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

Bhawna Piryani[*]♠**, Abdelrahman Abdallah**♠**, Jamshid Mozafari**♠**,**
Avishek Anand♡**, Adam Jatowt**♠
♠University of Innsbruck, ♡Delft University of Technology
{bhawna.piryani, abdelrahman.abdallah, jamshid.mozafari, adam.jatowt}@uibk.ac.at
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 Question Answering* (TQA), a research area that focuses on answering questions involving temporal constraints or context. As time-stamped content from sources like news articles, web archives, and knowledge bases continues to grow, TQA systems must address challenges such as detecting temporal intent, normalizing time expressions, ordering events, and reasoning over evolving or ambiguous facts. We organize existing work through a unified perspective that captures the interaction between corpus temporality, question temporality, and model capabilities, enabling a systematic comparison of datasets, tasks, and approaches. We review recent advances in TQA enabled by neural architectures, especially transformer-based models and Large Language Models (LLMs), highlighting progress in temporal language modeling, retrieval-augmented generation (RAG), and temporal reasoning. We also discuss benchmark datasets and evaluation strategies designed to test temporal robustness, recency awareness, and generalization.

https://github.com/DataScienceUIBK/TemporalQA-Survey

1 Introduction

Time fundamentally shapes how information is written, retrieved, and interpreted. As digital content continues to expand across time-stamped sources such as news archives, social media, and knowledge bases, the ability to reason about when events occur and how information evolves has become essential (Alonso et al., 2007). Temporal Question Answering (TQA) focuses on questions whose interpretation or answers depend on temporal context. Unlike standard Question Answering (QA), TQA requires systems to detect temporal intent, ground temporal references, infer event order, and reason over dynamic or ambiguous facts, rather than relying solely on surface-level retrieval.

These requirements introduce challenges that distinguish TQA from conventional QA tasks. One key challenge is temporal ambiguity resolution, where vague expressions such as *"recently"* or *"after the war"* must be interpreted relative to context. Another is cross-temporal reasoning, which involves understanding causal and sequential relationships across events. In addition, knowledge volatility refers to the evolution of facts over time, which renders static corpora and pre-trained models inadequate for answering time-sensitive queries. Temporal intent may be expressed explicitly or implicitly, requiring systems to infer appropriate timeframes. Addressing these challenges requires temporally aware reasoning, context-sensitive retrieval, and mechanisms for adapting to evolving knowledge (Berberich et al., 2010).

These challenges are illustrated in Figure 1, which presents two example questions. *Q1: "At what age did Obama win the Nobel Peace Prize?"* requires identifying and grounding two temporal anchors-Obama's birth year (1961) and the year he received the Nobel Peace Prize (2009)-and computing the interval between them to derive 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?" requires a more nuanced understanding of context. The phrase *"late years of service"* must be anchored to Obama's presidency timeline (2009-2017). Accurate answering requires access to contemporaneous policy documents from that specific period, rather than relying on retrospective analyses. This highlights that temporal context determines which

[*]Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28852–28881

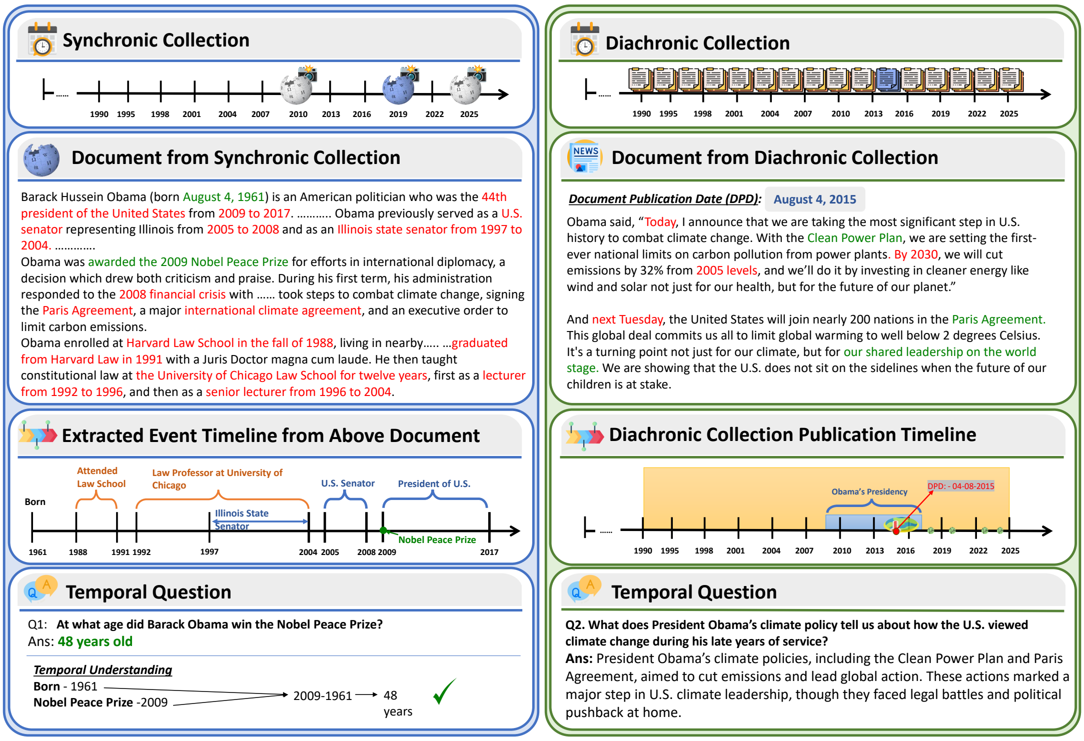

Figure 1: The figure illustrates an example of answering a temporal question using documents from a synchronic (left) and diachronic (right) collection. The Synchronic and Diachronic Collection panel highlights the specific document used to answer the question in each case in blue. Red marks temporal signals within the text, while Green highlights the answer-supporting spans. On the left, an event timeline is inferred from the internal structure of an undated document in the synchronic collection. On the right, the collection timeline displays the distribution of documents over time in the diachronic collection; Red dots indicate documents that contain an answer, while green dots denote documents related to the event described in Question Q2.

knowledge sources are relevant. As shown in the figure, the model must also interpret relative expressions such as *"today"* or *"next week"* within document content by anchoring them to the *document's publication date (August 4, 2015)*, thereby contextualizing Obama's climate policy statement within it's correct historical context.

These examples illustrate a broader point: effective TQA requires aligning three interacting dimensions: (i) the temporality of the underlying corpus (diachronic vs. synchronic), (ii) the temporal structure of the question (explicit vs. implicit intent, temporal orientation, and reasoning complexity), and (iii) the temporal capabilities of current models (temporal language modeling, temporally aware retrieval, and temporal reasoning). We adopt this three-dimensional perspective as the organizing framework of this survey, enabling us to synthesize prior work not just as datasets and systems, but as a landscape shaped by how different approaches handle mismatches across these dimensions.

Research in TQA has evolved substantially, moving from rule-based pipelines (Harabagiu and Bejan, 2005) and statistical models (Berberich et al., 2010; Wang et al., 2020) to neural systems capable of richer temporal reasoning capabilities (Dhingra et al., 2022; Wang et al., 2023). Early work focused on handcrafted normalization and event ordering rules, which struggled with scale and coverage. Pre-trained language models further shifted the field by supporting more robust temporal reasoning (Jain et al., 2023), event sequencing (Lin et al., 2021), and continual temporal adaptation (Han et al., 2021), enabling multi-hop temporal inference and retrieval-augmented updates.

Despite this progress, current systems still face challenges such as handling future-oriented questions, reasoning over temporally inconsistent documents, and mitigating knowledge decay. Addressing these problems requires a deeper understanding of the field's current capabilities and limitations.

While existing surveys[1] either focus on general QA/IR (Kolomiyets and Moens, 2011; Zhu et al., 2025, 2021) or on narrow aspects of temporal pro-

[1] See Appendix A for a detailed discussion.

