SECTION 3 · 15 min

Temporal QA/IR Datasets

Challenging temporal datasets

Presenter: Adam Jatowt

03

Temporal QA Datasets

Dataset	#Q	Source	Method	Answer Type	Time Frame	Temp. Meta	Multi-Hop
NewsQA	**119k**	News	CS	Freeform	2007–2015	✗	✗
TDDiscourse	6.1k	News	CS	Extractive	Unspecified	✗	✗
TORQUE	21k	News	CS	Abstractive	–	✗	✗
ArchivalQA	**532k**	News	AG	Extractive	1987–2007	✓	✗
TimeQA	41.2k	Wikipedia	AG	Extractive	1367–2018	✗	✗
TiQ	10k	Wikipedia	AG	Freebase	Unspecified	✗	✗
TempQuestions	1.2k	Freebase	AG	Extractive	Unspecified	✗	✓
TemporalQuestions	1k	News	CS	Extractive	1987–2007	✓	✗
TempLAMA	50k	News	CS	Extractive	2010–2020	✓	✗
ComplexTempQA	**100.2M**	Wikipedia	AG	Extractive	1987–2023	✓	✓
MenatQA	2.8k	Wikipedia	AG	Extractive	1367–2018	✗	✗
PAT-Question	6.1k	Wikipedia	CS	Extractive	–	✗	✓
TempTabQA	11.4k	Wiki Infobox	CS	Abstractive	–	✗	✗
SituatedQA	12.2k	Wikipedia	CS	–	≤2021	✗	✗
UnSeenTimeQA	3.6k	Synthetic	AG	Abstractive	–	✗	✓
ChroniclingAmericaQA	**485k**	News	AG	Extractive	1800–1920	✓	✗
FRESHQA	600	Google	CS	–	–	✗	✓
COTEMPQA	4.7k	Wikidata	CS	Abstractive	≤2023	✗	✓
Test of Time (ToT)	1.8k	Synthetic	AG	Abstractive	–	✗	✓
TIMEDAIL	1.1k	DailyDialog	CS	Multi-choice	–	✗	✗
TEMPO	1.7k	StackExch.	CS	Abstractive	≤2025	✓	✓
Complex-TR	10.8k	Wiki+Google	AG	Multi-answer	≤2023	✗	✓
StreamingQA	**147k**	News	CS	Extractive	2007–2020	✓	✓
TRACIE	5.4k	Wikipedia	CS	Abstractive	≤2020	✗	✗
ForecastQA	10.3k	News	CS	Multi-choice	2015–2019	✓	✓
TEMPREASON	52.8k	Wiki/Wikidata	SC	Abstractive	634–2023	✗	✗
TemporalAlignmentQA	20k	Wikipedia	AG	Abstractive	2000–2023	✗	✗
ReaLTimeQA	5.1k	Search	CS	Multi-choice	2020–2024	✗	✗

#Q = number of questions · CS = Crowdsourced, AG = Automatically Generated, SC = Semi-automatic · ✓ / ✗ = present / absent · ≤ year = Wikipedia-snapshot scope

TIQ: A Benchmark for Temporal Question Answering with Implicit Time Constraints

Why is TIQ needed?

Existing datasets mainly contain, **Explicit temporal questions**

- Who was the US president in **2010**?

But many real questions contain **implicit temporal references**

- Which football club did Messi join **after Paris Saint-Germain**?

These require understanding the time implied by events rather than explicit dates.

Dataset Overview	
Size	**10,000 questions**
Focus	Implicit temporal constraints
Sources	Wikipedia, Wikidata & Infoboxes
Relations	Before • After • During
Evidence	Supporting evidence snippets

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).. https://doi.org/10.1145/3589335.3651895

TIQ: A Benchmark for Temporal Question Answering with Implicit Time Constraints

Table 4: Example questions from the TIQ benchmark, including the information snippets they are derived from.

