scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A Question Answering Dataset for Temporal-Sensitive Retrieval-Augmented Generation

Ziyang Chen[1], Erxue Min[2], Xiang Zhao[1]✉, Yunxin Li[3], Xin Jia[2], Jinzhi Liao[1], Jichao Li[1], Shuaiqiang Wang[2], Baotian Hu[3] & Dawei Yin[2]

We introduce ChronoQA, a benchmark dataset for Chinese question answering focused on evaluating temporal reasoning in Retrieval-Augmented Generation (RAG) systems. Built from over 300,000 news articles published between 2019 and 2024, ChronoQA contains 5,176 questions covering absolute, aggregate, and relative temporal types, with both explicit and implicit time expressions. The dataset features both single- and multi-document scenarios, reflecting real-world requirements for temporal alignment and logical consistency. By providing structured evaluation across a wide range of temporal tasks, ChronoQA offers a dynamic, reliable, and scalable resource for benchmarking RAG systems in evolving knowledge environments.

Background & Summary

Large Language Models (LLMs), such as GPT-4[1], LLaMA[2], and GLM[3], have demonstrated remarkable capabilities across a broad spectrum of natural language understanding and generation tasks. However, LLMs remain inherently static, with their knowledge fixed at the time of training[4]. As the real-world evolves rapidly, there is an increasing demand for models that can process dynamic information[5–7]. Retrieval-Augmented Generation (RAG) has emerged as a solution, enabling LLMs to retrieve relevant documents from external sources to enhance response accuracy[8–10].

Despite its effectiveness in static knowledge retrieval, existing RAG systems face significant challenges when dealing with time-sensitive queries[11,12]. Their reliance on semantic matching often leads to retrieving outdated or irrelevant documents, failing to align properly with the temporal constraints embedded in user questions-such as implicit or relative time expressions. As a result, generating temporally coherent and accurate answers remains a major hurdle. Recently, the challenge of integrating temporal reasoning into RAG systems has attracted significant attention[11,13]. Numerous applications in finance, public policy, news analysis, and even scientific research demand accurate reasoning over evolving events. However, current evaluation efforts do not adequately reflect this need.

RAG datasets play a crucial role in evaluating retrieval-augmented methods, yet most existing benchmarks focus on static knowledge retrieval, lacking a systematic approach to temporal reasoning. Early QA datasets, such as Natural Questions (NQ)[14], TriviaQA[15], and MS MARCO[16], primarily assess open-domain retrieval, relying on web documents and knowledge graphs. More advanced RAG benchmarks like HotpotQA[17] introduce multi-hop retrieval, requiring models to synthesize information across multiple sources. However, these datasets assume static knowledge and overlook scenarios where answers evolve over time, a critical limitation for time-sensitive applications. Recent efforts have attempted to incorporate temporal awareness into RAG evaluation. For example, FreshQA[18] evaluates whether models retrieve the most temporally relevant evidence, while CRAG[19] and DomainRAG[20] introduce mechanisms for handling document updates over time. Nevertheless, these datasets are limited in scope: they typically support only direct temporal logic, lack diversity in question types (e.g., aggregate or implicit time expressions), and rarely require multi-document reasoning. Moreover, none offer a scalable or automated mechanism for dataset evolution. As shown in Table 1, existing datasets suffer from low temporal relevance coverage, limited reasoning complexity, and insufficient support for

[1]National Key Laboratory of Big Data and Decision, National University of Defense Technology, Changsha, China. [2]Baidu Inc., Beijing, China. [3]Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China. ✉e-mail: xiangzhao@nudt.edu.cn

Dataset	QA Scale	QA Source	Corpus Source	Corpus Scale	Temporal%	Temporal Logic	Multi-Doc
Natural Questions[14]	323K	Human	Web Search	15.8M	/	/	✗
TriviaQA[15]	95K	Human	Web Search	66K	/	/	✗
MIRAGE[24]	7,663	Human	Examination/Literature	65.3M	/	/	✗
FRESHQA[18]	600	Human	Web Search	/	/	Direct	✗
CRAG[19]	4,409	Human	Mock KG/Web Search	220K	/	Direct	✓
DomainRAG[20]	395	LLM	Admission Website	14K	16.4%	Direct	✓
MultiHop-RAG[25]	2,556	LLM	News Cropus	609	22.8%	Relative	✓
ChronoQA	5,176	LLM	News Cropus/Web Search	300K	100%	Multiple	✓

Table 1. Comparison of current RAG datasets.

multi-document contexts. This gap highlights the need for a benchmark that truly reflects the temporal dynamics of real-world QA tasks.

To bridge this gap, we present ChronoQA-a large-scale and systematically constructed dataset tailored for evaluating temporal-sensitive RAG systems. ChronoQA sets itself apart from prior work through several key innovations. First, it achieves 100% temporal relevance: every question requires temporal reasoning, encompassing both explicit and implicit time expressions and covering absolute, aggregate, and relative temporal types. Second, ChronoQA supports both single- and multi-document scenarios, mirroring real-world demands for temporal alignment and logical consistency across sources. Built from over 300,000 news articles published between 2019 and 2024, the dataset comprises 5,176 high-quality question-answer pairs, generated via a robust multi-stage pipeline that integrates LLM-based extraction, structured question synthesis, and rigorous validation. The dataset incorporates circuit-style question compositions-parallel and series reasoning circuits-to represent multi-step inference, cross-document alignment, and temporal dependency resolution. ChronoQA provides structured metadata and a temporal QA classification scheme, enabling detailed analysis of model performance across different temporal reasoning categories. The automated construction framework is designed for scalability, reproducibility, and updatability, supporting the dataset's adaptation to evolving knowledge.