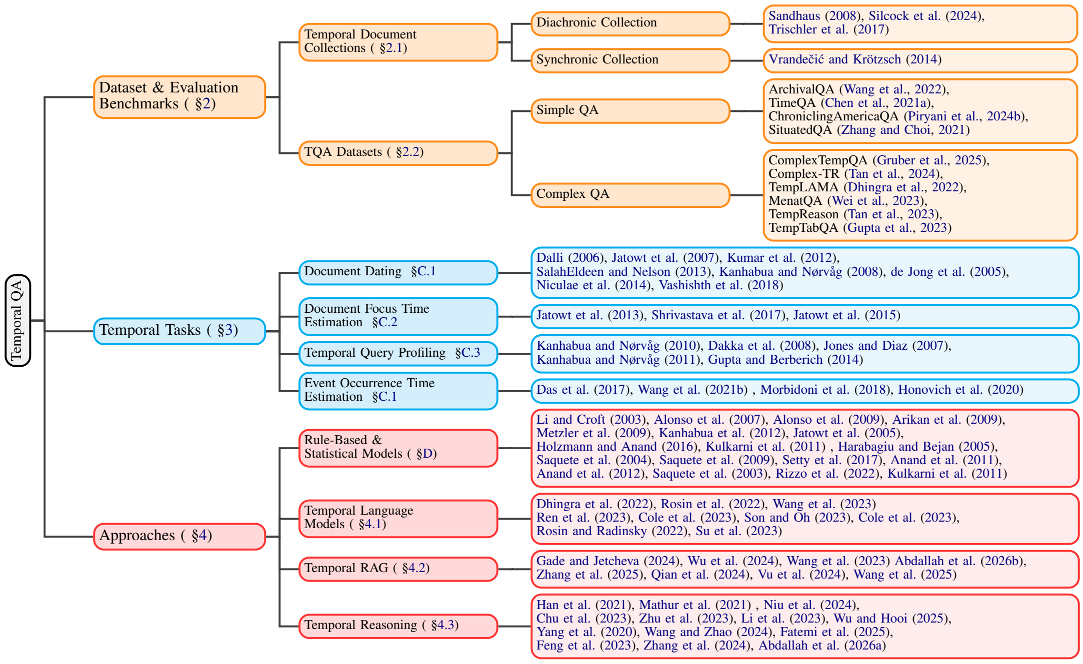

Figure 2: Taxonomy of Temporal Datasets and Evaluation benchmarks, Tasks, and Approaches.

cessing (Kobayashi and Takeda, 2000; Kanhabua et al., 2015; Campos et al., 2014), to the best of our knowledge, no prior work systematically examines TQA as a full pipeline spanning corpus, question, and model temporality. The last survey reviewing TQA (Campos et al., 2014) predates modern temporal LMs, RAG systems, and today's large temporal benchmarks, leaving a gap that we aim to fill.

In this survey, we provide a comprehensive synthesis of recent work in TQA over unstructured text. Our scope is limited to systems that operate on natural language questions and unstructured documents. We therefore exclude structured settings such as temporal KG-based QA (Su et al., 2024a; Saxena et al., 2021) and temporal QA over semi-structured tables (Shankarampeta et al., 2025). These settings assume explicit temporal structure (e.g., timestamped rows or evolving attribute records) and involve reasoning challenges that differ from grounding temporal expressions in free text. We also do not provide full coverage of temporal-reasoning benchmarks that are not grounded in textual evidence; for an overview of such resources, we refer readers to Wenzel and Jatowt (2023). This focus allows us to examine TQA as an end-to-end pipeline grounded in real-world text corpora, and to analyze how temporal reasoning interacts with retrieval, evidence selection, and evolving content.

We systematically review the evolution of TQA methods, benchmark datasets, evaluation strategies, and emerging challenges. To guide this discussion, we present a taxonomy, as illustrated in Figure 2, that organizes prior work into datasets and benchmarks, temporal tasks, and modeling approaches. Throughout the survey, we use the three-dimensional framework introduced above—corpus temporality, question temporality, and model temporality—as an analytical lens to compare methods, reveal mismatches, and identify open challenges.

2 Datasets and Evaluation Benchmarks

For readers unfamiliar with basic temporal concepts and terminology, such as **temporal expressions**, **temporal anchoring**, or **temporal signals**, we provide a concise overview in Appendix B. We now examine the datasets and evaluation benchmarks used to evaluate systems for temporally grounded question answering. From the three-dimensional perspective introduced earlier, existing benchmarks can be characterized along two primary dimensions: **corpus temporality** (diachronic vs. synchronic) and **question temporality** (explicit vs. implicit), which determine the temporal reasoning capabilities being evaluated. This section provides an overview of temporal datasets, organized into two categories: (i) Temporal Document Collections, which serve as knowledge sources, and (ii) TQA datasets, which enable systematic evaluation of temporal reasoning capabilities.

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., 2021a)	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	Extractive	1987-2007	✓	✗
TempLAMA (Dhingra et al., 2022)	50k	News	CS	Extractive	2010-2020	✓	✗
ComplexTempQA (Gruber et al., 2025)	100.2M	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	-	$\leq$ 2021	✗	✗
UnSeenTimeQA (Uddin et al., 2025)	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., 2024b)	4.7k	Wikidata	CS	Abstractive	$\leq$ 2023	✗	✓
Test of Time (ToT) (Fatemi et al., 2025)	1.8k	Synthetic	AG	Abstractive	-	✗	✓
TIMEDAIL (Qin et al., 2021)	1.1k	DailyDialog	CS	Multiple-choice	-	✗	✗
TEMPO (Abdallah et al., 2026a)	1.7k	Stack Exchange	CS	Abstractive	$\leq$ 2025	✓	✓
Complex-TR (Tan et al., 2024)	10.8k	Wikipedia+Google Search	AG	Multi-answer	$\leq$ 2023	✗	✓
StreamingQA (Liska et al., 2022)	147k	News	CS	Extractive	2007-2020	✓	✓
TRACIE (Zhou et al., 2021)	5.4k	Wikipedia	CS	Abstractive	$\leq$ 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 "$\leq$" 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.

2.1 Temporal Document Collections

Prior work has utilized three types of temporally structured document collections for temporal reasoning tasks, each offering distinct advantages.

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 serves as the basis for the `ArchivalQA` dataset (Wang et al., 2022). Another widely used resource is the *CNN/Daily Mail corpus* (2007–2015; 313K articles) (Hermann et al., 2015), used in datasets such as `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) extends temporal coverage, 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 spanning diverse domains such as politics, finance, and sports. Diachronic corpora are also used in 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).

Synchronic corpora represent a coherent snapshot of the world at a specific point in time. Unlike diachronic corpora, which typically span decades or years, synchronic collections capture a temporally aligned view, sometimes in conjunction with structured KBs. Wikipedia articles (Vrandečić and Krötzsch, 2014), for example, reflect a particular version of world knowledge at a specific time (when the dump was made) and can be linked to Wikidata timestamps. Datasets such as `TimeQA` (Chen et al., 2021a), `TEMPREASON` (Tan et al., 2023), and `ComplexTempQA` (Gruber et al., 2025) build on Wikipedia snapshots to support temporally-scoped QA grounded in a time-specific context.

Finally, **Annotated temporal corpora** with explicit temporal annotations enable more structured 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 this framework to historical narratives and causal relations, respectively. These corpora serve as gold-standard resources for temporal tagging, relation extraction, and for developing models capable of explicit temporal reasoning over text. Unlike diachronic and synchronic corpora, these resources emphasize explicit temporal structure, making them particularly suitable for modeling fine-grained temporal relations.

2.2 TQA Datasets

TQA datasets enable the evaluation of how well systems can answer questions that require temporal understanding and reasoning. They vary along multiple dimensions, including **Knowledge Source**, **Temporal Orientation**, **Temporal Explicitness**, and **Reasoning Complexity**. Table 1 provides a comprehensive comparison of major TQA datasets across these dimensions.

Knowledge Source: TQA datasets are commonly derived from diachronic or synchronic corpora. **Primary source datasets** derive from diachronic corpora, providing contemporaneous accounts written when events occurred. These datasets test models' ability to retrieve and reason over temporally anchored document collections with authentic historical perspectives. 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 historical news sources and support temporal reasoning over period-specific content.

In contrast, synchronic corpora such as Wikipedia constitute **Secondary sources**, as they provide a retrospective view of the past. They have been used to build datasets such as `TimeQA` (Chen et al., 2021a), `TEMPREASON` (Tan et al., 2023), `TiQ` (Jia et al., 2024), and `ComplexTempQA` (Gruber et al., 2025), which support fine-grained reasoning across temporally scoped, consistent knowledge bases.

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), `FutureContext` (Mutschlechner and Jatowt, 2025), 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., 2021a), `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 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. Other datasets, such as `ArchivalQA` (Wang et al., 2022), and `ComplexTempQA` (Gruber et al., 2025) 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., 2025) 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.