Topic entity	`Clarence Andrew Cannon`
Question	*What was Clarence Andrew Cannon's occupation before becoming a lawyer?*
Answer	teacher
Information snippet - Main	*"Clarence Andrew Cannon, occupation, teacher, start time, 1904, end time, 1908"* (KB)
Information snippet - Constraint	*"Clarence Cannon, He earned an LL.B. and joined the bar in 1908."* (TEXT)
Topic entity	`Robert Bosch GmbH`
Question	*Who was the chief executive officer at Robert Bosch GmbH before revenue reached €78.74 billion?*
Answer	Volkmar Denner
Information snippet - Main	*"Robert Bosch GmbH, chief executive officer, Volkmar Denner, start time, 2012, end time, 2021"* (KB)
Information snippet - Constraint	*"Robert Bosch GmbH, Revenue, € 78.74 billion (2021)"* (INFOBOX)
Topic entity	`Carlos Alberto Torres`
Question	*Which national football team did Carlos Alberto Torres manage before joining Flamengo?*
Answer	Oman national football team
Information snippet - Main	*"Carlos Alberto Torres, Managerial career, 2000–2001, Oman"* (INFOBOX)
Information snippet - Constraint	*"Carlos Alberto Torres, Managerial career, 2001–2002, Flamengo"* (INFOBOX)
Topic entity	`Alan Page`
Question	*What hall of fame did Alan Page become a member of while serving as Associate Justice of the Minnesota Supreme Court?*
Answer	College Football Hall of Fame
Information snippet - Main	*"Alan Page, In 1993, he was inducted into the College Football Hall of Fame."* (TEXT)
Information snippet - Constraint	*"Alan Page, Associate Justice of the Minnesota Supreme Court, In office January 4, 1993 – August 31, 2015"* (INFOBOX)

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).. https://doi.org/10.1145/3589335.3651895

ArchivalQA: A Large-scale Benchmark Dataset for Question Answering over Historical News Collections

Goal

Enable question answering over historical news archives instead of Wikipedia.

Key Contributions

- 532K question–answer pairs (both implicit and explicit)
- Covers 1987–2007
- Built from New York Times archive
- Four subsets based on difficulty and temporal expressions
- Automatic question generation with ambiguity filtering

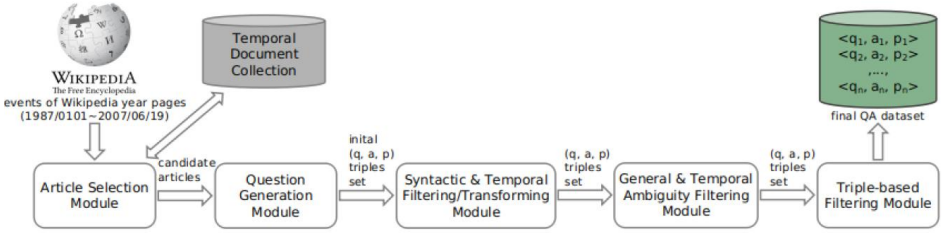
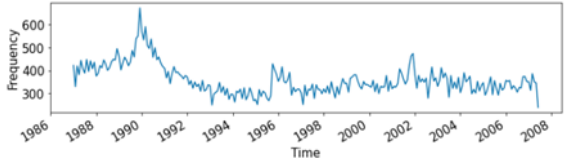


Figure 1: Dataset generation framework



id	question	answer	org_answer	answer_start	para_id	trans_que	trans_ans	source
train_0	Who claimed responsibility for the bombing of Bab Ezzouar?	Al Qaeda	Al Qaeda	184	1839755_20	0	0	wiki
train_4	When did Tenneco announce it was planning to sell its oil and gas operations?	May 26, 1988	today	103	148748_0	0	1	rand
val_45	What threat prompted Mr. Paik's family to flee to Hong Kong?	the Korean War	the Korean War	327	1736040_7	0	0	wiki
test_84	Along with the French Open, what other tournament did Haarhuis win in 1998?	Wimbledon	Wimbledon	527	1043631_15	1	0	rand

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). https://doi.org/10.1145/3477495.3531734

ArchivalQA: A Large-scale Benchmark Dataset for Question Answering over Historical News Collections

Temporally ambiguous questions: questions with several correct answers at different time points

- Examples:
  - (OK) *Who sent 20,000 American troops to Bosnia*? [**Clinton**]
  - (OK) *How old was Evelyn Sabin when she died*? [**90**]
  - (NG) ~~*Who was the Senator of West Virginia*? [**John D. Rockefeller**]~~
  - (NG) ~~*How many points does Ashley McElhiney have*? [**5**]~~
  - (NG) ~~*What country is the current U.S. policy*? [**Iran**]~~

- Filtering setup:
  - Dataset: 5,500 manually annotated questions
  - Classifier: BERT-base with an accuracy of 81.82%

TempAmbiQA: Detecting **Temporal Ambiguity** in Questions

Motivation

Many questions cannot be answered correctly without a temporal reference.

Example

- *Who was the president of NBC Universal?*
- Correct answer depends on when the question refers to.