ChronoQA addresses the limitations of existing RAG datasets by providing a resource with comprehensive temporal coverage and diverse reasoning requirements. The dataset is intended to support the evaluation and development of models for time-sensitive question answering and retrieval-augmented generation. In summary, the main contributions of this paper are as follows:

- This work defines the task of temporal-sensitive retrieval-augmented question answering, which requires models to retrieve and reason over temporally relevant evidence from dynamic corpora, handling both explicit and implicit temporal expressions.
- We introduce ChronoQA, a large-scale Chinese benchmark for this task, systematically covering diverse temporal reasoning types and supporting both single- and multi-document inference. The dataset is constructed through an automated pipeline that leverages LLMs for information extraction, question synthesis, and multi-document reasoning composition, enabling continuous update and scalability.
- ChronoQA includes comprehensive structural annotations-such as temporal type, scope, expression, answer type, and document reference-and has undergone multi-stage validation, including rule-based, LLM-based, and human evaluation, to ensure data quality and facilitate fine-grained model assessment.

Methods

In this section, we detail the construction process of the ChronoQA dataset. As illustrated in Fig. 1, the dataset is developed in three major steps: source article preparation, temporal question generation and verification.

Source Article Preparation. To construct a dataset reflecting real-world temporal dynamics, we required a source rich in evolving information and explicit time references. Publicly available news articles serve as an ideal basis due to their frequent updates and inherent temporal grounding. We began with a large corpus of text originating from diverse public news sources (e.g., Sina News), covering the period from January 1, 2019, to August 30, 2024. This initial collection represented a substantial volume of text, averaging content equivalent to approximately 171.8 articles per day, resulting in roughly 350k textual units (refer to Fig. 2 for yearly distribution).

Recognizing that raw news text contains noise and stylistic elements not conducive to direct question generation, and to ensure focus on factual content, we implemented a crucial processing step. Instead of using the raw article text directly, we systematically processed this initial corpus using `gpt-4o-mini`. The objective was to extract objective factual assertions, key entities, and associated temporal information (dates, times, durations, sequences) contained within the original texts. This LLM-driven extraction distilled the core temporal and factual essence of each news report into concise, structured summaries. Prior to LLM processing, standard text cleaning and deduplication were applied to the initial corpus to enhance data quality and remove redundancy. The output of the LLM extraction process yielded what we term "intensive temporal paragraphs" - focused textual units capturing verifiable facts and their temporal context. In total, 294,696 such distinct factual paragraphs were generated. Both the original news articles and these processed factual paragraphs are made publicly available in our repository (see the Data Records section for details). These derived paragraphs, rather than the

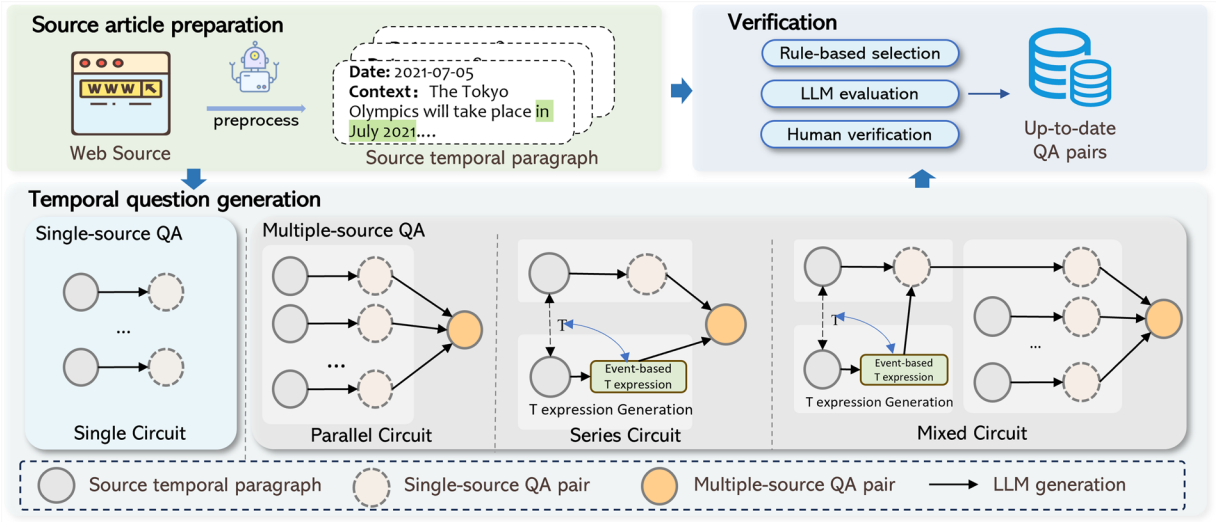

Fig. 1 Overview of construction process of the ChronoQA dataset.

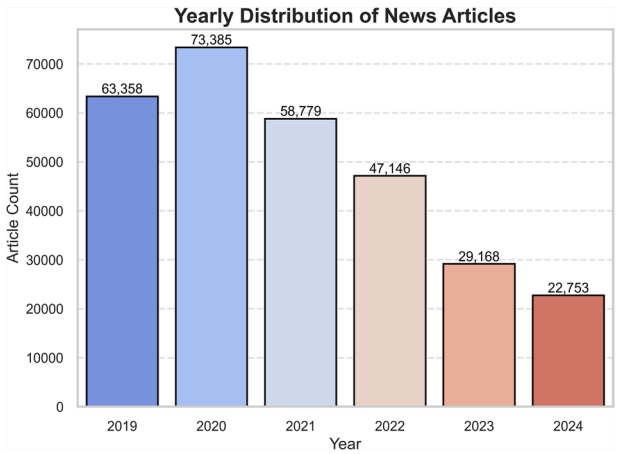

Fig. 2 Yearly distribution of collected articles from 2019 to 2024.

original full articles, formed the high-quality, manageable, and fact-centric foundation for the subsequent temporal question generation stages. This approach ensures that ChronoQA is built upon verifiable factual information extracted from real-world temporal narratives.