Comparative perspective across TQA datasets: TQA benchmarks exhibit systematic trade-offs depending on their construction and source. *Answer drift varies by dataset type*: crowd-sourced datasets (e.g., `SituatedQA`, `FRESHQA`) mitigate ambiguity through explicit temporal clarification, whereas automatically generated datasets (e.g., `ArchivalQA`) better capture factual change over time, including entity evolution.

Source type further shapes evaluation: news-based, primary-source datasets emphasize temporal robustness under evolving evidence and event-centric reasoning, whereas Wikipedia-based datasets offer broader topical coverage with more stable, retrospective temporal framing. Finally, *construction methods reflect scale–quality trade-offs*: crowdsourced datasets better capture natural language variation and implicit temporal intent, while automatically generated benchmarks enable large-scale evaluation of temporal sensitivity. A dataset selection guide summarizing these complementary strengths is provided in Table 4 in the Appendix.

Summary: TQA poses challenges that go beyond annotation quality and dataset scale. Unlike static QA tasks, the answers to temporal questions can shift over time, necessitating benchmarks that can account for answer drift and temporal volatility. Recent work such as RecencyQA (Piryani et al., 2026) addresses this limitation by introducing a recency–stationarity taxonomy, enabling fine-grained evaluation of how answer validity evolves over time and across contexts. However, most existing datasets remain static snapshots: notable exceptions include `RealTimeQA` (Kasai et al., 2023) and `FreshQA` (Vu et al., 2024), which incorporate periodic updates, though these updates are often resource-intensive, While datasets like `PATQA` (Meem et al., 2024) explore scalable, automated update mechanisms. *A persistent issue is temporal ambiguity* (Piryani et al., 2024a), where missing or implicit time references hinder both annotation and evaluation. Structural biases also emerge: *diachronic corpora tend to have more data surrounding major events and contain detailed information on past events*, yet they reflect contemporaneous perspectives. In contrast, *synchronic corpora provide broader coverage but often lack fine-grained event granularity.* Annotation strategies further involve trade-offs: *crowdsourced datasets are typically small but high-quality, whereas automatically generated datasets are larger but noisier*. Moreover, most benchmarks fail to isolate specific temporal reasoning skills, such as duration inference or event ordering, thereby limiting diagnostic evaluation.

While this section focuses on general-domain datasets, specialized domains such as medical, legal, and financial contexts introduce additional temporal reasoning challenges and require tailored dataset designs, as discussed in Appendix E.

3 Temporal Prediction Tasks

Temporal prediction tasks play a central role in the development of time-aware IR and QA systems. They focus on inferring implicit or missing temporal information from text, thereby improving alignment between queries, documents, and events. These tasks are especially important when explicit temporal metadata is sparse, noisy, or unavailable, and they support applications such as historical search, timeline construction, and time-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., 2026b; Liu and Quan, 2025; Yang et al., 2024). For a detailed review of task definitions, representative techniques, and evaluation strategies, we refer readers to Appendix C.

4 Approaches in Temporal QA

TQA approaches differ in how they model temporal information in questions, retrieve time-relevant evidence, and reason over evolving knowledge. A wide range of methods has been developed to address these challenges, ranging from early rule-based and statistical systems to neural architectures and large language models (LLMs). Tables 2 and 3 (see Appendix) provide a comprehensive comparison of temporal QA approaches across architectural paradigms, temporal representations, and methodological strategies. They highlight differences in temporal modeling, reasoning strategies, and adaptation to changing world knowledge. These approaches differ in how they address the interaction between corpus, question, and model temporality introduced earlier.

Many TQA systems additionally rely on auxiliary temporal prediction tasks (Section 3) to infer missing or implicit time information, particularly when explicit metadata is sparse or unreliable.

In this section, we focus on neural approaches, particularly temporal language models (TLMs), RAG, and reasoning for TQA. Detailed discussions of traditional rule-based and statistical QA methods are provided in Appendix D.

4.1 Temporal Language Models

Deep learning has significantly advanced TQA by enabling models to capture temporal dependencies and contextual dynamics. Recent research has led to the development of **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) integrate timestamps and temporal expressions into the training process, improving temporal generalization across downstream tasks, including semantic change detection and TQA.

Beyond explicit timestamp conditioning, a second line of work introduces architectural and objective-level modifications to better internalize temporal structure. Approaches such as TALM (Ren et al., 2023), SG-TLM (Su et al., 2023), Temporal Span Masking (Cole et al., 2023), and Temporal Attention (Rosin and Radinsky, 2022) encourage models to attend to time-related cues, durations, and event structure through specialized masking, hierarchical representations, or time-aware attention mechanisms.

Some TLMs are designed for generation tasks that require explicit temporal grounding. For instance, Cao and Wang (2022) proposes temporal prompts, both textual and continuous vector-based, that guide generation with time-specific context. TCQA (Son and Oh, 2023) introduces a synthetic QA dataset and a span-selection task that aligns answers with their temporal context, enabling models to simulate historical reasoning and maintain timeline consistency.

A common trend across TLMs is the shift from treating time as an auxiliary signal to making it a core component of the modeling process. Techniques such as timestamp conditioning, temporal pretraining tasks, and time-aware attention have improved temporal reasoning. However, most models still rely on static corpora and struggle with vague or implicit temporal cues. In practice, timestamp-conditioned or QA-supervised models align better with TQA than semantic change approaches focused on distributional drift. While these models represent a significant step forward, they still struggle with real-world temporal drift and complex reasoning.

4.2 Temporal RAG

While TLMs improve temporal understanding through pretraining, they remain constrained by the static nature of their training data. To address evolving information needs and mitigate temporal hallucinations, recent work has turned to **RAG** frameworks that couple neural retrieval with generation, enabling models to incorporate up-to-date, time-relevant evidence at inference time.

A key trend in this space is the integration of temporal signals directly into the retrieval process. Models such as TempRetriever (Abdallah et al., 2026b) and TsContriever (Wu et al., 2024) extend dense retrievers to account for temporal relevance by encoding queries and passages with timestamp-aware embeddings. These approaches improve alignment between temporally scoped questions and temporally valid evidence. Similarly, TempRALM (Gade and Jetcheva, 2024) enhances dense retrieval with temporal constraints to reduce factual drift and improve recency-grounded responses.

Beyond modifying retrieval scoring, several systems re-architect the retrieval–generation pipeline itself. TimeR4 (Qian et al., 2024) introduces a four-stage *Retrieve–Rewrite–Retrieve–Rerank* framework that transforms underspecified temporal queries into explicitly time-anchored formulations before retrieving and re-ranking documents based on temporal fit. MRAG (Zhang et al., 2025) further extends this direction by performing multi-hop reasoning across events, retrieving from multiple time-scoped sources, and aggregating cross-temporal evidence for complex QA. More recently, TimeRAG (Wang et al., 2025) proposes an iterative temporal–semantic query decomposition and time-aware answer generation, enabling multi-step retrieval and reasoning over rapidly evolving information.

Compared to timestamp-conditioned language models, temporal RAG systems are most effective when knowledge is volatile, temporal intent is implicit, or answers depend on recent or evolving facts. In temporal RAG pipelines, retrieval quality is often the primary bottleneck for TQA. Unlike standard QA, temporal questions require evidence that is both topically relevant and temporally aligned with the query. Retrieval can fail due to noisy or missing timestamps, recency bias, or mismatches between document creation time and the event time described in the text. This propagates errors downstream and yields temporally inconsistent answers. Despite recent advances, robust temporal retrieval remains an open challenge, especially for queries with implicit temporal intent or noisy metadata.

Together, these models signal a shift from static temporal representations to adaptive, retrieval-grounded temporal reasoning. While temporal RAG systems mitigate hallucinations and better reflect evolving knowledge, their effectiveness ultimately depends on reliable and temporally consistent evidence.

4.3 Temporal Reasoning Capabilities

While TLMs enhance time-aware representation, they remain largely limited to surface-level temporal associations. Many TQA tasks instead require **explicit temporal reasoning**-tracking event sequences, resolving durations, performing temporal arithmetic, and drawing logical inferences across time. These capabilities are essential for questions involving temporal transitions, causality, or evolving narratives, where retrieval alone cannot ensure temporally coherent answers.

