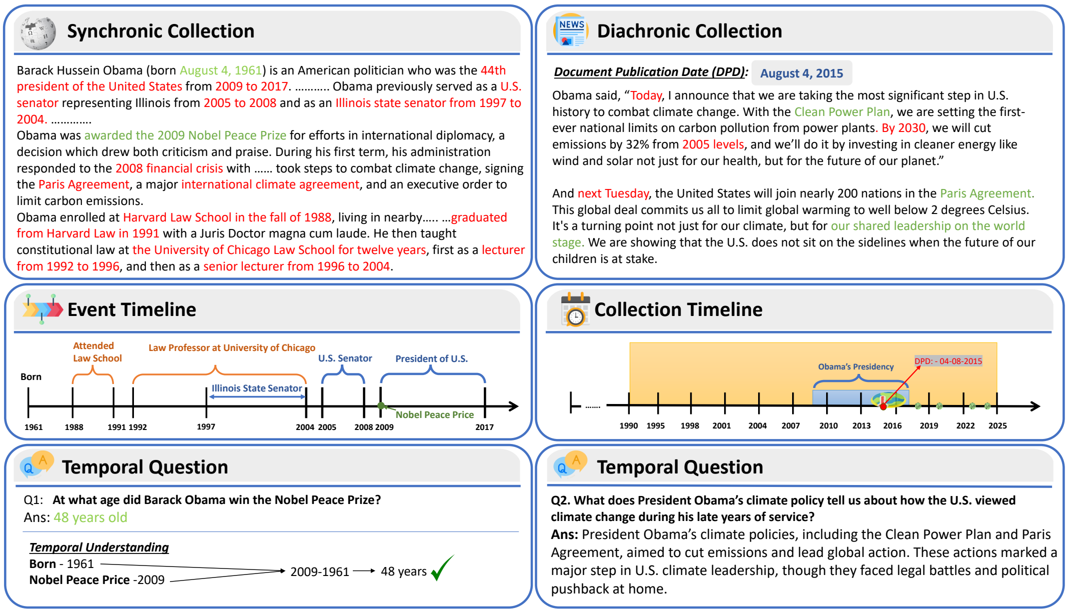

Figure 1: Examples of documents from synchronic (left) and diachronic (right) collections. Red highlights temporal signals present in the documents, while green indicates the answer to the questions (bottom). The event timeline built from the synchronic document on the left presents the inferred sequence and duration of events. On the other hand, the collection timeline represents the time span of the Diachronic collection. Red dots there mark documents that contain the answer, and green points indicate documents published related to the question Q2's event over time.

pre-trained language models has transformed the field by enabling robust temporal reasoning (Jain et al., 2023), event sequencing (Lin et al., 2021), and adaptation to evolving knowledge (Han et al., 2021). These advancements paved the way for more dynamic and scalable temporal systems.

While prior surveys[1] have explored general IR/QA methods (Robertson et al., 2009; Xiong et al., 2020; Formal et al., 2021), or focused narrowly on specific aspects of temporal processing within one of these fields (Kobayashi and Takeda, 2000; Kanhabua et al., 2015; Campos et al., 2014), a comprehensive and unified overview of Temporal IR/QA is long overdue. The most recent dedicated survey in this area was published nearly a decade ago (Campos et al., 2014). Since then, the field has grown substantially, driven by advances in Language Models, new datasets, and complex temporal tasks. A survey that captures recent advancements and outlines future research directions is then essential to foster progress and guide the community in developing time-aware systems. Our paper addresses this gap. We trace the evolution from traditional to neural approaches, highlight advances in tasks such as event dating, temporal modeling, and knowledge updating, and outline emerging challenges. In Figure 2, we portray a taxonomy overviewing temporal tasks, datasets, and approaches we will discuss in our review.

[1]For a discussion of previous related surveys, we refer the reader to Section A.1.

2 Key Concepts

We first introduce the core concepts related to Temporal IR/QA.

Temporal Information Retrieval (TIR) aims to retrieve documents that are not only topically relevant but also aligned with the query's **temporal intent**. Temporal intent may be explicit such as *"Olympics 2024"*, or implicit ones, such as *"latest Apple earnings"*. TIR relies on different **temporal signals** such as **document timestamps** (publication dates), **temporal expressions** (*"March 2023"*), and **event mentions** (*"2024 Olympics"*) to assess a document's **temporal relevance** indicating how well its temporal scope matches the query (Kanhabua and Nørvåg, 2008; Singh et al., 2016).

Temporal Question Answering (TQA) focuses on answering questions with **temporal constraints**, either explicitly stated, such as *"Who won the Nobel Prize in Physics in 2020?"* or implied, for instance, *"What are the latest US climate policies?"*. Success in TQA requires understanding the **question's temporal intent** and retrieving documents relevant to the corresponding time frame or ones published

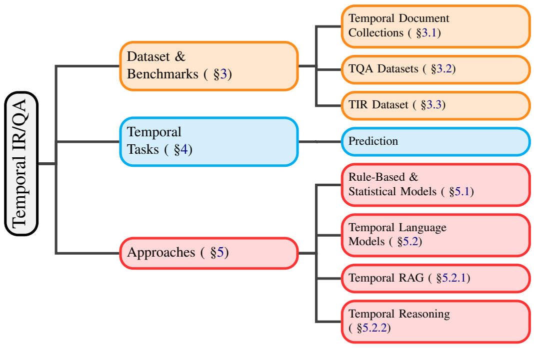

Figure 2: Taxonomy of temporal datasets and benchmarks, tasks, and approaches. For the complete version, please refer to Figure 3 in the Appendix.

around that time.

Temporal IR/QA rely on diverse temporal elements. *Temporal signals* are, in general, defined as features that convey time-related information in text. These include **explicit temporal expressions** like *"March 2023"* (used for indexing and filtering), implicit cues such as *"recently"* (requiring contextual interpretation), relative expressions like *"last week"* (necessitating to be anchored to a reference point), event-based references such as *"2024 Olympics"* (linking to known event timelines), and temporal metadata such as **document timestamps** which indicate publication time and often serve as proxies for judging freshness of content.

The concept of **document focus time** (Jatowt et al., 2013) is crucial here. It denotes the specific time point or interval a document relates to. For example, a 2013's publication discussing the 2010 Academy Awards has a focus time of 2010 while having a timestamp of 2013. Accurate focus time estimation of documents, using techniques like burst detection, temporal expression analysis, or timestamping named entities, enhances answer precision, especially in news or historical corpora (Wang et al., 2020).

In Appendix B we discuss other related concepts, such as temporal taggers, temponyms, granularity, temporal reasoning, timeline extraction, disambiguation, and robustness.

3 Datasets and Evaluation Benchmarks

The development of Temporal IR/QA systems fundamentally depends on the availability of temporally grounded datasets and robust evaluation methodologies used for training, testing, and benchmarking time-aware models. We provide an overview in this section of temporal datasets, organized into three categories: temporal document collections, TQA datasets, and TIR datasets.

3.1 Temporal Document Collection

Prior work has utilized diachronic and synchronic document collections as well as annotated temporal corpora.