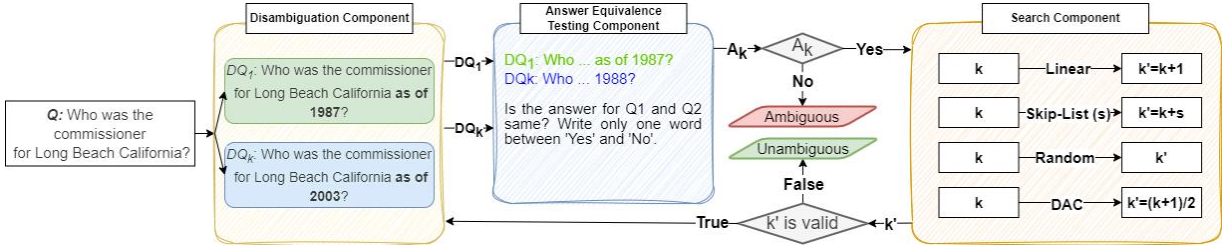


Figure: *Overview of different search strategies for detecting temporally ambiguous Questions. The Disambiguation Component generates questions DQ1 and DQk, referred to as Q1 and Q2 in the prompts, respectively. The Answer Equivalence Testing Component compares them, classifying Q as temporally ambiguous if the answer equivalence (Ak) is "No". If "Yes", the search proceeds to find the next valid year k' within the defined time range, generating the next disambiguation question DQk' to continue the classification process. If no valid k' is found, the question Q is classified as temporally unambiguous. A valid year k' is the one that falls within the specified time range (e.g., 2000-2024).*

Key Contributions

- First benchmark for temporal ambiguity detection (TempAmbiQA)
- 8,162 manually annotated questions: 3,879 ambiguous 4,283 unambiguous
- Built from ArchivalQA, SituatedQA, and AmbigQA
- Evaluates whether a question requires temporal clarification before answering

Piryani, Bhawna, et al. "**Detecting temporal ambiguity in questions.**" Findings of the Association for Computational Linguistics: EMNLP 2024. 2024.

ComplexTempQA: A Large-scale Benchmark for Complex Temporal Question Answering

Key idea

- 100M+ temporal question–answer pairs
- Built from Wikipedia + Wikidata
- Covers 1987–2023
- Each question has temporal metadata
- Designed for training and evaluating LLMs
- Focuses on complex temporal reasoning

Supports multiple reasoning skills

- ✓ Temporal comparison
- ✓ Counting over time
- ✓ Multi-hop reasoning
- ✓ Event ordering
- ✓ Cross-time reasoning
- ✓ Implicit event descriptions