Recent approaches have therefore sought to strengthen the reasoning capacity of pretrained language models (PLMs) through architectural innovations and specialized objectives. ECONET (Han et al., 2021) pioneers continual temporal adaptation, preserving event consistency as models are updated with new information. TIMERS (Mathur et al., 2021) and ConTempo (Niu et al., 2024) move beyond token-level reasoning, introducing structured inference layers that explicitly model temporal graphs, event hierarchies, and symbolic relations between actions. Despite these advances, such methods remain brittle in practice, improving local temporal ordering while still failing at multi-hop or abstract reasoning that spans long time.

Recent work has explored structured and logic-enhanced reasoning for text-based TQA, including temporal graph modeling over documents (Chu et al., 2023), executable program induction for time-sensitive question answering (Zhu et al., 2023), and logic- or timeline-based reasoning frameworks that explicitly formalize temporal inference for LLMs (Li et al., 2023; Wu and Hooi, 2025).

To expose these weaknesses more systematically, new benchmarks now evaluate LLMs along multiple reasoning dimensions. TRAM (Wang and Zhao, 2024) assesses event frequency, duration estimation, and timeline ordering, consistently revealing large performance gaps between humans and even the strongest models such as GPT-4. Test of Time (ToT) (Fatemi et al., 2025) isolates reasoning from memorization through synthetic tasks involving logical deduction, date arithmetic, and counterfactual inference—showing that models' temporal accuracy often collapses once factual recall cues are removed. TEMPO (Abdallah et al., 2026a) focuses on reasoning-intensive temporal retrieval. Complementary benchmarks, such as TODAY (Feng et al., 2023) and Narrative-of-Thought (Zhang et al., 2024), further show that LLMs struggle to maintain temporal coherence over evolving contexts.

Beyond accuracy, **temporal robustness** has emerged as a critical yet underexplored weakness. Studies such as (Wallat et al., 2024, 2025) identify persistent *temporal blind spots*, where even small perturbations to timestamps or event order cause sharp drops in performance. These findings suggest that models rely on shallow lexical or positional cues rather than genuine temporal reasoning, lacking invariance to time-related perturbations and failing to generalize across temporal shifts.

Despite steady progress, current systems remain far from robust temporal reasoners. A core limitation is that most architectures entangle temporal reasoning with retrieval or pattern matching, rather than representing time as an abstract, manipulable variable. They exhibit limited understanding of implicit temporal cues, weak abstraction beyond explicit dates, and poor generalization under temporal drift. Bridging these gaps will require integrating symbolic and logical reasoning modules, incorporating explicit event memory, and designing evaluation frameworks that can disentangle genuine reasoning competence from pattern recall or retrieval bias.

Summary: TQA has evolved from rule-based pipelines to neural architectures that integrate temporal information across input representations, model internals, and retrieval mechanisms.

TLMs improved contextualization by embedding timestamps and temporal cues, yet they remain limited by static training corpora and an over-reliance on explicit time expressions, often failing on vague or underspecified temporal references. Temporal RAG systems *mitigate these issues by dynamically retrieving time-sensitive evidence at inference, but their performance depends critically on the temporal quality, relevance, and freshness of retrieved content. Building on these advances, recent reasoning-oriented architectures extend capabilities toward event ordering, temporal inference, and multi-hop reasoning, yet even state-of-the-art models exhibit temporal blind spots and struggle to generalize across shifting temporal contexts.*

Collectively, the paradigms alleviate specific weaknesses of their predecessors, but none fully resolves the core challenges of temporal ambiguity, factual drift, and reasoning consistency. Future research must systematically target these failure modes, balancing parametric and retrieved knowledge, handling implicit or underspecified timeframes, and developing evaluation methodologies that disentangle genuine temporal reasoning from statistical pattern matching.

5 Future Research Directions

Despite recent advances in neural temporal reasoning, current TQA systems face fundamental limitations that require targeted research. We identify the following critical areas where future work is needed to build temporally aware, trustworthy QA systems. Detailed technical challenges and potential solutions for each direction are in Appendix F.

Dynamic Temporal Knowledge Management. Current TQA systems rely on static corpora, making them unable to respond to fast-evolving information needs. The temporal propagation problem, where small updates disrupt temporal dependencies across related events, exposes fundamental architectural limitations. Future systems must move beyond isolated fact updates toward scalable frameworks that track, edit, and reason over temporal dependencies in real time.

Temporally aware LLM agents. While LLMs demonstrate impressive general reasoning capabilities, they consistently underperform in temporal reasoning tasks. LLMs exhibit temporal hallucinations and fail to resolve context-dependent expressions such as "last Tuesday" or "since our previous discussion." Incorporating timeline tracking, event memory, and temporal reference resolution is essential for temporally coherent dialogue agents.

Diachronic and Synchronic Knowledge Integration. Temporal questions often require combining time-sensitive sources that capture change over time with stable sources summarizing information at specific points. Most current systems treat these separately, limiting their ability to answer questions involving both historical trends and current facts. Future systems should develop temporal alignment algorithms and cross-source reasoning frameworks.

Temporal Uncertainty and Confidence Modeling. Many historical events have unclear start and end dates, yet most TQA systems treat all dates as exact. This mismatch creates problems when systems give confident answers about uncertain information, with issues compounding in multi-step reasoning. Future systems must explicitly model temporal uncertainty and provide confidence for temporal answers.

Multilingual and Multimodal Temporal QA. Temporal expressions, date formats, and cultural references vary widely across languages and modalities, yet most TQA systems are developed primarily for English textual input. Current systems struggle with diverse temporal signals due to limited cultural grounding and inadequate multimodal integration. Future research should focus on developing multilingual temporal tagging models and cross-modal alignment techniques.

Implicit Temporal Intent Understanding. Many temporal questions hide their time requirements rather than stating them directly. Current systems struggle with this ambiguity because they lack mechanisms to infer intended timeframes from contextual clues or shared knowledge. As users increasingly expect models to infer nuanced, context-sensitive answers, detecting hidden temporal assumptions is critical to avoid plausible but misleading outputs.

Evaluation and Benchmarking for Temporal Reasoning. Standard metrics like accuracy and F1 fall short capturing temporal coherence or reasoning depth. Developing temporally aware evaluation metrics remains an important challenge for effectively assessing and comparing future systems.

6 Conclusion

TQA plays a critical role in retrieving and reasoning over time-sensitive information in dynamic contexts. We focus on recent advancements in the field, particularly modern TLMs and RAG approaches, examining core challenges including temporal expression recognition, event ordering, and implicit temporal reasoning. Despite advances, current systems struggle with answer drift, temporal uncertainty, and limited temporal granularity, while most benchmarks rely on static snapshots that hinder real-time and future-oriented reasoning. Viewed through the interaction among corpus temporality, question temporality, and model capabilities, these limitations highlight the need for more robust temporal representations, dynamic evaluation protocols, and adaptive learning mechanisms.

Limitations

This survey aims to provide a comprehensive overview of TQA. 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, CIKM and preprints on arXiv. On the other hand, due to page limitations, we provide only a brief summary of each method, without providing exhaustive technical details.

Acknowledgments

The authors would like to acknowledge the financial support provided by the Austrian Research Agency (FFG) for the project "AI Enabled Sustainability Jurisdiction Demonstrator" (Project No. 915229).

References

Abdelrahman Abdallah, Mohammed Ali, Muhammad Abdul-Mageed, and Adam Jatowt. 2026a. Tempo: A realistic multi-domain benchmark for temporal reasoning-intensive retrieval. *arXiv preprint arXiv:2601.09523*.

Abdelrahman Abdallah, Bhawna Piryani, Jonas Wallat, Avishek Anand, and Adam Jatowt. 2026b. Tempretriever: Fusion-based temporal dense passage retrieval for time-sensitive questions. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, WSDM '26, page 5–15, New York, NY, USA. Association for Computing Machinery.

Liesbeth Allein, Marlon Saelens, Ruben Cartuyvels, and Marie-Francine Moens. 2023. Implicit temporal reasoning for evidence-based fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 176–189, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

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.

Adrián Bazaga, Rexhina Blloshmi, Bill Byrne, and Adrià de Gispert. 2025. Learning to reason over time: Timeline self-reflection for improved temporal reasoning in language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28014–28033, Vienna, Austria. 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.

Marius Büttner and Ivan Habernal. 2024. Answering legal questions from laymen in German civil law system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2027, St. Julian's, Malta. Association for Computational Linguistics.

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).

Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023a. Equals: A real-world dataset for legal question answering via reading chinese laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 71–80, New York, NY, USA. Association for Computing Machinery.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. FinTextQA: A dataset for long-form financial question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6025–6047, Bangkok, Thailand. Association for Computational Linguistics.