Single Temporal QA Generation. Building on prior work[21,22], we leveraged `gpt-4o` to systematically generate temporal question-answer pairs from the processed source texts. To ensure the generation of high-quality and diverse questions, we developed a detailed, structured prompt, the template for which is shown in Fig. 3. This base template was programmatically adapted with different `{target_qa_type}` instructions to generate a diverse range of questions, from explicit time-based queries (e.g., "When did *Event X* occur?") to more complex implicit ones (e.g., "What event preceded *Event Y*?").

For each source paragraph, the LLM was prompted to generate multiple temporal QA pairs. After generation, all pairs underwent an automated filtering process to remove duplicates and ensure uniqueness. This structured approach resulted in an initial high-quality repository of over 10,000 standalone temporal QA pairs, which formed the basis for the subsequent composition and validation stages.

Multiple Temporal QA Composition. To evaluate a model's ability to reason across multiple documents, we developed a systematic process to compose complex, multi-document questions from the pool of single-document QA pairs. This process involves two main stages: (1) identifying suitable candidate pairs for composition, and (2) merging them into coherent reasoning circuits. As illustrated in Fig. 4, we define two primary composition patterns: **parallel circuits** and **series circuits**, each designed to test a distinct aspect of multi-document reasoning. In a parallel circuit, sub-questions are logically independent but collectively required to answer the main question. Each sub-question contributes unique information, and all must be resolved to produce a complete answer. In a series circuit, sub-questions are interdependent, forming a sequential reasoning chain where the answer to one sub-question serves as input or context for the next.

Prompt for Single Time-Sensitive QA Generation

 # Role
You are an expert data annotator. Your role is to analyze historical news snippets and generate high-quality, time-sensitive question-answer pairs based on a structured reasoning process.

 # Core Reasoning Steps

1. Deconstruct the provided news snippets to identify all key entities, events, and their specific attributes.
2. Extract all explicit and implicit temporal expressions to establish a clear chronological timeline of the events.
3. Synthesize the extracted facts and temporal relationships into a challenging, natural-language question that adheres to all constraints below.

 # Hard Constraints
Principles (✓)

- The answer must be explicitly stated in or logically derivable from the source text alone.
- The question must be self-contained, fully understandable without the source text, and emulate a genuine user query.
- The answer must be a specific, concise entity (e.g., a name, date, or number) with no ambiguity.

Strictly Prohibited (✗)

- Do not provide answers like 'unclear' or 'unknown'.
- The question must not require the source text to be understood.
- The answer must not require information from outside the provided snippets.

 # Input
Current Time:
{current_date}
News Snippets:
{news_snippets}

Target Question Type:
{target_qa_type}

 # Output Format
You should follow the following format:
{output_format}

Fig. 3 The prompt template guiding the LLM for single temporal QA generation.

Candidate Pair Selection. Before composition, we first identify promising single-document QA pairs that are suitable for merging. Our automated selection strategy is based on two key criteria:

- **Semantic Similarity:** We compute vector representations for all questions and their corresponding source paragraphs using the `bge-large-zh-1.5` embedding model. Using cosine similarity, we then identify pairs that are thematically related (e.g., both discuss financial indices) or involve overlapping entities, even if they originate from different documents.
- **Temporal Proximity:** We extract and normalize the key timestamps from each QA pair's context. Pairs whose events occur within a narrow time frame (e.g., the same day or year) are flagged as strong candidates for composition, as they often relate to the same overarching event.

This filtering process yields a high-quality candidate pool, enabling the efficient construction of logically sound multi-document questions.

Parallel Circuit. A parallel circuit question requires aggregating information from multiple, logically independent facts to form a comprehensive answer. From the candidate pool, we select two or more QA pairs that are semantically related but do not depend on each other (e.g., performance of different stock indices on the same day). We use a tailored prompt to instruct `o1-mini` to merge the independent questions into a single, natural-sounding query. We guides the model to create a question that necessitates all pieces of information for a complete answer. The ground-truth answers from the source QA pairs are concatenated or summarized to form the final answer.

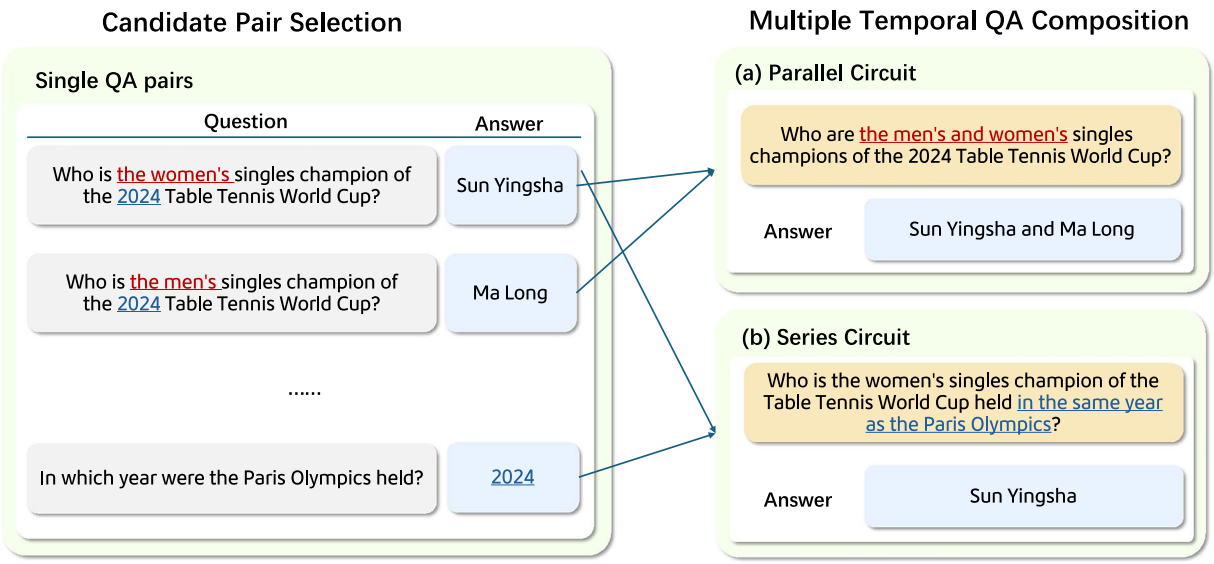

Fig. 4 The composition process for multi-document reasoning circuits. (**a**) In a Parallel Circuit, two independent QA pairs from different documents are synthesized into a single query that requires aggregating both pieces of information. (**b**) In a Series Circuit, the answer from one QA pair (e.g., a date) serves as a necessary input to resolve the second question, forming a dependency chain.