Diachronic corpora consist of time-stamped documents spanning extensive time periods. They support retrospective retrieval, diachronic analysis, and event-based reasoning. Prominent examples include the *New York Times Annotated Corpus* (1987–2007; 1.8m articles) (Sandhaus, 2008), which, for example, serves as the basis for `ArchivalQA` (Wang et al., 2022) dataset, and the *CNN/Daily Mail corpus* (2007–2015; 313k articles) (Hermann et al., 2015) used, e.g., in `NewsQA` (Trischler et al., 2017). The *Chronicling America* collection (1800–1920) offers digitized historical newspaper articles and supports long-range historical QA via `ChroniclingAmericaQA` (Piryani et al., 2024b). More recently, the *Newswire* corpus (Silcock et al., 2024) has expanded the length of time frames, providing 2.7 million newswire articles published between 1878 and 1977. It is enriched with metadata including geo-referenced datelines, Wikipedia/Wikidata entity links, and topical annotations, enabling fine-grained historical and spatio-temporal modeling. Another widely used corpus is *CUSTOMNEWS* (Lazaridou et al., 2021) (1969–2019), which consists of crawled English news sources and spans diverse domains including politics, finance, and sports.

Diachronic corpora are also used in a range of related temporal tasks, including semantic drift detection (Hamilton et al., 2016), event burst modeling (Radinsky and Horvitz, 2013), and timeline construction (Gutehrlé et al., 2022).

Synchronous corpora represent a coherent snapshot of the world at a specific point in time. Unlike diachronic corpora, which typically span decades or years, synchronous collections capture a temporally aligned view, sometimes in conjunction with structured knowledge bases. Wikipedia articles (Vrandečić and Krötzsch, 2014), for example, reflect a particular version of world knowledge at a certain time (when the dump was made) and can be linked to Wikidata timestamps. Datasets like `TimeQA` (Chen et al., 2021), `TEMPREASON` (Tan et al., 2023), and `ComplexTempQA` (Gruber et al., 2024) build on Wikipedia snapshots to support temporally-scoped QA grounded in a time-specific

Dataset	#Questions	Knowledge Source	Creation Method	Answer Type	Time Frame	Temporal Metadata	Multi-Hop
NewsQA (Trischler et al., 2017)	119k	News	CS	Freeform	2007-2015	✗	✗
TDDiscourse (Naik et al., 2019)	6.1k	News	CS	Extractive	Unspecified	✗	✗
TORQUE (Ning et al., 2020)	21k	News	CS	Abstractive	-	✗	✗
ArchivalQA (Wang et al., 2022)	532k	News	AG	Extractive	1987-2007	✓	✗
TimeQA (Chen et al., 2021)	41.2K	Wikipedia	AG	Extractive	1367-2018	✗	✗
TiQ (Jia et al., 2024)	10K	Wikipedia	AG	Freebase	Unspecified	✗	✗
TempQuestions (Jia et al., 2018)	1.2k	Freebase	AG	Extractive	Unspecified	✗	✓
TemporalQuestions (Wang et al., 2021a)	1K	News	CS	Extrcative	1987-2007	✓	✗
TempLAMA (Dhingra et al., 2022)	50k	News	CS	Extractive	2010-2020	✓	✗
ComplexTempQA (Gruber et al., 2024)	100,228k	Wikipedia	AG	Extractive	1987-2023	✓	✓
MenatQA (Wei et al., 2023)	2.8k	Wikipedia	AG	Extractive	1367-2018	✗	✗
PAT-Question (Meem et al., 2024)	6,1k	Wikipedia	CS	Extractive	-	✗	✓
TempTabQA (Gupta et al., 2023)	11.4k	Wikipedia Info box	CS	Abstractive	-	✗	✗
SituatedQA (Zhang and Choi, 2021)	12.2k	Wikipedia	CS	–	≤ 2021	✗	✗
UnSeenTimeQA (Uddin et al., 2024)	3.6k	Synthetic	AG	Abstractive	-	✗	✓
ChroniclingAmericaQA (Piryani et al., 2024b)	485k	News	AG	Extractive	1800-1920	✓	✗
FRESHQA (Vu et al., 2024)	600	Google Search	CS	-	-	✗	✓
COTEMPQA (Su et al., 2024)	4.7k	Wikidata	CS	Abstractive	≤ 2023	✗	✓
Test of Time (ToT) (Fatemi et al., 2024)	1.8k	Synthetic	AG	Abstractive	-	✗	✓
TIMEDAIL (Qin et al., 2021)	1.1k	DailyDialog	CS	Multiple-choice	-	✗	✗
Complex-TR (Tan et al., 2024)	10.8	Wikipedia+Google Search	AG	Multi-answer	≤ 2023	✗	✓
StreamingQA (Liska et al., 2022)	147k	News	CS	Extractive	2007-2020	✓	✓
TRACIE (Zhou et al., 2021)	5.4k	Wikipedia	CS	abstractive	≤ 2020	✗	✗
ForecastQA (Jin et al., 2021)	10.3k	News	CS	Multiple-Choice	2015-2019	✓	✓
TEMPREASON (Tan et al., 2023)	52.8k	Wikipedia/Wikidata	SC	Abstractive	634-2023	✗	✗
TemporalAlignmentQA (Zhao et al., 2024)	20k	Wikipedia	AG	Abstractive	2000-2023	✗	✗
ReaLTimeQA (Kasai et al., 2023)	5.1k	Search	CS	Multiple-choice	2020-2024	✗	✗

Table 1: Overview of Temporal QA datasets. Each dataset is characterized by the number of questions, the underlying knowledge source, the question creation method (CS = Crowdsourced, AG = Automatically Generated), the answer type, and the timeframe covered by the knowledge source. A "≤" symbol indicates that the dataset uses a snapshot of Wikipedia and inherits its temporal scope. We also indicate whether temporal metadata is available and whether questions require multi-hop temporal reasoning.

context.

Finally, **Annotated temporal corpora** with explicit temporal annotations facilitate more structured forms of temporal reasoning. *TimeBank* (Pustejovsky et al., 2003) introduced TimeML to annotate temporal expressions, events, and their temporal relations. Follow-up datasets like `WikiWars` (Mazur and Dale, 2010) and `RED` (O'Gorman et al., 2016) extended it to historical narratives and causal relations, respectively. Such corpora constitute gold-standard resources for temporal tagging and relation extraction.

3.2 TQA Datasets

TQA datasets allow evaluating how well systems can answer questions that require temporal reasoning. They vary along multiple dimensions, including Knowledge Source, Temporal Orientation, Temporal Explicitness, and Reasoning Complexity.

Knowledge Source TQA datasets are commonly derived from diachronic or synchronic corpora. Diachronic Corpora (also known as **Primary Sources**) tend to provide contemporaneous accounts written around the time when events occurred in the past. Datasets such as `NewsQA` (Trischler et al., 2017), `TDDiscourse` (Naik et al., 2019), `TORQUE` (Ning et al., 2020), `ArchivalQA` (Wang et al., 2022), `TKGQA` (Ong et al., 2023), `ChroniclingAmericaQA` (Piryani et al., 2024b), are curated from old news sources and can be used to evaluate models' abilities to retrieve and reason over temporally anchored document collections. Table 1 lists all the datasets and the types of knowledge sources used to generate their questions.

In contrast, Synchronic Corpora like Wikipedia basically constitute **Secondary Sources** since they provide retrospective view of the past. They have been used to build datasets like `TimeQA` (Chen et al., 2021), `TEMPREASON` (Tan et al., 2023), `TiQ` (Jia et al., 2024), and `ComplexTempQA` (Gruber et al., 2024), which support fine-grained reasoning across temporally scoped, consistent knowledge bases.

Recent advancements have also seen the emergence of purely synthetic datasets designed to specifically test models on controlled and complex temporal reasoning scenarios. For example, `UnSeenTimeQA` (Uddin et al., 2024) introduces a novel, data contamination-free benchmark that evaluates temporal reasoning independently from any pre-training knowledge.

Temporal Orientation While most datasets focus on past events, future-oriented QA datasets remain relatively rare. Still, they are increasingly important for evaluating models' ability to perform predictive and hypothetical reasoning. `ForecastQA` (Jin et al., 2021) and `TimeBench` (Chu et al., 2024) are among the few benchmarks that include questions about future events, testing models' ability to perform timeline projections and forecast-based inference.