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

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021b. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023b. 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.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy-yong Sohn, and Alejandro Lopez-Lira. 2025. Finder: Financial dataset for question answering and evaluating retrieval-augmented generation. In *Proceedings of the 6th ACM International Conference on AI in Finance*, ICAIF '25, page 638–646, New York, NY, USA. Association for Computing Machinery.

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.

Zheng Chu, Zekun Wang, Jiafeng Liang, Ming Liu, and Bing Qin. 2023. MTGER: Multi-view temporal graph enhanced temporal reasoning over time-involved document. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15218–15233, Singapore. 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.

Irwin Deng, Kushagra Dixit, Dan Roth, and Vivek Gupta. 2025. Enhancing temporal understanding in LLMs for semi-structured tables. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4936–4955, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

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. 2025. Test of time: A benchmark for evaluating LLMs on temporal reasoning. In *The Thirteenth International Conference on Learning Representations*.

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.

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

Yubin Ge, Salvatore Romeo, Jason Cai, Raphael Shu, Yassine Benajiba, Monica Sunkara, and Yi Zhang. 2025. TReMu: Towards neuro-symbolic temporal reasoning for LLM-agents with memory in multi-session dialogues. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18974–18988, Vienna, Austria. Association for Computational Linguistics.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2025. ComplexTempQA: A 100m dataset for complex temporal question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9100–9112, Suzhou, China. Association for Computational Linguistics.

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 *Proceedings 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 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.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

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.

Atharv Kulkarni, Kushagra Dixit, Vivek Srikumar, Dan Roth, and Vivek Gupta. 2025. LLM-symbolic integration for robust temporal tabular reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19914–19940, Vienna, Austria. Association for Computational Linguistics.

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.

Xingxuan Li, Liying Cheng, Qingyu Tan, Hwee Tou Ng, Shafiq Joty, and Lidong Bing. 2023. Unlocking temporal question answering for large language models with tailor-made reasoning logic. *arXiv preprint arXiv:2305.15014*.

Feng Liang, Weixin Zeng, Runhao Zhao, and Xiang Zhao. 2026. Nestr: A neuro-symbolic abductive framework for temporal reasoning in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):31907–31915.

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.

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.

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.

Gerrit Mutschlechner and Adam Jatowt. 2025. Analyzing the role of context in forecasting with large language models. *arXiv preprint arXiv:2501.06496*.

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.

Amy L. Olex and Bridget T. McInnes. 2021. Review of temporal reasoning in the clinical domain for timeline extraction: Where we are and where we need to be. *Journal of Biomedical Informatics*, 118:103784.

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, Zehra Mert, and Adam Jatowt. 2026. How often do answers change? estimating recency requirements in question answering. *arXiv preprint arXiv:2603.16544*.

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).

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.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2025. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *The Thirteenth International Conference on Learning Representations*.

Abhilash Shankarampeta, Harsh Mahajan, Tushar Kataria, Dan Roth, and Vivek Gupta. 2025. TRANSIENTTABLES: Evaluating LLMs' reasoning on temporally evolving semi-structured tables. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6526–6544, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

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.

Miao Su, Zixuan Li, Zhuo Chen, Long Bai, Xiaolong Jin, and Jiafeng Guo. 2024a. Temporal knowledge graph question answering: A survey. *arXiv preprint arXiv:2406.14191*.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min Zhang. 2024b. 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.

Cui Tao, Harold R Solbrig, Deepak K Sharma, Wei-Qi Wei, Guergana K Savova, and Christopher G Chute. 2010. Time-oriented question answering from clinical narratives sing semantic-web techniques. In *Proceedings of the 9th international semantic web conference on The semantic web-Volume Part II*, pages 241–256.

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.

Santosh T.y.s.s, Tuan-Quang Vuong, and Matthias Grabmair. 2024. ChronosLex: Time-aware incremental training for temporal generalization of legal classification tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3022–3039, Bangkok, Thailand. Association for Computational Linguistics.

Md Nayem Uddin, Amir Saeidi, Divij Handa, Agastya Seth, Tran Cao Son, Eduardo Blanco, Steven Corman, and Chitta Baral. 2025. UnSeenTimeQA: Time-sensitive question-answering beyond LLMs' memorization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1873–1913, Vienna, Austria. Association for Computational Linguistics.

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.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

Juraj Vladika and Florian Matthes. 2024. Improving health question answering with reliable and time-aware evidence retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4752–4763, Mexico City, Mexico. Association for Computational Linguistics.

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*, pages 15685–15705, 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.

Zhao Wang, Ziliang Zhao, and Zhicheng Dou. 2025. Timerag: Enhancing complex temporal reasoning with search engine augmentation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 3230–3239, New York, NY, USA. Association for Computing Machinery.

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.

Georg Wenzel and Adam Jatowt. 2023. An overview of temporal commonsense reasoning and acquisition. *arXiv preprint arXiv:2308.00002*.

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.

Jiaying Wu and Bryan Hooi. 2025. Chain-of-timeline: Enhancing LLM zero-shot temporal reasoning with SQL-style timeline formalization. In *Workshop on Reasoning and Planning for Large Language Models*.

Zhixuan Wu, Bo Cheng, Jiale Han, Jiabao Ma, Shuhao Zhang, Yuli Chen, and Changbo Li. 2025. Videoqata: Temporal-aware multi-modal video question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7239–7252.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024a. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

Siheng Xiong, Yuan Yang, Ali Payani, James C Kerce, and Faramarz Fekri. 2024b. 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.

Siyue Zhang, Yuxiang Xue, Yiming Zhang, Xiaobao Wu, Anh Tuan Luu, and Chen Zhao. 2025. MRAG: A modular retrieval framework for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3080–3118, Suzhou, China. 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.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

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.

Li Zhou and George Hripcsak. 2007. Temporal reasoning with medical data—a review with emphasis on medical natural language processing. *Journal of biomedical informatics*, 40(2):183–202.

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

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*.

Xinyu Zhu, Cheng Yang, Bei Chen, Siheng Li, Jian-Guang Lou, and Yujiu Yang. 2023. Question answering as programming for solving time-sensitive questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12775–12790, Singapore. Association for Computational Linguistics.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Large language models for information retrieval: A survey. *ACM Trans. Inf. Syst.*, 44(1).

A Related Surveys

Advances in temporal datasets, time-aware models, and temporal reasoning techniques have enabled systems capable of retrieving time-relevant documents, ordering events, and answering temporally constrained questions across applications such as historical analysis, fact-checking, and intelligent assistants. Although temporal information processing has received considerable attention in the literature, existing surveys have not provided a comprehensive, end-to-end analysis of TQA systems over unstructured text. Prior work can be grouped into broad areas, each with limitations relative to our focus.

While QA has 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., 2025), whereas QA surveys center on extractive, abstractive, or multi-hop answering (Zhu et al., 2021; Mavi et al., 2024). These works rarely consider temporal intent, dynamic or evolving information needs, or event sequencing, highlighting a key gap in understanding temporally conditioned QA.

Several earlier works provided foundational insights into TIR. 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. However, despite substantial progress in temporal reasoning with modern LLMs, there has been no recent systematic overview that integrates contemporary text-based TQA datasets, models, and evaluation paradigms.

Crucially, existing temporal IR and temporal processing surveys adopt a largely component-level perspective, focusing on tasks such as temporal tagging, document dating, or time-aware ranking, rather than temporal question answering as an end-to-end problem. Likewise, general QA surveys typically treat time as an implicit or secondary factor and do not systematically analyze temporally constrained queries or diachronic evidence selection. Our survey fills this gap by explicitly framing text-based TQA as a distinct problem setting and synthesizing recent datasets, models, and evaluation paradigms that capture the interaction between temporal reasoning, retrieval, and answer generation.

Temporal QA has also been studied in structured and semi-structured settings, including temporal knowledge graphs (Jia et al., 2021; Saxena et al., 2021; Chen et al., 2023b; Xiong et al., 2024b) and, more recently, evolving tables. Tabular temporal QA approaches reason over time-indexed records or attribute updates (Shankarampeta et al., 2025; Deng et al., 2025; Kulkarni et al., 2025), relying on explicit temporal schemata, row-level timestamps, and structured event representations. Their central challenges, such as temporal joins, value-update tracking, and schema-aware inference, differ fundamentally from grounding implicit temporal expressions, resolving document-relative timestamps, or retrieving diachronic evidence in natural language. Because our survey focuses on temporality as expressed in unstructured text, we treat KG-based and table-based TQA as out of scope while acknowledging their relevance as complementary research areas. Hybrid datasets such as TempTabQA (Gupta et al., 2023) sit at the boundary by extracting semi-structured information (e.g., infoboxes) from text, yet still rely primarily on natural-language context.

