IT'S HIGH TIME: A SURVEY OF TEMPORAL QUESTION ANSWERING

Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari,
Avishek Anand, Adam Jatowt

May 2025

HEIDELBERG UNIVERSITY
CROSS-TEMPORAL NLP
BUSE ERKIRAZ

13 January 2025

1. INTRODUCTION

Temporal Question Answering (TQA) is a subfield of QA that focuses on answering questions constrained by or dependent on time, requiring systems to go beyond surface-level retrieval and processing and engage in temporal understanding and reasoning (Piryani et al., 2025).

1. INTRODUCTION

Motivation

Time is a core dimension shaping information relevance and interpretation.

Growing volume of time-stamped text (news, social media, archives) intensifies temporal challenges

1. INTRODUCTION

Core Challenges

Temporal ambiguity resolution: vague expressions (e.g., “recently”, “after the war”) require contextual grounding

Cross-temporal reasoning: reasoning over causal and sequential relationships between events across time

Knowledge volatility: facts evolve, making static corpora and pre-trained models inadequate

Synchronic Collection

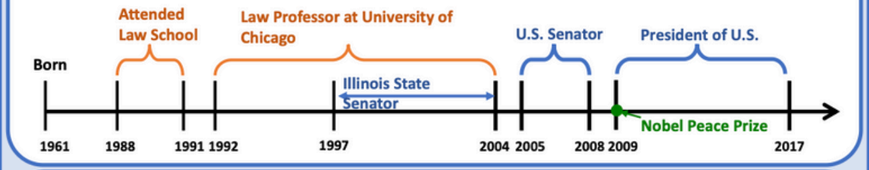
Document from Synchronic Collection

Barack Hussein Obama (born August 4, 1961) is an American politician who was the 44th president of the United States from 2009 to 2017. ..... Obama previously served as a U.S. senator representing Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004. .....

Obama was awarded the 2009 Nobel Peace Prize for efforts in international diplomacy, a decision which drew both criticism and praise. During his first term, his administration responded to the 2008 financial crisis with ..... took steps to combat climate change, signing the Paris Agreement, a major international climate agreement, and an executive order to limit carbon emissions.

Obama enrolled at Harvard Law School in the fall of 1988, living in nearby..... ...graduated from Harvard Law in 1991 with a Juris Doctor magna cum laude. He then taught constitutional law at the University of Chicago Law School for twelve years, first as a lecturer from 1992 to 1996, and then as a senior lecturer from 1996 to 2004.

Extracted Event Timeline from Above Document



Temporal Question

Q1: **At what age did Barack Obama win the Nobel Peace Prize?**

Ans: **48 years old**

Temporal Understanding

Born - 1961
Nobel Peace Prize -2009 → 2009-1961 → 48 years ✓

Diachronic Collection

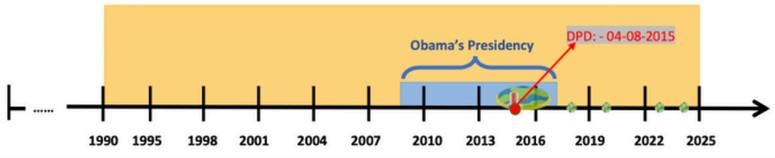
Document from Diachronic Collection

Document Publication Date (DPD): **August 4, 2015**

Obama said, "Today, I announce that we are taking the most significant step in U.S. history to combat climate change. With the Clean Power Plan, we are setting the first-ever national limits on carbon pollution from power plants. By 2030, we will cut emissions by 32% from 2005 levels, and we'll do it by investing in cleaner energy like wind and solar not just for our health, but for the future of our planet."

And next Tuesday, the United States will join nearly 200 nations in the Paris Agreement. This global deal commits us all to limit global warming to well below 2 degrees Celsius. It's a turning point not just for our climate, but for our shared leadership on the world stage. We are showing that the U.S. does not sit on the sidelines when the future of our children is at stake.