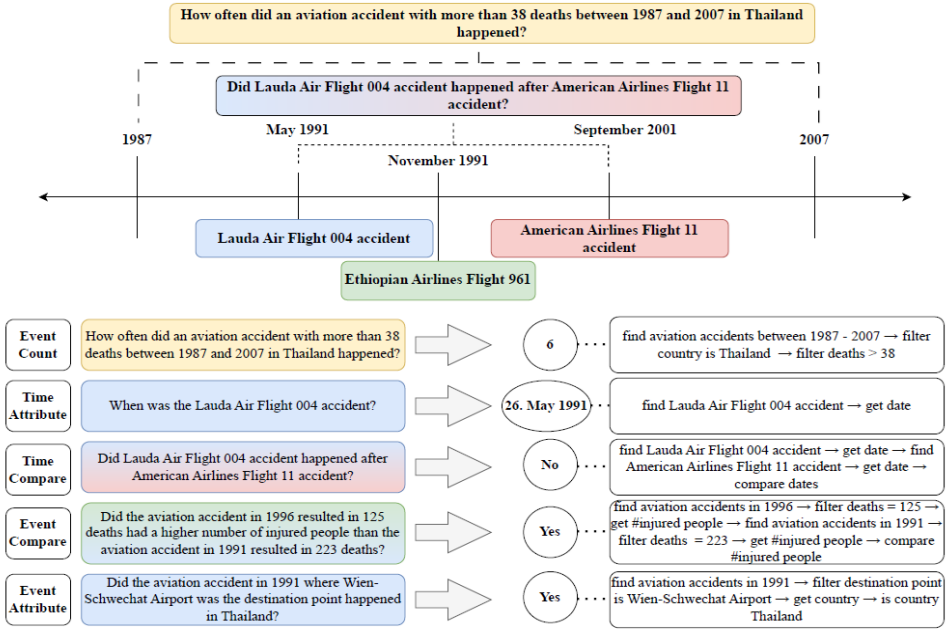
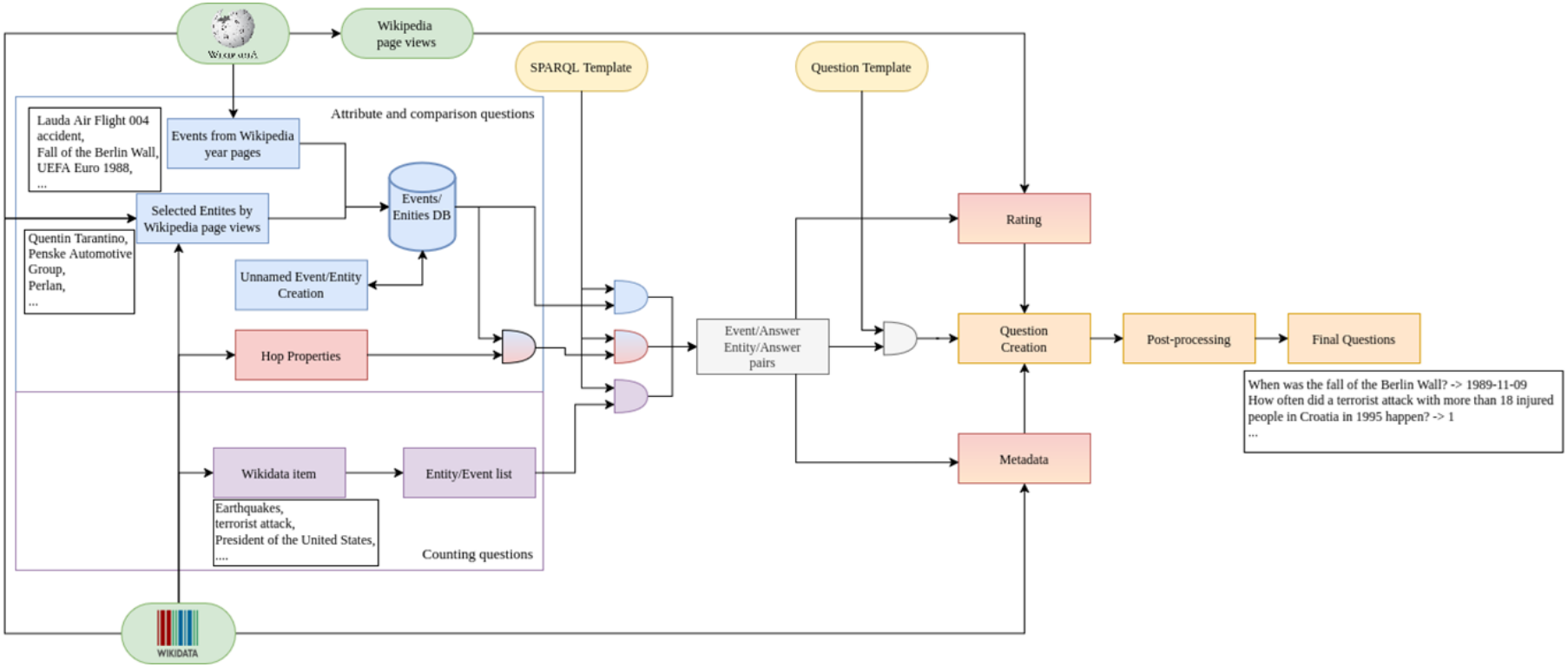


Figure 1: Example types of temporal reasoning in questions (left) with the required inference steps (right) and timeline-based visualization (above).

Gruber, Raphael, et al. **"COMPLEXTEMPQA: A 100m Dataset for Complex Temporal Question Answering."** Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025.

ComplexTempQA: A Large-scale Benchmark for Complex Temporal Question Answering



Gruber, Raphael, et al. **"COMPLEXTEMPQA: A 100m Dataset for Complex Temporal Question Answering."** Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025.

ComplexTempQA: A Large-scale Benchmark for Complex Temporal Question Answering

Method	Parameters	Precision	Recall	F1
Zephyr	7B	3.76	33.50	4.90
Falcon	7B	0.31	**34.22**	0.62
Llama-chat 7B	7B	3.68	33.94	6.09
Mistral	7B	3.73	**48.34**	6.33
LLama-chat 13B	13B	3.61	32.77	6.05
Vicuna	33B	2.02	37.27	3.63
Mixtral	8x7B	3.34	40.41	5.65
LLama-chat 70B	70B	**5.19**	39.30	**8.31**
Wizardlm	70b	1.63	27.63	2.80
GPT-3.5	175B	4.71	34.46	7.74