Our survey focuses on temporally aware QA over unstructured 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 TQA. 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 Key Concepts

We first introduce the core concepts related to TQA. **Temporal Information Retrieval (TIR)** aims to retrieve documents that are not only topically relevant but also aligned with the query's **temporal intent**. This intent may be explicit, such as *"Olympics 2024"*, or implicit, such as *"latest Apple earnings"*. TIR relies on different **temporal signals**, including **document timestamps** (publication dates), **temporal expressions** (*"March 2023"*), and **event mentions** (*"2024 Olympics"*) to assess a document's **temporal relevance**, which indicates how well its content aligns with the query's intended time frame (Kanhabua and Nørvåg, 2008; Singh et al., 2016). While TIR facilitates time-sensitive access to documents, it often lacks mechanisms for performing deeper temporal reasoning over the content itself.

On the other hand, **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 published around that time. Unlike TIR, TQA often involves more sophisticated processing, including interpreting **temporal expressions**, **ordering events**, and grounding answers in the appropriate time frame. As a result, it frequently requires multi-hop reasoning across documents and temporal signals.

TQA relies on diverse temporal cues. *Temporal signals* include **explicit temporal expressions** like *"March 2023"*, **relative expressions** such as *"last week"*, **implicit cues** like *"recently"*, and event-based references like *"2024 Olympics"*. These require contextual interpretation and sometimes external temporal knowledge. Temporal metadata, such as **document timestamps**, indicate publication time and often serve as proxies for judging the freshness of content.

A critical yet often underutilized concept is that of *document focus time*, the time period a document primarily discusses, which may differ from its publication timestamp. For example, a news article published in 2013 about the 2010 Academy Awards has a publication date of 2013 but a focus time of 2010. This distinction is crucial in TQA, where users often seek information tied to specific historical or future contexts rather than content merely anchored to when it was written.

Estimating focus time is non-trivial, often requiring temporal expression normalization, burst detection, or event timestamping. When accurately modeled, it improves retrieval quality in time-sensitive settings, especially for long-range archives or temporally anchored questions. Prior work has shown that focus time estimation can enhance performance in tasks such as historical QA, event ordering, and temporally grounded search (Jatowt et al., 2013; Wang et al., 2020).

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 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.

D Rule-based & Statistical Methods

Early work in Temporal 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 re-ranking 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 sub-questions using temporal expression taggers like TERSEO (Saquete et al., 2003). These approaches showed improved precision and generalizability across languages.

In retrospect, rule-based and statistical systems introduced many principles that remain relevant: time expression normalization, explicit modeling of document recency and user intent, and the integration of temporal structure into retrieval and reasoning pipelines. While their lack of adaptability to diverse and large-scale contexts eventually limited their impact, they provided a crucial foundation that modern neural and retrieval-augmented models continue to build upon.

E Domain-Specific Temporal QA

While much of the existing research in TQA focuses on open-domain settings using Wikipedia and news corpora, domain-specific applications present unique temporal reasoning challenges that remain largely underexplored. Unlike general TQA systems that handle diverse topics with explicit temporal markers, specialized domains require an understanding of domain-specific temporal conventions, implicit temporal relationships, and field-specific accuracy requirements.

Medical Domain Temporal reasoning in clinical narratives has been explored in foundational work. The TimeText system, for instance, demonstrated early success in temporal clinical QA, achieving 84% accuracy on 147 time-oriented questions about discharge summaries (Zhou et al., 2008). Other efforts have used semantic web techniques to answer time-based clinical questions, such as event sequencing in patient histories (Tao et al., 2010).

Despite comprehensive reviews of temporal reasoning in medical text processing (Olex and McInnes, 2021; Zhou and Hripcsak, 2007), modern temporal QA systems for medical applications remain limited. Recent work, such as Improving Health Question Answering with Reliable and Time-Aware Evidence Retrieval (Vladika and Matthes, 2024), incorporates recency and citation count for prioritizing reliable evidence, but falls short of full temporal reasoning. Open challenges remain in multi-hop reasoning over patient timelines, temporally scoped symptom progression, and treatment-event linking—tasks crucial to clinical decision support yet largely unexplored from a QA perspective.

Legal Domain Legal systems present inherent temporal complexity, shaped by evolving statutes, shifting precedents, and regulatory amendments. While several legal QA datasets exist, such as Büttner and Habernal (2024); Kien et al. (2020); Chen et al. (2023a); Zhong et al. (2020), most are focused on static question answering and do not address the temporal dynamics of legal knowledge. Recent advances like ChronosLex (T.y.s.s et al., 2024) introduce time-aware incremental training for better generalization across legal epochs, signaling early momentum toward temporal legal understanding.

However, most legal QA systems do not yet perform event ordering, precedent timeline alignment, or reasoning over legislative changes. For example, questions such as *"Which case law was applicable prior to the 2015 amendment of the privacy statute?"* remain out of scope for current systems. There is a pressing need for temporal-aware retrieval and reasoning architectures capable of navigating legal timelines and citations.

Financial Domain In the financial domain, temporal modeling is well-established through time-series forecasting methods used for stock price prediction, volatility estimation, and risk assessment. However, these approaches typically operate on structured numerical data and do not engage in temporal reasoning over natural language. As a result, financial temporal QA remains a largely unexplored frontier.

Several financial QA datasets have emerged in recent years, primarily focused on numerical reasoning. Notable examples include FinQA (Chen et al., 2021b), FinTextQA (Chen et al., 2024), and FinDER (Choi et al., 2025).

While these resources advance quantitative and factual QA, they fall short of supporting temporal inference. Complex questions such as *"How have regulatory changes since 2008 affected current banking policies?"* require systems to track policy evolution, integrate document publication dates, and synthesize trends across multiple temporal anchors (capabilities not currently addressed by existing models or datasets). Bridging this gap would require designing temporal-aware retrieval systems and multi-hop reasoning frameworks tailored to financial text, capable of answering questions grounded in historical context and regulatory shifts.

Real-World Deployment and Domain-Specific Challenges Deploying temporal QA systems in high-stakes, real-world settings presents challenges that require specialized reasoning capabilities. In the medical domain, errors in temporal interpretation can adversely affect clinical decision-making, demanding high standards of accuracy and interpretability. Legal information systems must navigate jurisdiction-specific temporal expressions (e.g., "Pre-Brexit regulatory regime") and shifting hierarchies of legal precedent that differ across regions and evolve over time. Financial applications require real-time temporal understanding for tasks such as regulatory compliance and market analysis, where even minor misalignments in temporal reasoning can lead to significant operational or economic consequences.

Unlike general-purpose systems that rely on clearly stated temporal cues, domain-specific QA often involves inferring timeframes from implicit, field-specific conventions such as the "perioperative period" in medicine, the "discovery phase" in legal contexts, or "earnings season" in financial reporting. These terms reflect temporally bounded phases that are well understood by domain experts but opaque to models lacking domain awareness. Addressing such challenges demands tailored temporal reasoning approaches that incorporate expert knowledge, event timelines, and contextual interpretation.

F Detailed Future Research Directions

This appendix provides a comprehensive analysis of the research challenges outlined in Section 5, detailing specific technical problems, examples, and potential research approaches for each identified direction.

F.1 Dynamic Temporal Knowledge Management

TQA systems face a persistent temporal consistency problem: their reliance on static corpora makes them unable to respond to fast-evolving information needs (Dhingra et al., 2022; Wallat et al., 2024). As real-world facts change, even small updates can disrupt temporal dependencies across related events, durations, and causal chains, a challenge known as the temporal propagation problem (Vu et al., 2024). This exposes a fundamental limitation in current architectures: they lack the modularity and temporal reasoning structure needed to adapt efficiently (Han et al., 2021). Future systems must move beyond isolated fact updates toward scalable frameworks that track, edit, and reason over temporal dependencies in real time.