Diachronic Collection Publication Timeline



Temporal Question

Q2. What does President Obama's climate policy tell us about how the U.S. viewed climate change during his late years of service?

Ans: President Obama's climate policies, including the Clean Power Plan and Paris Agreement, aimed to cut emissions and lead global action. These actions marked a major step in U.S. climate leadership, though they faced legal battles and political pushback at home.

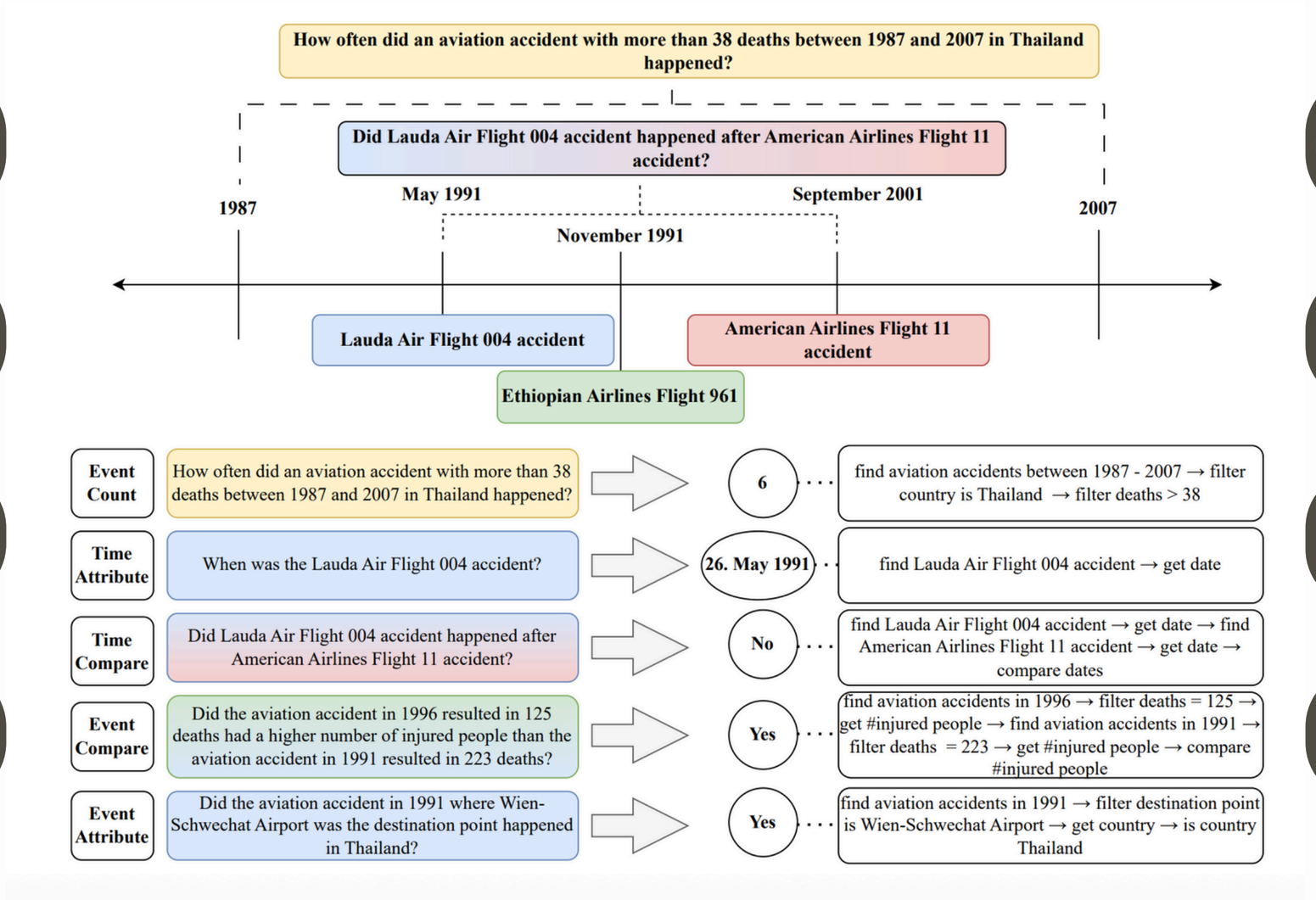
How complex could it be?