Question Type Temporal questions can be broadly classified by their explicitness in referencing time. Datasets like `TimeQA` (Chen et al., 2021), `SituatedQA` (Zhang and Choi, 2021) and `TempQuestions` (Jia et al., 2018) contain **Explicit Temporal Questions** with clear temporal markers, such as *"What happened in 1947?"*, signaling temporal intent directly.

In contrast, **Implicit Temporal Questions** omit direct time references but still require temporal inference. For instance, *"Who was Prime Minister of the UK when the Berlin Wall fell?"* requires inferring the date of the event and then linking it to a temporally relevant fact. Datasets such as `TiQ` (Jia et al., 2024) and `TORQUE` (Ning et al., 2020) focus on implicit reasoning, testing event-event and event-time relationships. Others like `ArchivalQA` (Wang et al., 2022), `TemporalQuestions` (Wang et al., 2021a), and `ComplexTempQA` (Gruber et al., 2024) combine both question types, offering a spectrum of temporal reasoning demands from explicit, time-anchored queries to implicit, event-based inference.

Temporal Reasoning Complexity TQA tasks also vary in the depth of reasoning they require. **Simple Temporal Questions** typically involve direct lookups, such as identifying the date of a specific event or the state of the world at a given time. Early datasets like `NewsQA` (Trischler et al., 2017) and `TempLAMA` (Dhingra et al., 2022) largely belong to this category. In contrast, **Complex Temporal Questions** demand more intricate processing such as multi-hop reasoning, temporal filtering, or synthesizing information across events. For example, the question *"What major international agreements were signed after World War I but before World War II?"* necessitates multi-hop temporal reasoning and contextual comparison. Datasets like `MenatQA` (Wei et al., 2023), `TempReason` (Tan et al., 2023), `Complex-TR` (Tan et al., 2024), and `ComplexTempQA` (Gruber et al., 2024) are explicitly designed to evaluate these advanced reasoning capabilities. Others like `TimeBench` (Chu et al., 2024) span both simple and complex reasoning levels, including tasks such as timeline construction or event duration inference. Table 1 compares various datasets for Temporal QA/IR.

3.3 TIR Datasets

While TQA datasets focus on answering time-sensitive questions, TIR datasets support tasks such as identifying time-sensitive documents, modeling temporal query intent, and ranking documents by temporal relevance or diversity. They typically pair queries with timestamped corpora and are designed to assess retrieval systems' performance across temporal dimensions.

The `Temporalia` series at NTCIR-11 and NTCIR-12 (Joho et al., 2014, 2016) established foundational benchmarks for TIR through two tasks: *Temporal Query Intent Classification (TQIC)*, which categorizes queries by temporal orientation (e.g., past, recency, future, atemporal), and *TIR*, which ranks documents based on their temporal relevance or diversity. The tasks use the *LivingKnowledge News/Blog Corpus* (Matthews et al., 2010), containing 3.8 million timestamped documents (2011—2013) annotated with time expressions and named entities. Apart from `Temporalia`, `TREC Temporal Summarization Track` (Diaz et al., 2015) offered datasets for a related task of real-time event summarization, testing systems' ability to rank documents by recency and relevance as well as emphasizing temporal diversity and freshness. In parallel, the `TempEval` series from the SemEval workshops (UzZaman et al., 2013; Verhagen et al., 2010, 2007) provided benchmark datasets for temporal information extraction such as temporal expression, event, and temporal relation, crucial for supporting TIR tasks.

4 Temporal Prediction Tasks

Temporal prediction tasks are essential for developing time-aware IR and QA systems. They focus on inferring implicit or missing temporal information from text, thereby improving the alignment between queries, documents, and events. These tasks are critical when explicit temporal metadata is sparse, noisy, or unavailable, and they support applications such as historical search, timeline construction, and temporally sensitive retrieval.

Key tasks include **Event Dating, Document Dating**, **Focus Time Estimation**, **Query Time Profiling**, and **Event Occurrence Prediction**. Traditional methods rely on statistical language models and handcrafted rules, while more recent techniques employ transformer-based encoders, temporal embeddings, and graph-based reasoning to improve generalization and robustness (Yang et al., 2023; Abdallah et al., 2025; Liu and Quan, 2025; Yang et al., 2024). For a detailed review of task definitions, representative techniques, and evaluation strategies, we refer the readers to Appendix C.

5 Approaches in Temporal IR/QA

A wide range of approaches have been developed to address the challenges of Temporal IR/QA, from early rule-based systems and statistical models to neural networks and large language models (LLMs). They differ in how they represent temporal information, reason over temporal relationships, and adapt to changing world knowledge.

5.1 Rule-based & Statistical Methods

Early work in Temporal IR/QA was dominated by rule-based systems and statistical models that laid the groundwork for core temporal tasks such as time expression normalization, event ordering, and temporal ranking. While limited in scalability and adaptability, they introduced many foundational concepts that remain relevant today.

In TIR, rule-based systems focused on extracting and normalizing time expressions to improve retrieval for time-sensitive queries (Arikan et al., 2009; Alonso et al., 2007). Models like TCluster (Alonso et al., 2009) and time-based language models (Li and Croft, 2003) used document timestamps and decay functions to model recency, while others like Berberich et al. (2010) combined metadata and vague expressions in probabilistic ranking models. To handle implicit temporal intent, techniques such as median timestamp analysis (Kanhabua and Nørvåg, 2010) and query log mining (Metzler et al., 2009) were introduced.

Other strategies focused on enhancing recency-aware retrieval. Jatowt et al. (2005) proposed reranking methods using archived web snapshots to favor fresher content, while Dong et al. (2010) incorporated real-time Twitter signals, and Setty et al. (2017) used news signals into crawling and ranking to support time-sensitive queries. Efficient indexing methods were also developed to support temporal queries over evolving corpora such as Wikipedia and web archives (Anand et al., 2011, 2012; Holzmann and Anand, 2016). Styskin et al. (2011) introduced a machine learning model to predict recency sensitivity, combining it with greedy diversification to balance freshness and topical relevance.

As TIR matured, researchers began modeling the temporal dynamics of both queries and documents. Kulkarni et al. (2011) analyzed how user intents evolve over time, highlighting the need for adaptive retrieval strategies that can respond to temporal drift in query behavior. Joho et al. (2013) studied the prevalence of different temporal orientations of user queries, and the strategies user apply to find temporally relevant content from the past, future or present. Later systems adapted ranking strategies to temporal query profiles using machine learning (Kanhabua et al., 2012) or temporal interval representations (Rizzo et al., 2022).

Early QA systems like Harabagiu and Bejan (2005) relied on TimeML and lexical resources like WordNet (Miller, 1992) for event reasoning. To handle complex temporal questions more effectively, Saquete et al. (2004, 2009) introduced a multi-layered QA architecture that decomposed questions into temporally constrained subquestions using temporal expression taggers like TERSEO (Saquete et al., 2003). These approaches showed improved precision and generalizability across languages.

Despite their simplicity, rule-based and statistical methods introduced key mechanisms of temporal intent modeling, expression normalization, and timeline reasoning that continue to influence more advanced systems.

5.2 Temporal Language Models

The emergence of deep learning has significantly advanced Temporal IR/QA by enabling models to capture temporal dependencies and contextual nuances. Recent research has led to the development of **Temporal Language Models (TLMs)** that explicitly incorporate temporal signals during pretraining or fine-tuning. Models such as TempoT5 (Dhingra et al., 2022), TempoBERT (Rosin et al., 2022), and BiTimeBERT (Wang et al., 2023) included timestamps and temporal expressions directly into their training inputs or used time-focused pretraining tasks, improving temporal generalization in downstream tasks such as semantic change detection and Temporal QA. Other approaches, like syntax-guided temporal language model (SG-TLM) (Su et al., 2023), enhance sensitivity to temporal structure by masking syntactic and semantic spans that carry temporal meaning.