For example, if a political leader's term end's date changes, this affects not only direct questions about their tenure but also questions about policies enacted *"during their presidency*" or events that occurred *"before they left office."* Current systems often fail to propagate such updates, leading to inconsistencies. To address this, future systems should: (1) identify temporal dependency chains when facts change, (2) propagate updates through related facts, (3) maintain multi-hop consistency, and (4) reconcile conflicting temporal evidence. Integrated temporal knowledge graphs offer a promising direction for such dynamic reasoning.

F.2 Temporally-Aware LLM Agents

Current LLM agents face severe limitations in temporal reasoning that become pronounced in interactive settings. They consistently exhibit temporal hallucinations, generating plausible but temporally incorrect information, and fail to resolve context-dependent expressions such as *"last Tuesday"* or "*since our previous discussion*" (Xiong et al., 2024a; Bazaga et al., 2025). These failures often lead to incoherent or contradictory answers across dialogue turns.

These failures stem from architectural constraints: transformers lack persistent temporal working memory and struggle with cross-turn temporal reference resolution (Ge et al., 2025). As a result, agents treat dialogue turns largely independently, causing temporal anchors to drift over extended interactions.

Addressing these issues requires explicit temporal modeling mechanisms, including timeline tracking across turns, event memory that preserves temporal relations, multi-turn consistency enforcement, and robust temporal reference resolution.

Concretely, a minimally viable temporal agent would extend a base LLM with three explicit components: (i) a persistent temporal memory that stores events and dialogue states indexed by time, (ii) a timeline tracking module that maintains and updates temporal anchors across interactions, and (iii) a consistency mechanism that verifies newly generated responses against stored temporal constraints. Even without full symbolic reasoning, such an agent would mitigate common failure modes such as temporal drift, inconsistent reuse of temporal references, and hallucinated event ordering.

Existing multi-session memory and neuro-symbolic approaches already implement individual components of this pipeline—e.g., timeline construction and self-verification (Bazaga et al., 2025), symbolic temporal constraint checking (Liang et al., 2026), and executable temporal reasoning (Zhu et al., 2023)—but integrating them into a unified temporal agent remains an open research challenge.

F.3 Diachronic and Synchronic Knowledge Integration

Temporal questions often require combining diachronic sources (which capture change over time) and synchronic sources (which reflect knowledge at a specific moment). Most TQA systems treat these separately (Wang et al., 2022; Gruber et al., 2025; Piryani et al., 2024b), limiting their ability to fully answer questions that span both historical context and present-day facts.

For instance, the question *"How has unemployment changed since 2008, and what is the current rate?"* requires integrating long-term trends from news archives (diachronic) with recent statistical summaries (synchronic). Current models often handle only one type of source effectively.

This integration requires more than merging sources; it demands alignment across different temporal granularities and anchoring schemes. Conflicts between historical accounts and retrospective summaries can further complicate the reasoning process.

Future systems should develop: (1) temporal alignment algorithms that can map events across different temporal representations, (2) cross-source reasoning frameworks that can weight and combine evidence from different temporal paradigms, (3) conflict resolution mechanisms for handling disagreements between diachronic and synchronic accounts, and (4) unified temporal representations that can accommodate both evolving and static knowledge seamlessly.

F.4 Temporal Uncertainty and Confidence Modeling

Many historical events involve uncertain or approximate dates, yet most TQA systems treat all temporal information as exact. This creates a mismatch with real-world complexity, particularly in multi-step reasoning where uncertainty can propagate and amplify.

For instance, the question *"What happened between the fall of Rome and the beginning of the Renaissance?"* involves fuzzy boundaries: the fall of Rome (often dated to 476 AD) and the Renaissance (spanning the 14th to 16th centuries) are debated and regionally variable. Systems must represent and reason over this imprecision, not resolve it artificially.

Temporal uncertainty also arises in natural language expressions like *"around that time"* or *"in the late period,"* and from conflicting historical accounts. Ignoring these leads to overconfident and potentially misleading answers.

Future systems should incorporate: (1) probabilistic temporal representations that can model uncertain dates and durations, (2) confidence propagation mechanisms for multi-step temporal reasoning, (3) uncertainty-aware answer generation that communicates temporal confidence to users, (4) conflict-aware reasoning that can handle and explain disagreements between sources, and (5) evaluation frameworks that reward appropriate uncertainty rather than false precision (Vashishtha et al., 2019; Zhou et al., 2021; Fatemi et al., 2025).

F.5 Multilingual and Multimodal Temporal QA

Temporal expressions, date formats, and cultural references vary widely across languages and modalities, yet most TQA systems are developed primarily for English and textual input. This poses challenges in multilingual and multimodal settings where temporal cues appear in varied forms such as non-Gregorian date formats in documents, seasonal imagery in videos, or handwritten timestamps in scanned texts. Current systems struggle to interpret these diverse signals due to limited cultural grounding and inadequate multimodal integration. For example, a model may fail to recognize a lunar calendar reference in Arabic or temporal context from a video of a snowstorm, suggesting winter.

Recent benchmarks like TimeDial (Qin et al., 2021) and TOMATO (Shangguan et al., 2025) highlight the need for culturally and visually grounded temporal reasoning. Future research should focus on developing multilingual temporal taggers, temporally annotated datasets in low-resource languages, and cross-modal alignment techniques that jointly reason over text, images, and video to capture time-related meaning in diverse cultural contexts (Qin et al., 2021; Shangguan et al., 2025; Wu et al., 2025).

F.6 Implicit Temporal Intent Understanding

Many temporal questions conceal their intended timeframes, making implicit temporal intent one of the most difficult challenges in TQA. For example, the question, *"What caused the economic crisis during Trump's presidency?"* could refer to events between 2017-2021 (e.g., COVID-19 or trade-related downturns) or more recent crises if the question is asked retrospectively. Current systems often fail to detect such ambiguity, either assuming the most recent crisis, misinterpreting the temporal anchor, or overlooking the ambiguity entirely.

This challenge stems from various factors: (1) conversational context, where prior dialogue establishes implicit timeframes; (2) cultural assumptions, such as shared historical reference points; (3) domain conventions, where temporal scope is understood implicitly; and (4) individual user perspectives or real-time situational context (e.g., ongoing events).

Existing models rely heavily on explicit temporal expressions and struggle with contextual inference (Zhou et al., 2021; Allein et al., 2023; Chu et al., 2024). Future work should focus on: (1) contextual intent detection models, (2) user and domain-aware temporal reasoning, (3) clarification strategies for resolving ambiguity, and (4) evaluation frameworks for temporal intent detection across diverse scenarios.

F.7 Evaluation and Benchmarking for Temporal Reasoning

Current evaluation practices for TQA largely rely on standard IR/NLP metrics such as accuracy, F1, MRR, or NDCG. While useful, these metrics often overlook the specific demands of temporal reasoning, such as correct temporal anchoring, event ordering, consistency across multi-hop inferences, and robustness to temporal ambiguity.

In Temporal QA, benchmarks such as TimeQA (Chen et al., 2021a), MenatQA (Wei et al., 2023), ComplexTempQA (Gruber et al., 2025), TimeBench (Chu et al., 2024) and TEMPO (Abdallah et al., 2026a) provide test sets that target temporally sensitive tasks, including temporal entailment, counterfactual reasoning, and timeline ordering. However, evaluation often reduces to span-level accuracy or multiple-choice correctness. These benchmarks represent early attempts to probe temporal reasoning beyond static fact recall, but their evaluation protocols remain largely inherited from non-temporal QA.

In Temporal IR, metrics sometimes integrate recency weighting or time decay functions (Berberich et al., 2010), but still conflate temporal and topical relevance.

Future work should aim to develop: (1) evaluation metrics that capture temporal grounding and coherence, (2) metrics sensitive to ambiguous or conflicting time signals, and (3) unified protocols for comparing temporal capabilities across models and tasks.

An important methodological concern in Temporal QA is the potential overlap between benchmark datasets and large language model pretraining corpora. Many TQA datasets (e.g., TimeQA (Chen et al., 2021a), ArchivalQA (Wang et al., 2022)) are derived from Wikipedia or news archives that are likely included in model training data, raising questions about whether reported performance reflects genuine temporal reasoning or memorization. Some benchmarks attempt to mitigate this issue through time-restricted or post-training evaluation: RealTimeQA (Kasai et al., 2023) and FreshQA (Vu et al., 2024) focus on events occurring after model training cutoffs, while Test of Time (Fatemi et al., 2025) employs synthetic tasks designed to minimize reliance on memorized world knowledge. However, most existing benchmarks lack systematic contamination analysis or explicit controls for pretraining overlap. Addressing this gap will require temporal decontamination protocols and evaluation on temporally out-of-distribution data.