Series Circuit. A series circuit question requires sequential reasoning, where the answer to one sub-question is a prerequisite for answering the next. This creates a dependency chain across documents. We identify candidate pairs where the answer of one QA pair (the "source," e.g., a specific date, person, or location) is mentioned in the question context of another QA pair (the "target"). This answer acts as the logical bridge. We prompt `o1-mini` to reformulate the target question by replacing the explicit mention of the "bridge" entity with a descriptive clause from the source question. This forces a two-step reasoning process. The ground-truth answer of the final, target question is retained as the answer for the newly composed series question.

By employing these strategies, we create a diverse and challenging set of multi-document QA pairs. These questions test a model's ability to aggregate independent information, perform sequential reasoning, and handle hybrid reasoning tasks. This diversity ensures the dataset serves as a robust benchmark for evaluating temporal multi-document reasoning.

Dataset Quality Verification. To ensure the quality of the ChronoQA dataset, we implemented a multi-step verification pipeline combining rule-based filtering, LLM evaluations, and manual verification. Rule-based filtering validated structural and logical consistency, such as ensuring multi-document questions referenced at least two documents. LLM evaluations assessed fluency, temporal relevance, and semantic coherence, filtering out poorly constructed or inconsistent QA pairs. Finally, manual evaluation of ~6000 samples confirmed that over 95% met quality standards, validating the pipeline's effectiveness. This rigorous process ensures ChronoQA is a comprehensive and reliable dataset for evaluating and benchmarking models on time-sensitive tasks.

Dataset Statistics. Figure 5 presents representative examples from ChronoQA, highlighting the diversity of temporal expressions and reasoning types. As summarized in Table 2, the dataset contains 5,176 question-answer pairs spanning three temporal types-absolute (2,529), aggregate (1,911), and relative (736)-and two time expression categories: explicit (2,000) and implicit (3,176). Notably, 37% of the questions (1,915) require multi-document reasoning, offering deeper evaluation capabilities compared to existing benchmarks that mostly focus on single-document settings. The dataset further covers a range of answer types-entity (2,556), time (864), numerical (507), judgment (1,045), and other (204)-enabling fine-grained performance analysis across modalities. Temporal scopes are categorized into long-term (1,946), mid-term (2,736), and short-term (494), reflecting diverse real-world scenarios. These characteristics position ChronoQA as a comprehensive, scalable, and challenging benchmark for evaluating temporal reasoning in retrieval-augmented systems.

Data Records

The ChronoQA dataset is available at Zenodo[23] and GitHub (https://github.com/czy1999/ChronoQA), released under the CC BY 4.0 license. The data is provided in JSON format (.json), where each item is a JSON object representing a single question-answer instance with rich metadata, as detailed in Table 3. Crucially, to ensure the traceability and verifiability of the source information, we have included a `golden_chunks_urls` field. This field contains an array of URLs that directly correspond to the evidence passages in the `golden_chunks` list, allowing users to reference the original news articles. The full dataset is distributed as a compressed archive occupying approximately 12 MB of disk space. The archive is organized into a directory containing the following files:

By temporal type: Absolute

2020年10月13日，是哪一支球队赢得了NBA总决赛?
On October 13, 2020, which team won the NBA Finals?

By time expression: Explicit

2020年10月13日，是哪一支球队赢得了NBA总决赛?
On October 13, 2020, which team won the NBA Finals?

By temporal type: Aggregate

中国在2023年宣布的住房贷款利率政策动态调整机制与境内企业境外发行证券的新规哪个更早生效?
Which took effect earlier: China's dynamic adjustment mechanism for housing loan interest rates announced in 2023, or the new regulations on domestic enterprises issuing securities overseas?

By time expression: Implicit

去年7月份，哪位中国选手进入了温网混双的决赛?
In July of last year, which Chinese player reached the Wimbledon mixed doubles final?

By temporal type: Relative

今年年初乐视开始执行的每周工作时间是多少小时?
At the beginning of this year, how many hours per week did LeEco start implementing as working hours?

By reference documents: Single

俄罗斯总理米舒斯京是什么时候访问了上海?
When did Russian Prime Minister Mishustin visit Shanghai?

By reference documents: Multiple

北京冬奥会开幕当天，彭斯在什么事件中对特朗普进行了强硬回击?
On the opening day of the Beijing Winter Olympics, in what event did Pence strongly counterattack Trump?

Fig. 5 Representative examples from ChronoQA.

Category	Subcategory	Count
Temporal Type	Absolute	2,529
	Aggregate	1,911
	Relative	736
Temporal Scope	Long-term	1,946
	Mid-term	2,736
	Short-term	494
Time Expression	Explicit	2,000
	Implicit	3,176
Referenced documents	Single	3,261
	Multiple	1,915
Answer Type	Entity	2,556
	Time	864
	Numerical	507
	Judgement	1,045
	Other	204
Total		5,176

Table 2. Statistics of Question Categories in ChronoQA.