Zero-shot performance

Method	Parameters	Shots	Precision	Recall	F1
Llama-chat	7B	0	3.675	**33.935**	6.09
		1	7.08	27.21	9.00
		2	23.05	30.65	23.655
		3	**23.67**	30.38	**24.22**
Llama-chat	13B	0	3.605	**32.76**	6.05
		1	22.865	27.91	23.345
		2	**31.645**	32.35	**31.56**
		3	30.37	32.03	30.495
Llama-chat	70B	0	5.191	39.30	8.31
		1	25.74	32.61	18.8
		2	34.01	44.21	35.26
		3	**37.09**	**46.57**	**38.44**
Mistral-Instruct	7B	0	3.73	**48.33**	6.325
		1	24.68	35.33	25.87
		2	32.74	35.91	33.145
		3	**35.28**	37.91	**35.68**
Mixtral	8x7B	0	3.34	40.41	5.65
		1	6.76	39.06	9.17
		2	14.27	41.89	16.44
		3	**15.49**	**44.03**	**17.83**

Few-shot performance

Method	Parameters	Context	Precision	Recall	F1
Llama-chat	7B	No Context	3.67	33.93	6.09
		Retriever	3.59	33.67	5.97
		True Context	**3.92**	**37.40**	**6.49**
Llama-chat	13B	No Context	3.60	32.76	6.05
		Retriever	3.50	34.22	5.84
		True Context	**3.75**	**37.09**	**6.28**
Llama-chat	70B	No Context	5.19	**39.30**	8.31
		Retriever	5.27	36.16	8.12
		True Context	**5.82**	38.59	**8.82**
Mistral-Instruct	7B	No Context	3.73	**48.33**	6.32
		Retriever	3.86	33.32	6.31
		True Context	**5.13**	35.26	**8.08**
Mixtral	8x7B	No Context	3.34	**40.41**	5.65
		Retriever	**4.23**	35.93	**6.62**
		True Context	3.65	38.02	5.88

RAG performance

Method	Parameters	Attribute-type				Comparison-type				Counting-type			
		Precision	Recall	F1	Con	Precision	Recall	F1	Con	Precision	Recall	F1	Con
Llama-chat	7B	6.55	60.70	9.52	96.86	5.30	58.05	8.83	63.20	0.34	10.50	0.66	80.24
Llama-chat	13B	6.17	61.26	10.67	96			8.75	62			0.30	86.13
Llama-chat	70B	**8.21**	65.85	**13.71**	97			11.25	64			0.84	48.99
Mistral-Instruct	7B	5.87	54.00	9.96	96.52	**8.33**	57.79	**12.73**	62.18	**0.68**	20.80	**1.31**	87.45
Mixtral	8*7B	4.04	**75.80**	7.21	97.25	5.18	**63.50**	8.86	72.10	0.60	**27.44**	1.17	**90.10**

Performance of different question types

Gruber, Raphael, et al. **"COMPLEXTEMPQA: A 100m Dataset for Complex Temporal Question Answering."** Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025.

ChroniclingAmericaQA: A Large-scale Question Answering Dataset based on Historical American Newspaper Pages

Goal

Extend historical QA to digitized newspapers with OCR errors.

Key Contributions

- 487K question–answer pairs
- 1800–1920 (120 years)
- Newspapers from 53 U.S. states
- Supports:
  - OCR text
  - Corrected OCR text
  - Newspaper images
- Evaluates QA under noisy OCR conditions