One question → multiple temporal reasoning operations

Not a single-hop QA task

Events must be grounded to absolute time

Relative comparisons depend on absolute dates



Different questions require different temporal skills

Failure at any step breaks the final answer

Assumes temporally consistent and complete evidence

Retrieval alone is insufficient

Figure from COMPLEXTEMPQA paper

1. INTRODUCTION

Evolution of TQA Research

Early approaches: rule-based pipelines and statistical models

→ poor scalability and brittleness across temporal contexts

Neural era: pre-trained language models enable

→ Temporal normalization and sequencing

→ Multi-hop temporal inference

→ Adaptation to evolving knowledge via retrieval and continual learning

Retrieval-augmented methods made increasingly central

→ multi-hop temporal inference

→ incorporate up-to-date information

1. INTRODUCTION

Focus

TQA over unstructured text (natural language questions + documents)

Contributions

- a review of TQA methods, benchmark datasets, evaluation strategies, and emerging challenges
- a taxonomy of TQA tasks, datasets, and approaches
- identification of open challenges

Goal

→ guide design of next-generation time-aware QA systems

2. KEY CONCEPTS

Temporal Information Retrieval (TIR)

- Retrieves documents aligned with a query's temporal intent
- Temporal intent can be explicit ("Olympics 2024") or implicit ("latest Apple earnings")

- Uses temporal signals:
  - document timestamps
  - temporal expressions
  - event mentions

- Limitation: typically does not perform deep temporal reasoning over content

Temporal Question Answering (TQA)

- Answers questions with explicit or implicit temporal constraints
- Requires understanding question's temporal intent and selecting temporally appropriate evidence

- Goes beyond retrieval to
  - interpret temporal expressions
  - order events in time
  - ground answers to the correct time frame

- Often involves multi-hop reasoning across documents and temporal signals

2. KEY CONCEPTS

Temporal Signals and Metadata

TQA relies on diverse temporal cues:

→ explicit dates (“March 2023”)

→ relative expressions (“last week”)

→ implicit cues (“recently”)

→ event-based references (“2024 Olympics”)

- Temporal cues require contextual interpretation
- Document timestamps indicate publication time and often proxy content freshness
- Publication time alone is insufficient for temporal grounding

Document Focus Time

The document’s focus time may differ from publication date.

→ Misalignment leads to incorrect temporal grounding

Estimating DFT requires

→ temporal expression normalization

→ burst detection

→ event timestamping

Accurate focus time modeling improves historical QA, event ordering and temporally grounded search.

3. DATASETS AND EVALUATION BENCHMARKS

Progress in TQA depends on temporally grounded datasets and evaluation protocols.

Datasets serve two roles:

- Knowledge sources (document collections)
- Evaluation benchmarks (question–answer datasets)

3.1 TEMPORAL DOCUMENT COLLECTIONS

Prior work used three types of temporally structured document collections.

1. Diachronic corpora

→ time-stamped documents spanning long periods

they support:

- retrospective retrieval
- diachronic analysis
- event-based reasoning

Commonly used corpora:

- News archives (decades of coverage)
- Large-scale crawled news collections

→ **New York Times Annotated Corpus** (1987–2007; 1.8m articles)

→ ArchivalQA (Wang et al., 2022)

→ **CNN/Daily Mail** (2007–2015; 313k articles)

→ NewsQA (Trischler et al., 2017)

→ **Chronicling America** (1800-1920)

→ ChroniclingAmericaQA (Piryani et al., 2024)