- **chronoqa.json**: The main dataset file in JSON format. Each JSON object includes the question, answer, extensive metadata, and source URLs for all evidence passages.
- **chronoqa.csv**: A tabular version of the dataset for convenient browsing and quick reference, containing the same fields as the JSON file.
- **README.md**: Documentation describing the dataset structure, field definitions, data provenance, usage instructions, and citation guidelines.
- **scripts/**: Utility scripts for source article preparation, temporal question generation and validation.

Field	Type	Description
question	String	The main temporal question text in Chinese. "COTODAMA歌词音箱和苹果停产 iPhone6演列哪个事件更早发生?" *(Which event occurred earlier: COTODAMA lyrics speaker or Apple discontinuing iPhone 6 series?)*
question_date	String	The reference date for resolving relative time expressions (e.g., "this year"), in YYYY-MM-DD format. E.g., "2024-10-30"
answer	String	The ground truth answer derived from the source documents. E.g., "Apple Inc."
temporal_expression_type	String	Indicates if the question contains explicit or implicit time references. One of: `'explicit'`, `'implicit'`. E.g., "implicit"
temporal_scope	String	Categorizes the time span relevant to the question. One of: `'short-term'`, `'mid-term'`, `'long-term'`. E.g., "long-term"
temporal_granularity	String	The level of time precision required (e.g., day, month, year). E.g., "day"
temporal_type	String	Classification of the temporal reasoning required. One of: `'absolute'`, `'aggregate'`, `'relative'`. E.g., "aggregate"
answer_type	String	Categorizes the expected format/type of the answer. One of: `'entity'`, `'time'`, `'numerical'`, `'judgement'`, `'other'`. E.g., "entity"
reference_document_count	String	Indicates if the answer requires single or multiple source documents. One of: `'single'`, `'multiple'`. E.g., "multiple"
golden_chunks	Array of Strings	List of relevant text passages (evidence) from the source news articles. ["2019年7月23日COTODAMA推出...", "2019年7月17日苹果公司宣布..."] *(["On July 23, 2019, COTODAMA launched...", "On July 17, 2019, Apple announced..."])*
golden_chunks_urls	Array of Strings	A list of source URLs, where each URL corresponds to a text passage in golden_chunks. ["https://tech.sina.com.cn/...", "https://finance.sina.com.cn/..."]

Table 3. JSON format of the ChronoQA benchmark dataset.

Additionally, to ensure full transparency and support further research, we also provide the complete source corpus used in this study. This supplementary data, available in the GitHub repository, includes both the original raw news articles and the 294,696 processed intensive temporal paragraphs.

Technical Validation

This section presents evidence supporting the technical quality, reliability, and representational validity of the ChronoQA dataset.

Validation of Dataset Correctness. ChronoQA underwent a rigorous, multi-stage validation process combining automated evaluation and manual verification. We first applied rule-based checks to ensure structural consistency, including correct document references and logical coherence. Next, `gpt-4o` was used to assess the fluency, factual accuracy, and temporal relevance of each QA pair. To further guarantee quality, approximately 6,000 examples were manually reviewed, achieving a correctness rate exceeding 95% with a high inter-annotator agreement (Cohen's Kappa = 0.85). All identified error pairs were removed from the final release. These validation results confirm that ChronoQA provides a reliable and high-quality benchmark for assessing temporal reasoning in retrieval-augmented systems.

Validation of Dataset Diversity. ChronoQA demonstrates rich diversity across multiple dimensions, making it a robust benchmark for evaluating temporal reasoning. As shown in Table 2, the dataset is well balanced across three core temporal types: absolute, aggregate, and relative, and includes both explicit and implicit time expressions. This composition ensures broad coverage of diverse reasoning patterns and varying degrees of temporal ambiguity. A notable feature of ChronoQA is its substantial inclusion of multi-document questions, which account for 37% of the dataset. These questions require models to synthesize information from multiple sources, an essential yet underrepresented capability in existing benchmarks.

To further assess and validate the topical diversity of the source articles, we performed a thematic analysis using a pre-trained news category classifier. The results, illustrated in Fig. 7, confirm that ChronoQA spans a wide range of real-world domains. While Social Affairs constitutes the largest portion at 31.9%, there is substantial representation from other key areas, including Finance (16.3%), Entertainment (16.0%), International news (13.6%), Politics (9.0%), and Technology (8.2%). This thematic breadth ensures that ChronoQA can be used to evaluate models on time-sensitive queries across various fields, directly addressing the risk of topical bias.

Furthermore, as illustrated in Fig. 6, ChronoQA exhibits a wide distribution of question lengths, ranging from short, direct queries to longer, multi-part formulations. Questions with explicit time expressions tend to be more concise, while those involving implicit references or multiple documents are generally more complex and verbose. Similarly, numerical and judgment-based questions are typically shorter, whereas entity- and time-oriented questions often involve more elaborate phrasing. This variation in structure and complexity underscores ChronoQA's ability to comprehensively evaluate models across different dimensions of temporal reasoning.

Validation of Direct LLM Performance. We evaluated several state-of-the-art LLMs on ChronoQA to assess their ability to perform temporal reasoning. As shown in Fig. 8, while the models achieve moderate