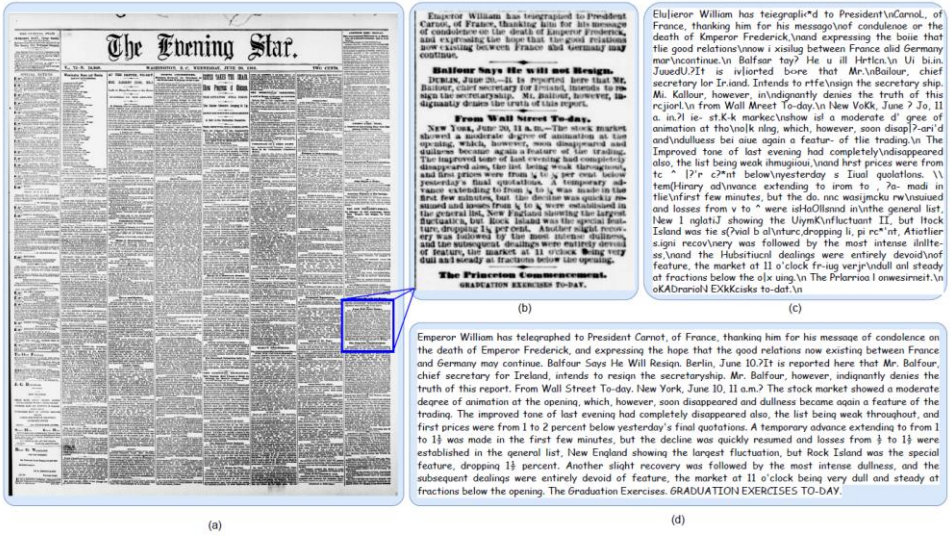


Figure 1: An example of the scanned newspaper page from Chronicling America Collection. a) depicts the entire newspaper page published in Evening Star on 1803-02-07 in the District of Columbia, b) depicts the zoomed-in paragraph of the newspaper page shown in (a), c) shows the original OCR text of the zoomed-in paragraph that is available in the Chronicling America, and d) displays the OCR text corrected by GPT 3.5 Turbo.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. **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). https://doi.org/10.1145/3626772.3657891

RecencyQA: Estimating **Recency Requirements** in Question Answering

Motivation

Existing temporal QA datasets ask*: Is the answer current?*
But they do not ask: *How often does the answer change?*

Key Contributions

- **RecencyQA,** first benchmark to model recency requirements
- Introduces Recency–Stationarity Taxonomy
- 4,031 open-domain questions
- Context-aware temporal evaluation
- Supports recency-aware QA and retrieval

Questions differ not only in whether their answers change, but also in **how frequently they change** and whether that **behavior depends on context**.

Recency Class	Expected Time Until Answer Change
An-Hour	Within an hour
A-Few-Hours	Within a few hours
A-Day	Within a day
A-Few-Days	Within a few days
A-Week	Within a week
A-Few-Weeks	Within a few weeks
A-Month	Within a month
A-Few-Months	Within a few months
A-Year	Within a year
A-Few-Years	Within a few years
Many-Years	After many years
Never	Not expected to change

Table 2: The proposed Recency Taxonomy consists of 12 classes, ordered from highly volatile (top) to temporally stable (bottom). Each class reflects the expected time until a question's answer first changes.

Piryani, B., Mert, Z., & Jatowt, A. (2026). **How often do Answers Change? Estimating Recency Requirements in Question Answering**. arXiv preprint arXiv:2603.16544.

RecencyQA: Estimating **Recency Requirements** in Question Answering

Question	Majority Recency	Stationarity	Representative Context
Who has the most liked post on Instagram?	A-Day	Non-Stationary	"As the Instagram Awards ceremony is about to announce the winner of the most liked post of the year..."
Who is the starting running back for the San Francisco 49ers?	A-Week	Non-Stationary	"With training camp wrapping up, the team's final roster cuts are imminent..."
Who is the richest man on earth?	A-Few-Months	Non-Stationary	"As the annual Forbes billionaire list is about to be released..."
What was the shortest war in history?	Never	Stationary	"During a trivia night at a local pub, a group of friends wondered about the shortest war in history."
Where is the deepest place on Earth?	Never	Stationary	"During a geography lecture, a student asks about the deepest place on Earth."
What is Geoff Hintons h-index?	A-Year	Non-Stationary	"During a panel discussion on the impact of deep learning at a major tech conference..."

Table 8: Representative examples from RecencyQA illustrating diverse recency and stationarity behaviors.

Question	RL_1	RL_2
Who has the most liked post on Instagram?	A-Few-Hours	A-Day
Which country is currently at the top rank at the FIBA Men's World Ranking?	A-Few-Months	A-Few-Years
How many asteroids have been discovered before impacting Earth?	A-Day	A-Few-Years
What was the shortest war in history?	Many-Years	Never
The town that Sigmund Freud was born in is currently called by what name?	Many-Years	Never

Table 9: Examples of recency label transitions induced by contextual variation.

Piryani, B., Mert, Z., & Jatowt, A. (2026). **How often do Answers Change? Estimating Recency Requirements in Question Answering**. arXiv preprint arXiv:2603.16544.

TRAM: Benchmarking Temporal Reasoning for Large Language Models

Key Contributions

- Comprehensive benchmark for temporal reasoning
- 10 tasks spanning multiple reasoning abilities
- 38 subtasks
- 526K+ multiple-choice questions
- Evaluates both traditional models and LLMs

TRAM extends evaluation beyond temporal QA to a broad range of temporal reasoning skills.