→ **Newswire corpus** (1878-1977)

→ **CUSTOMNEWS** (1969-2019)

3.1 TEMPORAL DOCUMENT COLLECTIONS

2. Synchronic corpora

→ secondary sources

→ snapshots of the world at a specific time

Can be derived from Wikipedia dumps

- TimeQA (Chen et al., 2021)
- TEMPREASON (Tan et al., 2023)
- ComplexTempQA (Gruber et al., 2024)

→ Used in datasets for temporally scoped QA

3.1 TEMPORAL DOCUMENT COLLECTIONS

3. Annotated Temporal Corpora

→ Include explicit temporal annotations for more structured forms of temporal reasoning.

TIMEBANK - an annotated corpus to indicate events, times, and temporal relations (Pustejovsky et al., 2003).
TimeML - a standard annotation schema for representing temporal information in text.

Following datasets:
WikiWars (Mazur and Dale, 2010) - historical narratives
RED (O'Gorman et al., 2016) - causal relations

3.2 TQA DATASETS

TQA datasets asses systems' ability to answer temporally grounded questions.

They differ along multiple dimensions:

- knowledge source
- temporal orientation
- temporal explicitness
- reasoning complexity

Knowledge source

Primary-source datasets (diachronic corpora):

→ Built from contemporaneous historical documents

→ Test retrieval and reasoning over temporally anchored document collections

→ Support authentic historical perspectives

→ **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., 2024)

Secondary-source datasets (synchronic corpora):

→ Built from retrospective snapshots (e.g., Wikipedia)

→ Provide temporally consistent but reconstructed views

→ Enable fine-grained reasoning within a fixed temporal scope

→ **TimeQA**, **TEMPREASON**, **TiQ** (Jia et al., 2024), **ComplexTempQA**

3.2 TQA DATASETS

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	Extrcative	1987-2007	✓	✗
TempLAMA (Dhingra et al., 2022)	50k	News	CS	Extractive	2010-2020	✓	✗
ComplexTempQA (Gruber et al., 2024)	100,228k	Wikipedia	AG	Extractive	1987-2023	✓	✓
MenatQA (Wei et al., 2023)	2.8k	Wikipedia	AG	Extractive	1367-2018	✗	✗
PAT-Question (Meem et al., 2024)	6,1k	Wikipedia	CS	Extractive	-	✗	✓
TempTabQA (Gupta et al., 2023)	11.4k	Wikipedia Info box	CS	Abstractive	-	✗	✗
SituatedQA (Zhang and Choi, 2021)	12.2k	Wikipedia	CS	-	≤2021	✗	✗
UnSeenTimeQA (Uddin et al., 2024)	3.6k	Synthetic	AG	Abstractive	-	✗	✓
ChroniclingAmericaQA (Piryani et al., 2024b)	485k	News	AG	Extractive	1800-1920	✓	✗
FRESHQA (Vu et al., 2024)	600	Google Search	CS	-	-	✗	✓
COTEMPQA (Su et al., 2024b)	4.7k	Wikidata	CS	Abstractive	≤2023	✗	✓
Test of Time (ToT) (Fatemi et al., 2024)	1.8k	Synthetic	AG	Abstractive	-	✗	✓
TIMEDAIL (Qin et al., 2021)	1.1k	DailyDialog	CS	Multiple-choice	-	✗	✗
Complex-TR (Tan et al., 2024)	10.8	Wikipedia+Google Search	AG	Multi-answer	≤2023	✗	✓
StreamingQA (Liska et al., 2022)	147k	News	Cs	Extractive	2007-2020	✓	✓
TRACIE (Zhou et al., 2021)	5.4k	Wikipedia	CS	abstractive	≤2020	✗	✗
ForecastQA (Jin et al., 2021)	10.3k	News	CS	Multiple-Choice	2015-2019	✓	✓
TEMPREASON (Tan et al., 2023)	52.8k	Wikipedia/Wikidata	SC	Abstractive	634-2023	✗	✗
TemporalAlignmentQA (Zhao et al., 2024)	20k	Wikipedia	AG	Abstractive	2000-2023	✗	✗
ReaLTimeQA (Kasai et al., 2023)	5.1k	Search	CS	Multiple-choice	2020-2024	✗	✗

3.2 TQA DATASETS

Temporal Orientation

Most datasets focus on **past events**.