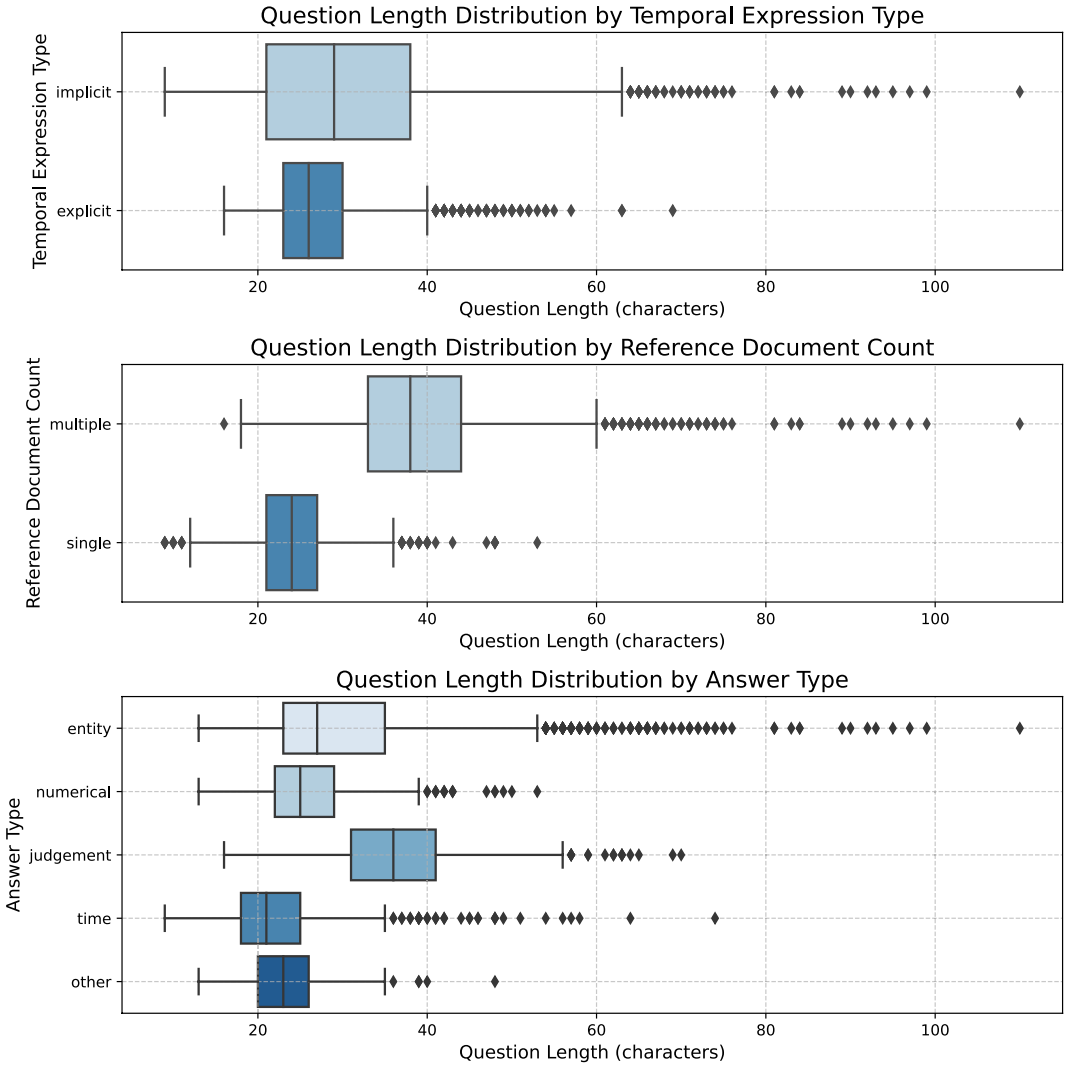

Fig. 6 Distribution of question lengths (number of characters) in ChronoQA.

performance on single-document questions, their accuracy drops significantly on multi-document questions. This gap highlights two key limitations: first, these models lack access to up-to-date knowledge, which is crucial for answering time-sensitive queries; second, they struggle with complex temporal reasoning, especially when multiple events need to be temporally aligned and integrated. These findings underscore the difficulty of ChronoQA and its effectiveness as a benchmark for advancing retrieval-augmented and temporally-aware question answering systems.

Retrieval Baseline Evaluation. To further validate the utility and challenge of ChronoQA, we conduct retrieval experiments using several representative methods under two retrieval depths ($K = 5$ and $K = 10$). As shown in Table 4, we compare four approaches: Native RAG, Temporal Filter, Query Rewrite, and Query Decomposition. Evaluation metrics include Recall, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG), reported for the overall dataset as well as for the multiple- and single-document subsets.

The results demonstrate clear performance differences among methods, indicating that ChronoQA can effectively distinguish between retrieval strategies. Notably, the Query Decomposition method achieves the best overall performance on most metrics, especially in the more challenging multiple-document setting. This suggests that ChronoQA not only requires robust temporal reasoning, but also benefits from advanced retrieval strategies capable of handling temporal constraints and multi-hop evidence aggregation. These baseline results provide a reference for future research and highlight the importance of temporal-aware retrieval in the context of time-sensitive question answering.

Error Analysis. To validate the challenges posed by ChronoQA and to identify key research opportunities, we analyzed 100 error cases from our baseline experiments. This analysis of failure modes serves not just to measure

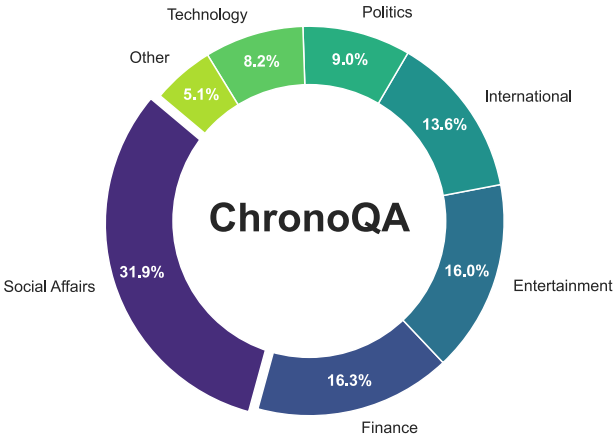

Fig. 7 Thematic distribution of source news articles in ChronoQA. The chart displays the proportion of articles across seven major categories, confirming the dataset's topical diversity.