Frequency (Commonsense)	**Q:** It is also a love story , between Ace and Tobio, a trans woman. How often do they break up? A. Once ✓ B. Always ✗ C. Once per week ✗
Ambiguity Resolution (Interpretation)	**Q:** A historic event is documented to have happened 'before you know it'. When did it take place? A. The next day ✗ B. Without hesitation ✗ C. Before long ✓
Temporal Causality (Cause)	**Q:** She noticed that all the wall clocks in the store were set to ten past ten. What's the more plausible CAUSE? A. It is a common display setting for clocks and watches. ✓ B. B. It was ten minutes past ten at that moment. ✗
Temporal Storytelling	**Q:** I woke up so late this morning. I was panicked when I saw what time it was. I had to be at work on time. I threw myself together quickly. Which of the two endings is the most plausible correct ending to the story? A. I was able to get a job at a local restaurant. ✗ B. I was still thirty minutes late. ✓
Arithmetic (24-hour Adjustment)	**Q:** What is 00:18 - 23:50? A. 0:28 ✓ B. 1:44 ✗ C. 22:15 ✗ D. 1:35 ✗

Figure 1: Example questions in TRAM.

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.

MenatQA: A New Dataset for Testing the Temporal Comprehension and Reasoning Abilities of Large Language Models

Motivation

Existing temporal QA benchmarks mainly evaluate **fact retrieval** or **simple temporal reasoning.**

MenatQA evaluates **three complementary temporal reasoning abilities:**

- Scope
- Order
- Counterfactual reasoning

Key Contributions

- First benchmark to jointly evaluate **Scope, Order**, and **Counterfactual** reasoning
- **2,853** temporal reasoning questions
- Includes **answerable** and **unanswerable** questions
- Built from **Wikidata temporal knowledge**
- Designed to evaluate **temporal reasoning in LLMs**

Context: [1] *Twitter was created by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams in March 2006 and launched in July of that year.* [2] *On October 16, 2008, Evan Williams became the CEO, and Dorsey became the chairman of the company.* [3] *Jack Dorsey rejoined Twitter in March 2011 as Executive Chief of Product Development.* [4] *In June 2020, Twitter announced that Patrick Pichette would succeed Omid Kordestani as chairman.* [5] *In November 2021, Jack Dorsey stepped down as CEO and was replaced by Parag Agrawal, the chief technology officer.* [6] *On October 27, 2022, business magnate Elon Musk acquired Twitter for US$44 billion, gaining control of the platform.* [7] *On May 12, 2023, Musk announced that he will resign as CEO of Twitter in approximately six weeks.*

Multiple Sensitive Factors Time QA

Scope Factor:
Who was the CEO of Twitter from May 2013 to 2020 ?
Order Factor:
Shuffle the order of the sentences [1-7] in the context .
Counterfactual Factor:
Who was the CEO of Twitter from March 2011 to July 2022 , if Jack Dorsey stepped down as CEO in November 2022 ?

Answer: *Jack Dorsey*

Figure 1: Yellow front indicates the time specifiers of events in the context. Scope Factor refers to the time specifiers would be different between the question and the given context. Order Factor is where the complete events in the context are shuffled in chronological order. Counterfactual Factor is a question with hypothetical propositions.

MenatQA: A New Dataset for Testing the Temporal Comprehension and Reasoning Abilities of Large Language Models (https://aclanthology.org/2023.findings-emnlp.100/) (Wei et al., Findings 2023)

NTCIR Temporalia 1/2

Temporal Query Intent Classification (TQIC)

Query class	Query example
Past	price hike in bangladesh 2008
Past	Who Was Martin Luther
Past	when did the titanic sink
Past	Yuri Gagarin Cause of Death
Past	History of Coca-Cola
Recency	apple stock price
Recency	Number of Millionaires in USA
Recency	time in london
Recency	Trendy Plus Size Clothing
Recency	Did the Pirates Win Today
Future	2013 MLB Playoff Schedule
Future	release date for ios7
Future	College Baseball Regional Projections
Future	disney prices 2014
Future	long term weather forecast
Atemporal	blood pressure monitor
Atemporal	distance from earth to sun
Atemporal	how to start a conversation
Atemporal	New York Times
Atemporal	lose weight quickly

Temporal Information Retrieval (TIR)

	Girl with the Dragon Tattoo
Description	I've recently watched a film called Girl with the Dragon Tattoo, and really liked it. Therefore, I would like to gather information about the movie.
Past question	How did the casting of the film develop?
Recency question	What did the recent reviews say about the film?
Future question	Is there any plan about its sequel?
Atemporal question	What are the names of main actors and actresses of the film?
Search date	28 Feb 2013 GMT+0:00