→ **Future-oriented QA** is rare but increasingly important.
→ they test predictive and hypothetical reasoning

FutureContext (Mutschlechner and Jatowt, 2025) and **TimeBench** (Chu et al., 2024) include questions on future events to test timeline projections and forecast-based inference

Question Type

Explicit temporal questions:

- contain clear temporal markers
- temporal intent directly specified

→ "What happened in 1947?"
→ **TimeQA**, **SituatedQA** (Zhang and Choi, 2021), **TempQuestions** (Jia et al., 2018)

Implicit temporal questions:

- omit explicit time references
- require inferring time from events or context

→ "Who was Prime Minister of the UK when the Berlin Wall fell?"
→ **TiQ** (Jia et al., 2024), **TORQUE** (Ning et al., 2020)

3.2 TQA DATASETS

Temporal Reasoning Complexity

Simple temporal questions:

- direct lookups
- identify dates or facts valid at a given time

→ **NewsQA**, **TempLAMA** (Dhingra et al., 2022)

Complex temporal questions:

- Require multi-hop reasoning, temporal filtering, or synthesizing info across events

→ **MenatQA** (Wei et al., 2023), **TempReason** (Tan et al., 2023), **Complex-TR** (Tan et al., 2024), and **ComplexTempQA** (Gruber et al., 2024)

"What major international agreements were signed after World War I but before World War II?"

Limitations and challenges

→ Answers to temporal questions change over time

→ Most datasets are static snapshots

- exceptions are **RealTimeQA** (Kasai et al., 2023), **FreshQA** (Vu et al., 2024) and **PATQA** (Meem et al., 2024)

→ temporal ambiguity

→ structural biases

→ annotation trade-offs: small, high-quality crowdsourced data vs. large, noisy automatic data

4. APPROACHES IN TEMPORAL QA

TQA approaches span rule-based, statistical, and neural/LLM-based paradigms

Methods differ in:

- temporal representations
- reasoning over temporal relations
- adaptation to changing world knowledge

TQA relies on some core **Temporal Prediction Tasks**:

- Event Dating, Document Dating, Focus Time Estimation, Query Time Profiling, and Event Occurrence Prediction

4.1 TEMPORAL LANGUAGE MODELS

Specialized neural models that explicitly incorporate temporal signals during pretraining or fine-tuning to capture temporal dependencies and contextual nuances.

Input-Level Integration

→ TempoT5, TempoBERT, and BiTimeBERT prepend timestamps or temporal expressions directly to training inputs.

Architectural Modifications

→ **TALM** (Ren et al., 2023) learns distinct short-term and long-term temporal word representations, models time hierarchically.

→ **SG-TLM** (Su et al., 2023) masks syntactic spans related to time to force the model to learn implicit temporal cues.

→ **TSM** (Cole et al., 2023) and **Temporal Attention** (Rosin and Radinsky, 2022) incorporate temporal supervision into attention heads.

Temporal Grounding

→ Cao and Wang (2022) proposes both textual and continuous vector-based prompts to provide time-specific context for generative tasks.

→ **TCQA** (Son and Oh, 2023) uses synthetic datasets and span-selection to enforce consistency between answers and their historical/temporal context.

4.2 TEMPORAL RAG

To overcome the "static knowledge" limitation of TLMs and reduce temporal hallucinations, recent work turned to T-RAG

→ Integrates neural retrieval with generation to fetch up-to-date, time-relevant evidence at inference

→ **TempRetriever** (Abdallah et al., 2025) and **TsContriever** (Wu et al., 2024) encode queries and passages with timestamp-aware embeddings to improve dense retrieval.