On the other hand, Cao and Wang (2022) explored time-aware generation by introducing temporal prompts, including both natural language timestamp descriptions and continuous vector (linear) representations of timestamps. Beyond input-level integration, time-aware language models like TALM (Ren et al., 2023) incorporate time-specific word representations through hierarchical modeling and temporal adaptation, achieving strong results in historical text dating. TCQA (Son and Oh, 2023) employs synthetic data and a time-context span selection task to train models that align time-aware representations with contextually grounded answers. Further, techniques such as Temporal Span Masking (TSM) (Cole et al., 2023) and temporal attention mechanisms (Rosin and Radinsky, 2022) incorporate explicit temporal annotations into transformer architectures to improve time sensitivity.

5.2.1 Temporal RAG

While TLMs improve temporal understanding through pretraining, they remain limited by the static nature of training data. To address evolving information needs and reduce temporal hallucinations, recent work has turned to **Retrieval-Augmented Generation (RAG)** that integrates neural retrieval with generation to incorporate up-to-date, time-relevant evidence at inference time.

Recent temporal RAG systems extend this idea by embedding temporal signals directly into retrieval and generation pipelines. TempRALM (Gade and Jetcheva, 2024) introduces temporal signals into dense retrieval, enhancing recency and factual grounding for time-sensitive queries. TempRetriever (Abdallah et al., 2025) and TsContriever (Wu et al., 2024) encode temporal relevance directly into dense retrievers, improving alignment between temporal queries and evidence. TimeR4 (Qian et al., 2024) proposes a Retrieve-Rewrite-Retrieve-Rerank pipeline that transforms implicit temporal queries into explicit ones, retrieves from time-anchored knowledge sources, and reranks based on temporal constraints. MRAG (Siyue et al., 2024) adapts RAG with multi-source and multi-hop temporal retrieval for event-centric QA. To mitigate hallucinations and outdated generations, FRESHPROMPT (Vu et al., 2024) integrates real-time signals into the prompting and retrieval process. Together, these models make RAG more responsive to temporal dynamics in IR/QA.

5.2.2 Temporal Reasoning Capabilities

While Temporal Language Models enhance time-aware representations and retrieval, many Temporal IR/QA tasks demand more sophisticated reasoning, such as understanding event sequences, temporal constraints, and durations.

Temporal reasoning capabilities in pre-trained language models (PLMs) have seen notable improvements, with recent efforts focusing on enhancing zero-shot generalization and temporal robustness. Continual temporal adaptation methods, including ECONET (Han et al., 2021), enhance temporal relational coherence and consistency across evolving contexts. Structural temporal reasoning models like TIMERS (Mathur et al., 2021), and ConTempo (Niu et al., 2024) address multi-hop and document-level inference with specialized architectures. Moreover, event duration and ordering prediction have benefited from task-specific temporal pretraining objectives (e.g., E-PRED, R-PRED) (Yang et al., 2020) and transfer learning strategies (Virgo et al., 2022).

Despite these advancements, modeling temporal relationships in LLMs remains challenging. Recent benchmarks such as TRAM (Wang and Zhao, 2024) evaluate LLMs on tasks like event ordering, arithmetic, frequency, and duration, revealing that even strong models like GPT-4 fall short of human-level performance. To isolate genuine reasoning abilities from memorization, Test of Time (ToT) (Fatemi et al., 2024) introduces synthetic tasks targeting temporal logic and inference. Additionally, TODAY (Feng et al., 2023) challenges models with subtle temporal shifts and differential analysis. Methods like Narrative-of-Thought (Zhang et al., 2024) guide models to generate structured temporal narratives. Finally, Wallat et al. (2024, 2025) study temporal blind spots of LLMs and their resiliency to changes in time-related elements (e.g., altering a date in a query, or its position) elucidating missing knowledge and showing that current models are still vulnerable to adversarial or other perturbations.

6 Future Directions

Despite two decades of research and significant progress, Temporal IR/QA systems still struggle to adapt to evolving real-world events, shifting user needs, and dynamic data streams. To advance the development of time-aware systems, we propose future directions organized into three core themes: *System Design* (architectures and real-time capabilities), *Knowledge Management* (updating and representing time-sensitive knowledge), and *Evaluation and Robustness* (metrics and generalization).

These directions address gaps identified throughout this survey, such as temporal bias, rigidity in models, and the limited scope of existing evaluations.

6.1 System Design

Real-Time Information Integration. Most IR systems depend on periodically updated corpora, leaving them blind to rapidly unfolding events like elections, protests, or trending information. Future work should treat data as a continuous stream, enabling real-time indexing (Baeza-Yates and Ribeiro-Neto, 2011), burst detection (Wang et al., 2021a), and responsive re-ranking (Tran et al., 2015), as well as supporting applications like live event tracking or misinformation detection (V et al., 2024).

Development of Temporally-Aware LLM Agents. Current LLM agents prioritize task completion or dialogue but lack structured temporal understanding (Wallat et al., 2024). Future systems should include dedicated temporal understanding methods for better understanding temporal references, semantics, and test-time reasoning.

6.2 Knowledge Management

Advanced Temporal Knowledge Editing. Static models struggle to keep up with real-world change. Instead of retraining, future systems could use modular, trackable edit layers for local updates, preserving historical facts.

Integration of Diachronic and Synchronic Knowledge. Temporal questions often require combining evolving facts (e.g., event timelines) with stable knowledge (e.g., definitions). Future systems should integrate diachronic sources with synchronous sources to provide comprehensive answers. For example, answering "How has the unemployment rate changed since 2008?" requires diachronic trends from datasets like **ArchivalQA** (Section 3) and synchronous explanations from Wikipedia, addressing the aggregation needs (Section 2).

Multilingual Temporal IR/QA. Temporal expressions vary across languages and cultures, posing challenges for globalized systems. For instance, date formats differ (e.g., DD/MM/YYYY vs. MM/DD/YYYY), and cultural references (e.g., "post-Meiji era" in Japanese) require context-specific interpretation. Future research should develop cross-lingual temporal taggers, multilingual benchmarks, and culturally adaptive models, building on multilingual taggers like HeidelTime (Strötgen and Gertz, 2010).

6.3 Evaluation and Robustness

Implicit Temporal Intent Understanding. Many queries imply but do not state a time frame. Future work should improve models' ability to infer latent temporal scopes using derived labels or event grounding. This addresses the implicit reasoning challenges in datasets like `TORQUE` (Ning et al., 2020) and `TiQ` (Jia et al., 2024).

Robustness to Temporal Drift and Misalignment. Performance drops when models are applied to data from different time periods, which can reduce accuracy (Shin et al., 2025; Zhang and Choi, 2023; Luu et al., 2022; Wallat et al., 2025). Future work should enhance model resilience to temporal misalignment, building on the robustness challenges in `Test of Time`.

7 Conclusion

Temporal IR/QA is critical for retrieving and reasoning over time-sensitive information in dynamic, evolving contexts. In this survey, we have traced the field's progression from early rule-based systems to TLMs and RAG approaches. We identified core challenges, including temporal tagging, temporal intent detection, event ordering, and robustness to evolving facts and implicit temporal signals.

Our review highlights persistent limitations such as reliance on static knowledge, limited capabilities for future-oriented reasoning, and dataset bias toward past events. We show that temporal complexity, vague expressions, knowledge drift, and real-time demands significantly impact system behavior and evaluation.