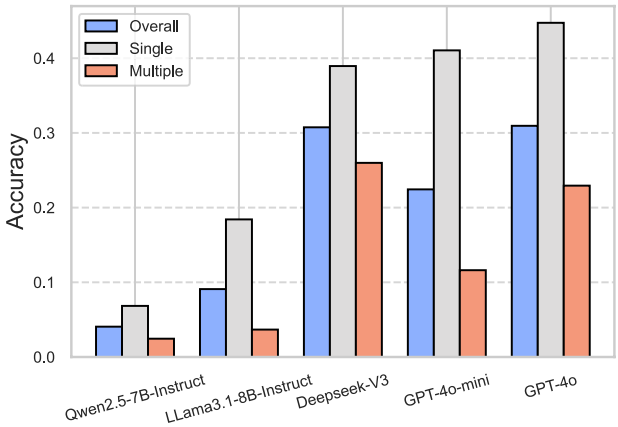

Fig. 8 Direct LLM Performance on ChronoQA.

difficulty, but to illuminate the specific, built-in features of our dataset that push the boundaries of current RAG systems.

Challenge 1: Resolving Complex Temporal Expressions. A core design principle of ChronoQA is its rich inclusion of both explicit and implicit temporal expressions. Our analysis confirms that this diversity effectively probes the limits of current models. We found that systems are particularly challenged by relative (implicit) time expressions, which account for a striking 69% of all analyzed failures, compared to just 31% for absolute (explicit) questions. This disparity demonstrates that ChronoQA successfully moves beyond simple date matching, creating complex scenarios that require deeper semantic and computational reasoning. This highlights a critical area for future research: developing models with more robust capabilities for parsing, grounding, and calculating based on natural language temporal phrases.

Challenge 2: Reasoning with Different Temporal Granularities. ChronoQA is intentionally designed to test reasoning across multiple levels of temporal granularities, revealing another significant research gap. Our findings show that reasoning at the fine-grained day level is the most difficult task for models, causing 62% of errors, far more than the month (27%) or year (11%) levels. As detailed in Table 5, this challenge is pervasive across different question types. The ability of ChronoQA to surface this weakness validates its utility as a benchmark for high-precision QA and calls for future investigation into models that are inherently sensitive to varying temporal granularities, a core requirement for many real-world applications.

Challenge 3: The Bottleneck of Temporal-Aware Retrieval. The multi-document and time-sensitive nature of questions in ChronoQA exposes a fundamental bottleneck in modern RAG pipelines: the lack of temporal awareness in the retrieval phase. Our analysis reveals that a failure to retrieve the correct evidence is the single largest source of error, responsible for a staggering 72% of incorrect answers. Even more telling, in 7% of cases, models failed even with perfect retrieval, indicating that reasoning remains a distinct challenge. The ability of ChronoQA to clearly separate and quantify these two failure modes is a key contribution. It strongly indicates

K	Method	Overall			Multiple			Single		
		Recall	MAP	NDCG	Recall	MAP	NDCG	Recall	MAP	NDCG
5	Native RAG	54.6 ± 1.1	52.5 ± 0.9	62.9 ± 1.3	26.9 ± 1.5	42.5 ± 1.8	49.5 ± 2.1	70.6 ± 1.0	58.3 ± 0.8	70.6 ± 1.0
	Temporal Filter	49.0 ± 1.4	45.1 ± 1.2	56.8 ± 1.5	15.5 ± 1.9	23.2 ± 2.2	30.8 ± 2.5	68.5 ± 1.2	57.8 ± 1.1	70.9 ± 0.9
	Query Rewrite	55.7† ± 1.0	53.3† ± 0.8	64.0† ± 1.1	29.0† ± 1.4	44.2† ± 1.6	51.6† ± 1.9	71.3† ± 0.9	58.6 ± 0.7	71.3† ± 0.9
	Query Decomposition	**61.9**† ± 0.9	**56.4**† ± 0.7	**71.8**† ± 1.0	**45.2**† ± 1.2	**52.4**† ± 1.4	**72.1**† ± 1.5	**71.6**† ± 0.9	**58.6** ± 0.7	**71.6**† ± 0.9
10	Native RAG	61.8 ± 1.0	53.5 ± 0.9	71.8 ± 1.1	35.6 ± 1.6	43.7 ± 1.7	62.6 ± 1.9	77.1 ± 0.9	59.1 ± 0.8	77.1 ± 0.9
	Temporal Filter	51.7 ± 1.3	43.7 ± 1.3	60.7 ± 1.6	17.5 ± 2.0	20.6 ± 2.4	35.5 ± 2.7	71.6 ± 1.1	57.2 ± 1.2	74.3 ± 1.0
	Query Rewrite	62.6† ± 0.9	53.8 ± 0.8	72.3† ± 1.0	36.7† ± 1.5	44.3† ± 1.6	63.2† ± 1.8	**77.7**† ± 0.8	**59.3** ± 0.7	**77.7**† ± 0.8
	Query Decomposition	**68.2**† ± 0.8	**57.6**† ± 0.7	**78.9**† ± 0.9	**53.9**† ± 1.3	**53.1**† ± 1.5	**83.2**† ± 1.4	76.5 ± 0.9	59.3 ± 0.7	76.5 ± 0.9

Table 4. Retrieval performance metrics for different models at K = 5 and K = 10. All metrics are reported in percentage (%). Results are reported as mean ± standard deviation. The best result in each column is **bolded**. † indicates a statistically significant improvement over the Native RAG baseline ($p < 0.05$).

Temporal Expression Type	Day	Month	Year	Total
Absolute (Explicit)	26	5	0	31
Relative (Implicit)	36	22	11	69
Total	62	27	11	100

Table 5. Cross-analysis of error distribution by temporal expression type and time granularity (N=100).

that the most urgent direction for future work is the development of novel retrieval strategies specifically designed to handle the temporal constraints embedded in user queries.