→ **TempRALM** (Gade and Jetcheva, 2024) applies temporal constraints during the search phase to mitigate factual drift and prioritize recency.

→ **TimeR4** (Qian et al., 2024) utilizes a four-stage process (Retrieve-Rewrite-Retrieve-Rerank) to transform vague queries into time-anchored formulations.

→ **MRAG** (Siyue et al., 2024) enables multi-hop reasoning across events by retrieving from multiple time-scoped sources.

4.3 TEMPORAL REASONING CAPABILITIES

Retrieval alone is insufficient for tasks involving event sequences, durations, temporal arithmetic, and causality.

So, studies turned to improving the reasoning capacity of PLMs through **architectural design** and **specialized training objectives**.

→ **ECONET** (Han et al., 2021) focuses on continual adaptation to maintain consistency across temporal updates.
→ **TIMERS** (Mathur et al., 2021) & **ConTempo** (Niu et al., 2021) utilize structured reasoning layers, temporal graphs, or symbolic abstractions for multi-hop document-level inference.

Specialized Benchmarks
→ **TRAM** (Wang and Zhao, 2024) evaluates event frequency, duration estimation, and timeline ordering.
→ **Test of Time (ToT)** (Fatemi et al., 2024) isolates core logic and counterfactuals from simple memorization.
→ **TODAY** (Feng et al., 2023) **Narrative-of-Thought** (Zhang et al., 2024) probes reasoning over structured narratives and temporal shifts.

Even SOTA models (e.g., GPT-4) significantly underperform compared to human baselines in these structured tasks.

4.3 TEMPORAL REASONING CAPABILITIES

Paradigm	Core Method	Strength	Limitation
TLMs	Pretraining with timestamps	Time-aware representations	Static knowledge; lacks implicit cues
T-RAG	Inference-time retrieval	Real-time factuality	Evidence reliability; grounding issues
Reasoning Models	Symbolic/Graph layers	Multi-hop & Event ordering	Fragility to adversarial shifts

5. FUTURE RESEARCH DIRECTIONS

Dynamic temporal knowledge management: move beyond isolated fact updates toward scalable frameworks

Temporally-aware LLM agents: incorporate timeline tracking, event memory and temporal reference

Diachronic and synchronic knowledge integration: develop temporal alignment algorithms and cross-source reasoning frameworks

Temporal uncertainty and confidence modeling: handle uncertain dates and provide confidence scores for temporal answers

Multilingual and multimodal temporal QA: develop multilingual temporal taggers and cross-modal alignment techniques

Implicit temporal intent understanding: detect hidden temporal assumptions

Evaluation and benchmarking for temporal reasoning: develop temporally-aware evaluation metrics

6. CONCLUSION

- Evolution from rule-based systems to TLMs and RAG has improved time-sensitive retrieval, yet core reasoning remains a hurdle.
- Current models are still vulnerable to answer drift, temporal uncertainty, and a lack of cultural/linguistic diversity.
- Reliance on static snapshots prevents models from mastering real-time adaptation and future-oriented reasoning.
- Advancing the field necessitates a shift toward adaptive learning and representations that treat time as a fluid, continuous dimension.
- Closing the "reasoning gap" is essential for building trustworthy systems capable of navigating complex, evolving narratives.

Thank you for your Time!

Any (temporal) questions?

REFERENCES

Piryani, B., Abdallah, A., Mozafari, J., Anand, A., & Jatowt, A. (2025). It’s high time: A survey of temporal question answering. arXiv. https://arxiv.org/abs/2505.20243

Gruber, R., Abdallah, A., Färber, M., & Jatowt, A. (2025). COMPLEXTEMPQA: A 100m dataset for complex temporal question answering. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 9100–9112). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.463