Despite notable progress, current systems often struggle with temporal uncertainty, maintaining consistency across time, and adapting to multilingual or culturally diverse temporal expressions. As real-world applications increasingly require temporally adaptive systems, these gaps point to the need for richer evaluation protocols, improved temporal representations, and continual learning strategies. We anticipate future progress toward robust, time-aware IR and QA systems capable of understanding not just what happened, but also when, why, and how information evolves over time.

Limitations

This survey aims to provide a comprehensive overview of Temporal IR/QA. There are a few important limitations to acknowledge.

We made our best efforts to be thorough, but it is possible that some relevant works may have been missed. We conducted an extensive literature review using forward and backward snowballing techniques, with particular attention to papers published in major venues such as ACL, SIGIR, EMNLP, NeurIPS, ECIR, and preprints on arXiv. On the other hand, due to page limitations, we provide only a very brief summary of each method without exhaustive technical details.

References

Abdelrahman Abdallah, Bhawna Piryani, Jonas Wallat, Avishek Anand, and Adam Jatowt. 2025. Tempretriever: Fusion-based temporal dense passage retrieval for time-sensitive questions. *arXiv preprint arXiv:2502.21024*.

Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2007. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2):35–41.

Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2009. Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 97–106, New York, NY, USA. Association for Computing Machinery.

Omar Alonso, Jannik Strötgen, Ricardo Baeza-Yates, and Michael Gertz. 2011. Temporal information retrieval: Challenges and opportunities. In *Temporal Web Analytics Workshop TWAW 2011*, page 1.

Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. 2011. Temporal index sharding for space-time efficiency in archive search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 545–554, New York, NY, USA. Association for Computing Machinery.

Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. 2012. Index maintenance for time-travel text search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 235–244.

Irem Arikan, Srikanta Bedathur, and Klaus Berberich. 2009. Time will tell: Leveraging temporal expressions in ir. In *Second ACM International Conference on Web Search and Data Mining*. ACM.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval: The concepts and technology behind search*, 2nd edition. Addison-Wesley Publishing Company, USA.

Anab Maulana Barik, Wynne Hsu, and Mong-Li Lee. 2024. Time matters: An end-to-end solution for temporal claim verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 657–664, Miami, Florida, US. Association for Computational Linguistics.

Harsimran Bedi, Sangameshwar Patil, Swapnil Hingmire, and Girish Palshikar. 2017. Event timeline generation from history textbooks. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 69–77, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Klaus Berberich, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. 2010. A language modeling approach for temporal information needs. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval*, ECIR'2010, page 13–25, Berlin, Heidelberg. Springer-Verlag.

Ricardo Campos, Gaël Dias, Alípio M. Jorge, and Adam Jatowt. 2014. Survey of temporal information retrieval and related applications. *ACM Computing Survey*, 47(2).

Shuyang Cao and Lu Wang. 2022. Time-aware prompting for text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7231–7246, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angel X. Chang and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul, Turkey. European Language Resources Association (ELRA).

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11378–11392, Toronto, Canada. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Wisam Dakka, Luis Gravano, and Panagiotis G. Ipeirotis. 2008. Answering general time sensitive queries. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 1437–1438, New York, NY, USA. Association for Computing Machinery.

Angelo Dalli. 2006. Temporal classification of text and automatic document dating. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 29–32, New York City, USA. Association for Computational Linguistics.

Supratim Das, Arunav Mishra, Klaus Berberich, and Vinay Setty. 2017. Estimating event focus time using neural word embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2039–2042, New York, NY, USA. Association for Computing Machinery.

Franciska de Jong, Henning Rode, and Djoerd Hiemstra. 2005. Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, pages 161–168. Koninklijke Nederlandse Academie van Wetenschappen.

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.

Fernando Diaz, Matthew Ekstrand-Abueg, Richard McCreadie, Virgil Pavlu, and Tetsuya Sakai. 2015. Trec 2014 temporal summarization track overview.

Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. 2010. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 331–340, New York, NY, USA. Association for Computing Machinery.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv preprint arXiv:2406.09170*.

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. Association for Computational Linguistics.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Anoushka Gade and Jorjeta G Jetcheva. 2024. It's about time: Incorporating temporality in retrieval augmented language models. *CoRR*.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2024. Complextempqa: A large-scale dataset for complex temporal question answering. *arXiv preprint arXiv:2406.04866*.

Dhruv Gupta and Klaus Berberich. 2014. Identifying time intervals of interest to queries. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2431–2453, Singapore. Association for Computational Linguistics.

Nicolas Gutehrlé, Antoine Doucet, and Adam Jatowt. 2022. Archive TimeLine summarization (ATLS): Conceptual framework for timeline generation over historical document collections. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 13–23, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021. ECONET: Effective continual pretraining of language models for event temporal reasoning. In *Pro-*

ceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5367–5380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sanda Harabagiu and Cosmin Adrian Bejan. 2005. Question answering based on temporal inference. In *Proceedings of the AAAI-2005 workshop on inference for textual question answering*, pages 27–34.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Helge Holzmann and Avishek Anand. 2016. Tempas: Temporal archive search based on tags. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 207–210.

Or Honovich, Lucas Torroba Hennigen, Omri Abend, and Shay B. Cohen. 2020. Machine reading of historical events. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7486–7497, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Adam Jatowt, Ching-Man Au Yeung, and Katsumi Tanaka. 2013. Estimating document focus time. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2273–2278, New York, NY, USA. Association for Computing Machinery.

Adam Jatowt, Ching Man Au Yeung, and Katsumi Tanaka. 2015. Generic method for detecting focus time of documents. *Information Processing & Management*, 51(6):851–868.

Adam Jatowt, Yukiko Kawai, and Katsumi Tanaka. 2005. Temporal ranking of search engine results. In *Web Information Systems Engineering–WISE 2005: 6th International Conference on Web Information Systems Engineering, New York, NY, USA, November 20-22, 2005. Proceedings 6*, pages 43–52. Springer.

Adam Jatowt, Yukiko Kawai, and Katsumi Tanaka. 2007. Detecting age of page content. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management*, WIDM '07, page 137–144, New York, NY, USA. Association for Computing Machinery.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1057–1062, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zhen Jia, Philipp Christmann, and Gerhard Weikum. 2024. Tiq: A benchmark for temporal question answering with implicit time constraints. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1394–1399, New York, NY, USA. Association for Computing Machinery.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 792–802, New York, NY, USA. Association for Computing Machinery.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. ForecastQA: A question answering challenge for event forecasting with temporal text data. 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 4636–4650, Online. Association for Computational Linguistics.

Hideo Joho, Adam Jatowt, and Roi Blanco. 2014. Ntcir temporalia: a test collection for temporal information access research. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, page 845–850, New York, NY, USA. Association for Computing Machinery.

Hideo Joho, Adam Jatowt, Roi Blanco, Haitao Yu, and Shuhei Yamamoto. 2016. Building test collections for evaluating temporal ir. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 677–680, New York, NY, USA. Association for Computing Machinery.

Hideo Joho, Adam Jatowt, and Blanco Roi. 2013. A survey of temporal web search experience. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 1101–1108, New York, NY, USA. Association for Computing Machinery.

Rosie Jones and Fernando Diaz. 2007. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3):14–es.

Nattiya Kanhabua and Avishek Anand. 2016. Temporal information retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1235–1238, New York, NY, USA. Association for Computing Machinery.

Nattiya Kanhabua, Klaus Berberich, and Kjetil Nørvåg. 2012. Learning to select a time-aware retrieval model. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 1099–1100, New York, NY, USA. Association for Computing Machinery.