Temporally Diversified Retrieval (TDR)

	Junk food health effect
Description	I am concerned about the health effects of junk food in general. I need to know more about their ingredients, impact on health, history, current scientific discoveries and any prognoses.
Past question	When did junk foods become popular?
Recency question	What are the latest studies on the effect of junk foods on our health?
Future question	Will junk food continue to be popular in the future?
Atemporal question	How junk foods are defined?
Search date	29 May 2013 GMT+0:00

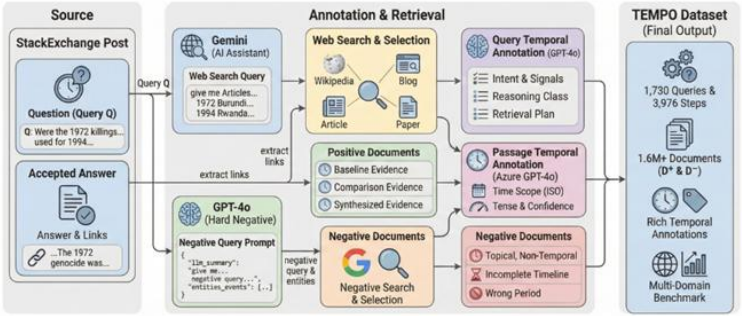
Temporal Intent Disambiguation (TID)

Query	Past	Recency	Future	Atem.
Australian Open	0.091	0.0	0.455	0.455
motorcycle accident june	0.7	0.0	0.3	0.0
NBA Finals	0.1	0.0	0.4	0.5
NBA playoff schedule	0.0	0.2	0.6	0.2
price of oil	0.0	0.9	0.0	0.1
how to lose weight	0.0	0.1	0.0	0.9
time in India	0.0	1.0	0.0	0.0
history of volleyball	1.0	0.0	0.0	0.0

https://ntcirtemporalia.github.io/

TEMPO: A Realistic Multi-Domain Benchmark for Temporal Reasoning-Intensive Retrieval

1.7k complex temporal questions from Stack Exchange with reasoning steps and positive/negative documents for reasoning-intensive retrieval from 1.6 m passages



Query

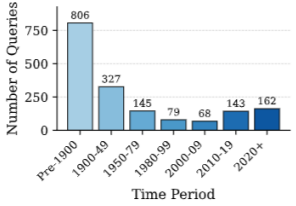
How accurate was the measurement of the period of Earth's orbit in the 19th Century?
There was a section on my textbook on history of theories of sun's energy source.
It talks about how the Meteorite Theory was dismissed, as it would decrease the period of Earth's orbit by 2 seconds per year due to increased mass of the Sun.
This theory was dismissed due to disagreeing with observation. And the textbook says the change in period is "easily measurable"
My question is how is 2 seconds difference easily measurable in Nineteenth Century?

Query

Was the Tsar's property separated from state property in Russian Empire in early 19th century (regarding land)?

I mean, were there Tsar's serfs who were not state serfs?

Dataset	Total Number Q	Total Number $\mathcal{D}$	Total Number $\mathcal{D}^+$	Avg. Length Q	Avg. Length $\mathcal{D}$	Avg. Steps
Blockchain						
Bitcoin	100	153,291	3.3	222.0	596.9	2.93
Cardano	51	87,201	2.5	161.1	647.2	2.84
Iota	10	10,372	3.8	148.6	1,036.5	3.20
Monero	65	85,093	2.6	171.8	703.3	2.72
Social Sciences						
Economics	83	93,756	3.6	290.2	495.9	3.08
Law	35	43,288	3.0	258.5	500.7	3.23
Politics	150	183,394	2.7	343.2	476.3	3.35
History	801	356,493	4.5	374.2	682.7	3.42
Applied						
Quant	34	28,785	2.4	422.5	477.1	2.68
Travel	100	177,677	2.6	264.5	374.8	3.11
Workplace	36	64,659	2.8	291.8	368.7	2.42
Genealogy	115	156,228	2.8	359.6	629.2	3.78
STEM						
HSM	150	213,818	2.5	303.5	563.6	3.25
Total	**1,730**	**1,654,055**	–	–	–	–



Abdallah, Abdelrahman, Mohammed Ali, Muhammad Abdul-Mageed, and Adam Jatowt. **"TEMPO: A Realistic Multi-Domain Benchmark for Temporal Reasoning-Intensive Retrieval."** arXiv preprint arXiv:2601.09523 (2026).

RETECO: SemEval 2027 Task on Reasoning-oriented Retrieval with Temporal & Conversational Context