In summary, this analysis validates that ChronoQA is a challenging benchmark that pinpoints key weaknesses in existing systems. By requiring models to handle implicit time, reason at fine granularities, and perform temporal-aware retrieval, our dataset paves the way for the next generation of RAG research. Future work should focus on designing temporal-aware retrieval models that can interpret temporal expressions to filter and rank documents by relevance and timeliness, building advanced temporal reasoners with stronger intrinsic capabilities for date calculation, event sequencing, and duration understanding, and investigating hybrid systems that combine robust retrievers with specialized temporal reasoning modules to tackle both information access and synthesis challenges.

Data availability

The ChronoQA dataset, including the main QA pairs and supplementary source data, can be accessed via Zenodo at https://doi.org/10.5281/zenodo.17163857[23]. All data are released under the CC BY 4.0 license.

Code availability

Complete scripts and algorithms used for dataset construction, evaluation, and validation are available openly at https://github.com/czy1999/ChronoQA. These resources ensure full reproducibility of the dataset and offer researchers the flexibility to extend or customize the dataset generation processes for specialized requirements or updated scenarios.

Received: 12 May 2025; Accepted: 6 October 2025;
Published online: 21 November 2025

References

1. OpenAI. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
2. Touvron, H. *et al.* LLaMA: Open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
3. Du, Z. *et al.* GLM: General language model pretraining with autoregressive blank infilling. In *Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (Vol. 1: Long Papers)*, 320–335 (2022).
4. Sun, K., Xu, Y. E., Zha, H., Liu, Y. & Dong, X. L. Head-to-Tail: How knowledgeable are large language models (LLMs)? A.K.A. Will LLMs replace knowledge graphs? In *Proc. 2024 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long Papers)*, 311-325 (2024).
5. Laszlo, K. C. & Laszlo, A. Evolving knowledge for development: the role of knowledge management in a changing world. *J. Knowl. Manag.* **6**, 400–412 (2002).
6. Yang, C. C., Shi, X. & Wei, C.-P. Discovering event evolution graphs from news corpora. *IEEE Trans. Syst. Man Cybern. Part A* **39**, 850–863, https://doi.org/10.1109/TSMCA.2009.2015885 (2009).
7. Choudhury, N. & Uddin, S. Time-aware link prediction to explore network effects on temporal knowledge evolution. *Scientometrics* **108**, 745–776, https://doi.org/10.1007/s11192-016-2003-5 (2016).
8. Chen, Z., Li, D., Zhao, X., Hu, B. & Zhang, M. Temporal knowledge question answering via abstract reasoning induction. In *Proc. 62nd Annu. Meet. Assoc. Comput. Linguist. (Vol. 1: Long Papers)*, 4872–4889 (2024).
9. Chen, Z. *et al.* An adaptive framework for generating systematic explanatory answer in online Q&A platforms. Preprint at https://arxiv.org/abs/2410.17694 (2024).
10. Gao, Y. *et al.* Retrieval-augmented generation for large language models: a survey. Preprint at https://arxiv.org/abs/2312.10997 (2023).

11. Wu, F. *et al*. Time-sensitive retrieval-augmented generation for question answering. In *Proc. 33rd ACM Int. Conf. on Information and Knowledge Management (CIKM)*, 2544-2553 (2024).
12. Zhang, S. *et al*. MRAG: A modular retrieval framework for time-sensitive question answering. Preprint at https://arxiv.org/abs/2412.15540 (2024).
13. Abdallah, A., Piryani, B., Wallat, J., Anand, A. & Jatowt, A. Extending dense passage retrieval with temporal information. Preprint at https://arxiv.org/abs/2502.21024 (2025).
14. Kwiatkowski, T. *et al*. Natural Questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* **7**, 452–466, https://doi.org/10.1162/tacl_a_00276 (2019).
15. Joshi, M., Choi, E., Weld, D. S. & Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proc. 55th Annu. Meet. Assoc. Comput. Linguist. (Vol. 1: Long Papers)*, 1601–1611 (2017)
16. Nguyen, T. *et al*. MS MARCO: A human generated machine reading comprehension dataset. In *Proc. Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches, co-located with NIPS 2016*, vol. 1773 (2016).
17. Yang, Z. *et al*. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proc. 2018 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2369–2380 (2018).
18. Vu, T. *et al*. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Assoc. Comput. Linguist., ACL 2024*, 13697-13720 (2024).
19. Yang, X. *et al*. CRAG: Comprehensive RAG Benchmark. Adv. Neural Inf. Process **37**, 10470–10490 (2024).
20. Wang, S. *et al*. DomainRAG: A Chinese benchmark for evaluating domain-specific retrieval-augmented generation. Preprint at https://arxiv.org/abs/2406.05654 (2024).
21. Xu, Z. *et al*. Let LLMs take on the latest challenges! A Chinese dynamic question answering benchmark. *Proc. 31st Int. Conf. Comput. Linguist. (COLING 2025)*, 10435–10448 (2025).
22. Xiao, S., Liu, Z., Zhang, P. & Muennighoff, N. C-Pack: Packaged resources to advance general Chinese embedding. Preprint at https://arxiv.org/abs/2309.07597 (2023).
23. Chen, Z. *et al*. ChronoQA: A Question Answering Dataset for Temporal-Sensitive Retrieval-Augmented Generation [Data set]. Zenodo. https://doi.org/10.5281/zenodo.17163857 (2025).
24. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Assoc. Comput. Linguist., ACL 2024*, 6233–6251 (2024).
25. Tang, Y. & Yang, Y. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. Preprint at https://arxiv.org/abs/2401.15391 (2024).

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (Nos. U23A20296, 62272469, 72301284), and The Science and Technology Innovation Program of Hunan Province (No. 2023RC1007).

Author contributions

Z.C., E.M., and X.Z. designed the study and methodology. Z.C. led the implementation, data generation, and analysis, with assistance from E.M., J.L., X.J., and J.C. Y.L. performed dataset validation. Z.C. drafted the manuscript. X.Z., B.H., S.W., and D.Y. supervised the project. All authors contributed to manuscript revision and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.Z.