Nattiya Kanhabua, Roi Blanco, and Kjetil Nørvåg. 2015. Temporal information retrieval. *Found. Trends Inf. Retr.*, 9(2):91–208.

Nattiya Kanhabua and Kjetil Nørvåg. 2008. Improving temporal language models for determining time of non-timestamped documents. In *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '08, page 358–370, Berlin, Heidelberg. Springer-Verlag.

Nattiya Kanhabua and Kjetil Nørvåg. 2010. Determining time of queries for re-ranking search results. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'10, page 261–272, Berlin, Heidelberg. Springer-Verlag.

Nattiya Kanhabua and Kjetil Nørvåg. 2011. A comparison of time-aware ranking methods. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 1257–1258, New York, NY, USA. Association for Computing Machinery.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime qa: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pages 49025–49043. Curran Associates, Inc.

Mei Kobayashi and Koichi Takeda. 2000. Information retrieval on the web. *ACM Comput. Surv.*, 32(2):144–173.

Anagha Kulkarni, Jaime Teevan, Krysta M. Svore, and Susan T. Dumais. 2011. Understanding temporal query dynamics. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, page 167–176, New York, NY, USA. Association for Computing Machinery.

Abhimanu Kumar, Jason Baldridge, Matthew Lease, and Joydeep Ghosh. 2012. Dating texts without explicit temporal cues. *arXiv preprint arXiv:1211.2290*.

Erdal Kuzey, Vinay Setty, Jannik Strötgen, and Gerhard Weikum. 2016a. As time goes by: Comprehensive tagging of textual phrases with temporal scopes. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 915–925, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Erdal Kuzey, Jannik Strötgen, Vinay Setty, and Gerhard Weikum. 2016b. Temponym tagging: Temporal scopes for textual phrases. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 841–842, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363. Curran Associates, Inc.

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.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. *ACM Trans. Inf. Syst.* Just Accepted.

Xiaoyan Li and W. Bruce Croft. 2003. Time-based language models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, page 469–475, New York, NY, USA. Association for Computing Machinery.

Shih-Ting Lin, Nathanael Chambers, and Greg Durrett. 2021. Conditional generation of temporally-ordered event sequences. 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 7142–7157, Online. Association for Computational Linguistics.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13604–13622. PMLR.

Zefang Liu and Yinzhu Quan. 2025. Retrieval of temporal event sequences from textual descriptions. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 37–49, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. TIMERS: Document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533, Online. Association for Computational Linguistics.

Michael Matthews, Pancho Tolchinsky, Roi Blanco, Jordi Atserias, Peter Mika, and Hugo Zaragoza. 2010. Searching through time in the new york times. *HCIR 2010*, page 41.

Vaibhav Mavi, Anubhav Jangra, Adam Jatowt, et al. 2024. Multi-hop question answering. *Foundations and Trends® in Information Retrieval*, 17(5):457–586.

Pawel Mazur and Robert Dale. 2010. WikiWars: A new corpus for research on temporal expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 913–922, Cambridge, MA. Association for Computational Linguistics.

Jannat Meem, Muhammad Rashid, Yue Dong, and Vagelis Hristidis. 2024. PAT-questions: A self-updating benchmark for present-anchored temporal question-answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13129–13148, Bangkok, Thailand. Association for Computational Linguistics.

Donald Metzler, Rosie Jones, Fuchun Peng, and Ruiqiang Zhang. 2009. Improving search relevance for implicitly temporal queries. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 700–701, New York, NY, USA. Association for Computing Machinery.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Christian Morbidoni, Alessandro Cucchiarelli, and Domenico Ursino. 2018. Leveraging linked entities to estimate focus time of short texts. In *Proceedings of the 22nd International Database Engineering & Applications Symposium*, IDEAS '18, page 282–286, New York, NY, USA. Association for Computing Machinery.

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, Stockholm, Sweden. Association for Computational Linguistics.

Vlad Niculae, Marcos Zampieri, Liviu Dinu, and Alina Maria Ciobanu. 2014. Temporal text ranking and automatic dating of texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 17–21, Gothenburg, Sweden. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.

Jingcheng Niu, Saifei Liao, Victoria Ng, Simon De Montigny, and Gerald Penn. 2024. ConTempo: A unified temporally contrastive framework for temporal relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1521–1533, Bangkok, Thailand. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Ryan Ong, Jiahao Sun, Ovidiu Șerban, and Yi-Ke Guo. 2023. Tkgqa dataset: Using question answering to guide and validate the evolution of temporal knowledge graph. *Data*, 8(3).

Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, and Adam Jatowt. 2024a. Detecting temporal ambiguity in questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9620–9634, Miami, Florida, USA. Association for Computational Linguistics.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024b. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2038–2048, New York, NY, USA. Association for Computing Machinery.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. 2024. TimeR^4 : Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6942–6952, Miami, Florida, USA. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 255–264, New York, NY, USA. Association for Computing Machinery.

Han Ren, Hai Wang, Yajie Zhao, and Yafeng Ren. 2023. Time-aware language modeling for historical text dating. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13646–13656, Singapore. Association for Computational Linguistics.

Stefano Giovanni Rizzo, Matteo Brucato, and Danilo Montesi. 2022. Ranking models for the temporal dimension of text. *ACM Trans. Inf. Syst.*, 41(2).

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 833–841, New York, NY, USA. Association for Computing Machinery.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.

Hany M. SalahEldeen and Michael L. Nelson. 2013. Carbon dating the web: estimating the age of web resources. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 1075–1082, New York, NY, USA. Association for Computing Machinery.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

E. Saquete, P. Martínez-Barco, R. Muñoz, and J. L. Vicedo. 2004. Splitting complex temporal questions for question answering systems. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 566–es, USA. Association for Computational Linguistics.

Estela Saquete, Rafael Munoz, and Patricio Martínez-Barco. 2003. Terseo: Temporal expression resolution system applied to event ordering. In *International Conference on Text, Speech and Dialogue*, pages 220–228. Springer.

Estela Saquete, Jose L. Vicedo, Patricio Martínez-Barco, Rafael Muñoz, and Hector Llorens. 2009. Enhancing qa systems with complex temporal question processing capabilities. *Journal of Artificial Intelligence Research*, 35(1):755–811.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. 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 6663–6676, Online. Association for Computational Linguistics.

Vinay Setty, Abhijit Anand, Arunav Mishra, and Avishek Anand. 2017. Modeling event importance for ranking daily news events. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 231–240.

Changho Shin, Xinya Yan, Suenggwan Jo, Sungjun Cho, Shourjo Aditya Chaudhuri, and Frederic Sala. 2025. Tardis: Mitigating temporal misalignment via representation steering. *arXiv e-prints*, pages arXiv–2503.

Shashank Shrivastava, Mitesh Khapra, and Sutanu Chakraborti. 2017. A concept driven graph based approach for estimating the focus time of a document. In *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings*, page 250–260, Berlin, Heidelberg. Springer-Verlag.

Emily Silcock, Abhishek Arora, Luca D'Amico-Wong, and Melissa Dell. 2024. Newswire: A large-scale structured database of a century of historical news. In *Advances in Neural Information Processing Systems*, volume 37, pages 49768–49779. Curran Associates, Inc.

Jaspreet Singh, Wolfgang Nejdl, and Avishek Anand. 2016. History by diversity: Helping historians search news archives. In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval*, pages 183–192.

Zhang Siyue, Xue Yuxiang, Zhang Yiming, Wu Xiaobao, Luu Anh Tuan, and Zhao Chen. 2024. Mrag: A modular retrieval framework for time-sensitive question answering. *arXiv preprint arXiv:2412.15540*.