Method	Type	Task	Temporal Representation	Temporal Signals	Architecture
TempoT5 (Dhingra et al., 2022)	Encoder-Decoder	TQA	Explicit timestamps	Document timestamps + temporal context	T5 + temporal conditioning
TempoBERT (Rosin et al., 2022)	Encoder	Semantic Change Detection, Sentence Time PredictionSemantic Change Detection, Sentence Time Prediction	Explicit timestamps	Time tokens	BERT + temporal masking
BiTimeBERT (Wang et al., 2023)	Encoder	Event Time Estimation, Document Dating, TQA	Explicit timestamps and content temporal expressions	Document timestamps + temporal expressions in text	BERT + dual temporal encoding
TALM (Ren et al., 2023)	Temporal adaptation + Hierarchical modeling	Historical Dating	Implicit (Time-specific word variants)	Time-specific representations	BERT + Temporal Adaptation + Hierarchical Document Encoder
SG-TLM (Su et al., 2023)	TLM	Temporal Language Modeling, TQA	Implicit Temporal Information	Syntactic role distributions + timestamps	BERT + syntax-guided masking + temporal-aware masking
TSM (Cole et al., 2023)	Encoder	Temporal language modelling, TQA	Implicit temporal spans	Temporal expressions (dates, durations)	T5 + temporal span masking
Temporal Attention (Rosin and Radinsky, 2022)	Encoder	Semantic Change Detection	Explicit Document timestamp	Document timestamp	Transformer encoder with time-aware attention via time matrix T
TempRetriever (Abdallah et al., 2026b)	Neural Embedding	TQA	Explict+Implicit temporal information	Query timestamps + document timestamps	Semantic + Temporal Encoder
TsContriever (Wu et al., 2024)	Neural Retrieval	TQA	Explict+Implicit temporal information	Query timestamps + document timestamps	Semantic + Temporal Encoder
TempRALM (Gade and Jetcheva, 2024)	RAG	TQA	Explicit temporal Information	Query + document timestamps	RAG + temporal enhancement
TimeR4 (Qian et al., 2024)	RAG	TQA	Explicit temporal facts	Question constraints + TKG timestamps	Retrieve-Rewrite-Retrieve-Rerank with contrastive time-aware retriever + LLaMA2 reasoning
MRAG (Zhang et al., 2025)	Neural + Symbolic	TQA	Explicit temporal information	Query constraints + Document timestamps	Modular RAG framework (Question Processing, Retrieval+ Summarization +Semantic-Temporal Hybrid Ranking

Table 2: Comprehensive summary of prominent approaches to TQA and temporally-aware language understanding. The table classifies each method by: *Type*, indicating the architectural paradigm, *Task*, specifying the temporal reasoning or language understanding objective; *Temporal Representation*, describing whether temporal information is explicitly or implicitly modeled; *Temporal Signals*, referring to the temporal cues utilized, and *Architecture*, summarizing the model components and any temporal-specific adaptations. This overview highlights the variety of strategies employed to capture, encode, and reason over temporal context.

Method	Input	Output	Knowledge	Processing	Methodology
TempoT5 (Dhingra et al., 2022)	Text + temporal prefix	Masked span prediction	CUSTOMNEWS	Temporal conditioning + uniform sampling	Input prefixing for time-aware modeling and efficient updates
TempoBERT (Rosin et al., 2022)	Text + timestamps	Time tokens	News corpus	Implicit temporal reasoning via time masking	Time-aware pretraining with masking
BiTimeBERT (Wang et al., 2023)	News articles with timestamps	Time predictions (day/-month/year granularity)	NYT corpus	Bi-temporal reasoning using both timestamp and content time	Time-Aware Masked Language Modeling, Document Dating
TALM (Ren et al., 2023)	Historical text	Time period classification	Chinese Twenty-Four Histories + English Royal Society Corpus	Temporal word adaptation and contextual learning	Learn separate word representations per time period + alignment
SG-TLM (Su et al., 2023)	Text + timestamp	Masked token + time prediction	WMT News Crawl + Reddit Time Corpus	Syntax-guided masking (SGM) + temporal-aware masking (TAM)	Syntax-based lexicon selection for efficient temporal adaptation
TSM (Cole et al., 2023)	Text with temporal expressions	Masked span prediction	Wikipedia	Temporal span masking using SUTIME parser + intermediate training	SUTIME parser identifies temporal expressions (Time, Duration, Set, Date) for targeted masking during intermediate pre-training
Temporal Attention (Rosin and Radinsky, 2022)	Text + Document timestamp	Time-specific contextualized embeddings	Time-annotated corpora (SemEval datasets)	Time-aware self-attention mechanism with additional time matrix	Extends self-attention with time matrix T to condition attention weights on temporal context via time-aware dot products.
TempRetriever (Abdallah et al., 2026b)	Query + Query timestamp + Document + Document timestamp	passages with temporal +semantic relevance	NYT + ChroniclingAmerica Corpus	Temporal alignment + Temporal understanding	Learned temporal embeddings with fusion techniques + Time-based negative sampling
TsContriever (Wu et al., 2024)	Time-sensitive questions	Top-k temporally relevant documents	Nobel Prize benchmark corpus + Wikidata QA pairs	Contrastive learning + Query-side fine-tuning + Query routing	Supervised contrastive learning with time-aware query tuning and routing
TempRALM (Gade and Jetcheva, 2024)	Time-sensitive queries	Answer	Knowledge base	Temporal proximity-based document ranking	Temporal scoring
TimeR4 (Qian et al., 2024)	Time-sensitive queries	Answer (entity or timestamp)	Knowledge base	Retrieve-Rewrite-Retrieve-Rerank pipeline with contrastive learning and LLM finetuning	Temporal fact retrieval, question rewriting, time-aware reranking, and LLM finetuning
MRAG (Zhang et al., 2025)	Temporal questions + Document	Ranked passages	Wikipedia	Symbolic temporal scoring + Semantic-temporal hybrid ranking	Question decomposition (MC + TC) + Symbolic temporal

Table 3: Detailed comparison of TQA methods across key system dimensions. Each method is described in terms of its *Input* (the structure and temporal components of the data provided to the model), *Output* (the nature of predictions or generated responses), *Knowledge* (the underlying corpora or knowledge sources used), *Processing* (the mechanisms used to handle temporal information, such as conditioning, masking, or alignment), and *Methodology* (the core modeling strategy or training paradigm employed). This table highlights the diverse ways in which temporal signals and knowledge are integrated to enhance reasoning over time-sensitive questions.

Evaluation Goal	Recommended Datasets
Temporal Robustness / Answer Drift	StreamingQA (Liska et al., 2022), RealTimeQA (Kasai et al., 2023), PAT-Questions (Meem et al., 2024), FreshQA (Vu et al., 2024)
Temporal Retrieval (Diachronic Evidence)	ArchivalQA (Wang et al., 2022), ChroniclingAmericaQA (Piryani et al., 2024b), NewsQA (Trischler et al., 2017)
Reasoning-focused Temporal Retrieval	TEMPO (Abdallah et al., 2026a)
Implicit Temporal Intent	TiQ (Jia et al., 2024), TemporalQuestions (Wang et al., 2021a), ArchivalQA (Wang et al., 2022)
Event Ordering	TORQUE (Ning et al., 2020), TRACIE (Zhou et al., 2021), TimeBench (Chu et al., 2024)
Duration Inference	MenatQA (Wei et al., 2023), TEMPREASON (Tan et al., 2023), TimeBench (Chu et al., 2024)
Multi-hop Temporal Reasoning	ComplexTempQA (Gruber et al., 2025), Complex-TR (Tan et al., 2024) , TempQuestions (Jia et al., 2018)
Knowledge Updating	TempLAMA (Dhingra et al., 2022), StreamingQA (Liska et al., 2022), PAT-Questions (Meem et al., 2024)
End-to-End Temporal QA	ComplexTempQA (Gruber et al., 2025), ChroniclingAmericaQA (Piryani et al., 2024b) TimeQA (Chen et al., 2021a), ArchivalQA (Wang et al., 2022)
Temporal Ambiguity Resolution	SituatedQA (Zhang and Choi, 2021), TEMPAMBIQA (Piryani et al., 2024a)

Table 4: Dataset selection guide for temporal question answering. Each row summarizes which benchmarks are most suitable for evaluating specific temporal capabilities.