Daivik Sojitra, Raghav Jain, Sriparna Saha, Adam Jatowt, and Manish Gupta. 2024. Timeline summarization in the era of llms. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2657–2661, New York, NY, USA. Association for Computing Machinery.

Jungbin Son and Alice Oh. 2023. Time-aware representation learning for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 70–77, Singapore. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.

Andrey Styskin, Fedor Romanenko, Fedor Vorobyev, and Pavel Serdyukov. 2011. Recency ranking by diversification of result set. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 1949–1952, New York, NY, USA. Association for Computing Machinery.

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. Association for Computational Linguistics.

Zhaochen Su, Juntao Li, Zikang Zhang, Zihan Zhou, and Min Zhang. 2023. Efficient continue training of temporal language model with structural information. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6315–6329, Singapore. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Tuan A Tran, Claudia Niederée, Nattiya Kanhabua, Ujwal Gadiraju, and Avishek Anand. 2015. Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1201–1210.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Md Nayem Uddin, Amir Saeidi, Divij Handa, Agastya Seth, Tran Cao Son, Eduardo Blanco, Steven R Corman, and Chitta Baral. 2024. Unseentimeqa: Time-sensitive question-answering beyond llms' memorization. *arXiv preprint arXiv:2407.03525*.

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. Association for Computational Linguistics.

Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 650–660, New York, NY, USA. Association for Computing Machinery.

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. Dating documents using graph convolution networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1605–1615, Melbourne, Australia. 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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Felix Virgo, Fei Cheng, and Sadao Kurohashi. 2022. Improving event duration question answering by leveraging existing temporal information extraction data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4451–4457, Marseille, France. European Language Resources Association.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Jonas Wallat, Abdelrahman Abdallah, Adam Jatowt, and Avishek Anand. 2025. A study into investigating temporal robustness of llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal blind spots in large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 683–692, New York, NY, USA. Association for Computing Machinery.

Jiexin Wang, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Answering event-related questions over long-term news article archives. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*, pages 774–789. Springer.

Jiexin Wang, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2021a. Improving question answering for event-focused questions in temporal collections of news articles. *Inf. Retr.*, 24(1):29–54.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2021b. Event occurrence date estimation based on multivariate time series analysis over temporal document collections. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 398–407, New York, NY, USA. Association for Computing Machinery.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3025–3035, New York, NY, USA. Association for Computing Machinery.

Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa, and Yi Cai. 2023. Bitimebert: Extending pre-trained language representations with bi-temporal information. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 812–821, New York, NY, USA. Association for Computing Machinery.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, Bangkok, Thailand. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Feifan Wu, Lingyuan Liu, Wentao He, Ziqi Liu, Zhiqiang Zhang, Haofen Wang, and Meng Wang. 2024. Time-sensitve retrieval-augmented generation for question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2544–2553, New York, NY, USA. Association for Computing Machinery.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Siheng Xiong, Yuan Yang, Ali Payani, James C Kerce, and Faramarz Fekri. 2024. Teilp: Time prediction over knowledge graphs via logical reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):16112–16119.

Sen Yang, Xin Li, Lidong Bing, and Wai Lam. 2023. Once upon a *time* in *graph*: Relative-time pretraining for complex temporal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11879–11895, Singapore. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Zonglin Yang, Xinya Du, Alexander Rush, and Claire Cardie. 2020. Improving event duration prediction via time-aware pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3370–3378, Online. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14213–14226, Singapore. Association for Computational Linguistics.

Xinliang Frederick Zhang, Nicholas Beauchamp, and Lu Wang. 2024. 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. Association for Computational Linguistics.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah Smith. 2024. Set the clock: Temporal alignment of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15015–15040, Bangkok, Thailand. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A Appendix

A.1 Related Surveys

Advances in temporal datasets, time-aware models, and reasoning techniques have enabled systems capable of retrieving time-relevant documents, ordering events, and answering temporally constrained questions, benefiting applications such as historical analysis, fact-checking, and intelligent assistants.

While IR and QA have been widely surveyed, most existing reviews focus on general techniques, often neglecting temporal aspects. IR surveys emphasize ranking functions, neural retrieval models, and query understanding (Li et al., 2025; Zhu et al., 2023), while QA surveys center on extractive, abstractive, or multi-hop answering over static knowledge sources (Zhu et al., 2021; Mavi et al., 2024). They rarely consider temporal intent, dynamic or evolving information, or event sequencing, highlighting a key gap that remains unaddressed.

Several earlier works provided foundational insights into Temporal IR. Alonso et al. (2011) discusses challenges such as real-time streams, exploratory temporal search, and spatio-temporal retrieval. Campos et al. (2014) offers a broad overview of document dating, time-aware ranking, and query understanding, covering both explicit and implicit time signals. Kanhabua and Anand (2016) complements these with a tutorial on temporal indexing and ranking, emphasizing the detection of temporal query intent. There has been however no recent systematic overview of the field despite much research interest.

As a parallel research line, TQA over knowledge graphs has gained considerable attention (Jia et al., 2021; Saxena et al., 2021; Chen et al., 2023; Xiong et al., 2024). Our survey focuses on temporally aware IR/QA over text. We review both traditional and neural approaches to core tasks such as temporal tagging, event dating, time-aware retrieval, and temporal reasoning. To our knowledge, no prior survey brings together recent developments across these tasks in the context of text-based IR/QA. Other related topics, including temporal fact verification (Barik et al., 2024) and timeline summarization (Sojitra et al., 2024), are discussed only when directly relevant.

B Temporal Processing Concepts

We mention here other concepts broadly related to temporal processing.

Temporal taggers are essential tools in temporal information processing; they identify and standardize time expressions in text, such as *"March 15, 2021"* or *"yesterday,"* converting them into formats like YYYY-MM-DD and categorizing them (e.g., DATE, DURATION). Popular taggers like HeidelTime (Strötgen and Gertz, 2010), SUTime (Chang and Manning, 2012), Temponym tagger (Kuzey et al., 2016b), CogCompTime (Ning et al., 2018) support a range of languages and domains, forming the foundation for downstream tasks including

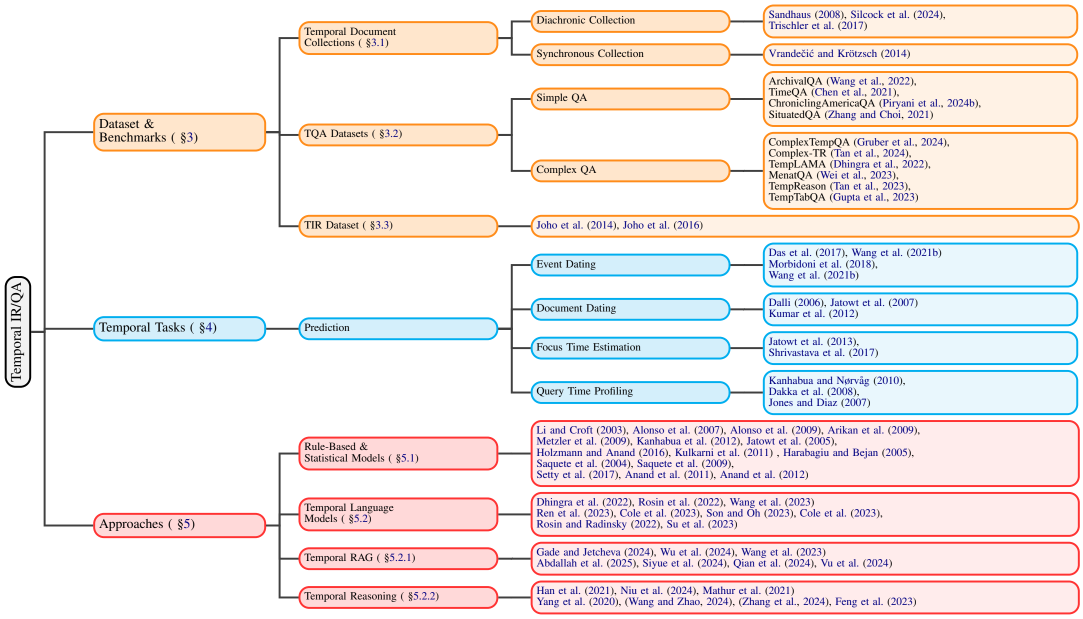

Figure 3: Taxonomy of temporal datasets and benchmarks, tasks, and approaches (Complete version of Figure 2).

TQA, event ordering, and timeline construction.

Additionally, **Temponyms** (Kuzey et al., 2016a) are free-text phrases that implicitly refer to specific time periods or events but are not recognized as standard temporal expressions, for Instance, *"Greek referendum"* or *"Clinton's presidency"*. Recognizing and resolving these expressions is essential for comprehensive temporal understanding. Other related concepts include **temporal granularity** (typically ranging from day to decade), **temporal proximity** (the temporal closeness of a document to the query's target time, influencing ranking), and **temporal distribution** patterns in retrieval results. Effectively leveraging these signals is key to building time-aware systems (Campos et al., 2014).

Temporal Disambiguation resolves ambiguous time references (e.g., identifying which "Tuesday" is being discussed), addressing **temporal ambiguity** in both queries and documents (Piryani et al., 2024a). **Temporal Co-reference** involves identifying and linking different mentions of the same temporal entity within or across documents, such as connecting "that year" to "2020" (Ning et al., 2018). **Timeline Extraction** automatically constructs a chronological sequence of events or facts from text, to answer questions requiring event ordering, such as constructing a historical timeline (Bedi et al., 2017).

More advanced reasoning tasks include **Temporal Reasoning**, which infers time-related relationships, such as determining the order of events or calculating durations between them. It is crucial for answering complex questions like *"What happened in Poland after World War II and before 1960?"* (Leeuwenberg and Moens, 2019). **Temporal Aggregation** synthesizes information from multiple time periods to answer broad or comparative questions (e.g., *"How has climate policy evolved over the last decade?"*). **Temporal Robustness** (Wallat et al., 2025) refers to the resiliency of systems to adversarial changes in time-related elements (e.g., altering a date in a query, or its position in a sentence) in the form of **temporal perturbations**. It is used in evaluation to assess temporal reasoning stability.

C Temporal Prediction Tasks

Temporal prediction tasks are crucial in understanding and organizing time-sensitive textual data. Despite sharing the common objective of grounding text in time, these tasks differ in focus, granularity, and application. In this section, we explore related temporal prediction tasks—document dating, document focus time estimation, temporal query profiling, and event occurrence time estimation, which provide complementary insights and support distinct applications. Each task addresses unique aspects of temporal analysis, from inferring document creation times to profiling query intent. Below, we review these tasks, their methodologies, and key contributions, emphasizing their roles in temporal IR and QA.

C.1 Document Dating

Document dating refers to the task of estimating a document's creation time (e.g., publication date) based on its textual content, especially when metadata is missing, unreliable, or unavailable. The input is the full document text, and the output is a timestamp, typically at year or month granularity.

Early approaches, such as that by de Jong et al. (2005), leveraged unigram language models trained over distinct time periods to determine when a document's vocabulary was most prevalent. Building on this, Kanhabua and Nørvåg (2008) integrated additional linguistic features such as part-of-speech tags, tf-idf scores, and collocations to better capture temporal patterns. Dalli (2006) introduced an unsupervised method for automatic document dating using periodic word usage. Kumar et al. (2012) trained language models over discretized time intervals (chronons) using Wikipedia biographies. Niculae et al. (2014) model document dating as a pairwise ranking problem using logistic regression. More recently, Vashishth et al. (2018) introduced a neural method employing Graph Convolutional Networks (GCNs) to model syntactic and temporal relations jointly.

Document dating is crucial in temporal indexing, digital preservation, and metadata recovery, particularly for historical or noisy corpora. Beyond textual content analysis, several methods estimate the creation date of web resources. Jatowt et al. (2007) was the first approach for dating content of web pages. The authors estimated timestamps of individual content elements of web pages using their archived snapshots. SalahEldeen and Nelson (2013) developed Carbon Date, a tool that aggregates signals from multiple online sources, such as first tweets, archive snapshots, URL shorteners, and search engine crawls, to estimate a webpage's creation date.

C.2 Document Focus Time Estimation

Document focus time estimation aims to identify the historical time periods that a document discusses, which may differ from its actual publication date. For example, a news article published in 2021 that analyzes the 9/11 attacks would have a focus time centered around September 2001. The input to this task is the document's full text, and the output consists of one or more temporal intervals that represent the document's narrative temporal scope. Jatowt et al. (2013) proposed a graph-based method that models co-occurrences between terms and dates to identify salient temporal associations within the text. Building on this, Jatowt et al. (2015) introduced a method that estimates focus time using statistical evidence from external corpora, even when explicit temporal expressions are limited. Shrivastava et al. (2017) further advanced this line of work by linking documents to Wikipedia concepts, leveraging their temporal relations to estimate focus times. This task supports historical analysis, event-centric retrieval, and timeline generation, providing insights into the temporal context of textual content.

C.3 Temporal Query Profiling

Temporal query profiling determines a query's temporal intent and time of interest, such as whether it refers to the past, future, or is atemporal. The input is a short keyword query (e.g., "Ukraine-Russia war"), and the output is an inferred time or temporal distribution. Kanhabua and Nørvåg (2010) estimated query time by analyzing timestamps of top-k retrieved documents, while Dakka et al. (2008) and Jones and Diaz (2007) modeled temporal distributions of relevant documents. Kanhabua and Nørvåg (2011) conducted a comparative evaluation of five temporal ranking approaches (LMT, LMTU, TS, TSU, FuzzySet), evaluating their ability to model uncertainty and adapt to temporal variance. Gupta and Berberich (2014) combined timestamp metadata with temporal expressions in document content to infer precise time intervals. Temporal query profiling is essential for time-aware IR, as it enables query disambiguation, improves temporal relevance ranking, and supports applications such as event-centric search and timeline construction.

C.4 Event Occurrence Time Estimation

Event occurrence time estimation aims to predict the specific date on which an event occurred, given a short textual description (e.g., "Plane crash in Armenia kills 36"). Unlike document-centric tasks, this focuses on the event mention itself and typically requires high-granularity outputs—such as day- or month-level timestamps.

Das et al. (2017) introduced time vectors combining word and global temporal embeddings, estimating dates via cosine similarity. Morbidoni et al. (2018) leveraged structured knowledge bases such as DBpedia and Wikipedia to link event descriptions to temporally grounded entities. Honovich et al. (2020) proposed a neural approach with sentence extraction, LSTM with attention, and an MLP classifier for date prediction. More recently, Wang et al. (2021b) introduced TEP-Trans, a Transformer-based model that formulates event time prediction as a multivariate time series forecasting problem using features extracted from temporal news collections.

Summary: While these temporal prediction tasks are highly interrelated, each aiming to anchor textual information within a temporal context, they address distinct facets of temporal understanding. Document dating predicts when a document was created, whereas document focus time estimation identifies when the content is about, which may precede or differ from the creation time. Temporal query profiling focuses on the user's intent, inferring when the query is directed in time rather than analyzing any specific document. Finally, event occurrence time estimation deals with precise, often fine-grained dating of event mentions, requiring models to infer real-world event timelines from sparse input. Together, these tasks form a complementary suite of temporal reasoning capabilities, enabling robust time-aware information retrieval and question answering systems.