Unfolding the Headline: Iterative Self-Questioning for News Retrieval and Timeline Summarization

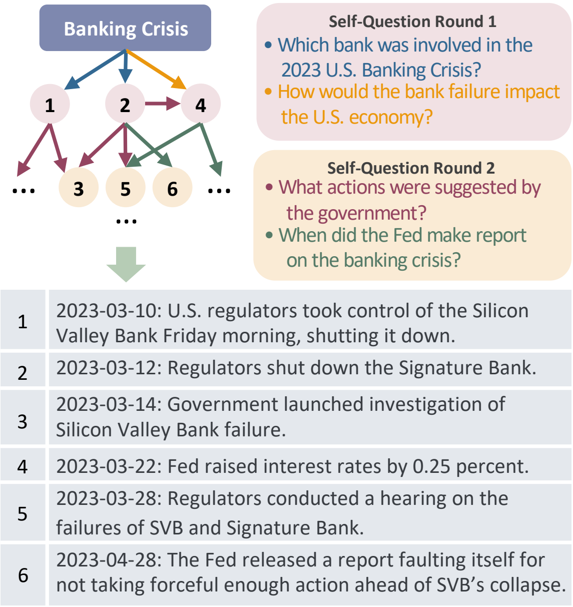
Weiqi Wu[1,3,4†], **Shen Huang**[2], **Yong Jiang**[2*], **Pengjun Xie**[2], **Fei Huang**[2], **Hai Zhao**[1,3,4*]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University,
[2]Tongyi Lab, Alibaba Group,
[3]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University,
[4]Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3
wuwq1022@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn
{pangda,yongjiang.jy,chengchen.xpj}@alibaba-inc.com

Abstract

In the fast-changing realm of information, the capacity to construct coherent timelines from extensive event-related content has become increasingly significant and challenging. The complexity arises in aggregating related documents to build a meaningful event graph around a central topic. This paper proposes **CHRONOS** - **C**ausal **H**eadline **R**etrieval for **O**pen-domain **N**ews Timeline SummarizatiO**n** via Iterative **S**elf-Questioning, which offers a fresh perspective on the integration of Large Language Models (LLMs) to tackle the task of Timeline Summarization (TLS). By iteratively reflecting on how events are linked and posing new questions regarding a specific news topic to gather information online or from an offline knowledge base, LLMs produce and refresh chronological summaries based on documents retrieved in each round. Furthermore, we curate Open-TLS, a novel dataset of timelines on recent news topics authored by professional journalists to evaluate open-domain TLS where information overload makes it impossible to find comprehensive relevant documents from the web. Our experiments indicate that CHRONOS is not only adept at open-domain timeline summarization, but it also rivals the performance of existing state-of-the-art systems designed for closed-domain applications, where a related news corpus is provided for summarization.[1]

1 Introduction

The exponential growth of news information in the digital era has made the task of understanding complex event narratives more critical. Timeline Summarization (TLS) (Yan et al., 2011; Wang et al., 2015; Chen et al., 2019; Gholipour Ghalandari and Ifrim, 2020) aims to extract and order the pivotal events from a multitude of textual sources over time, providing a structured view of historical developments. Despite the complexities inherent in extracting and organizing news events from multiple documents, the advent of Large Language Models (LLMs) (Kojima et al., 2022; OpenAI, 2023; Yang et al., 2023a; Bai et al., 2023) as powerful tools in understanding and generating high-quality text shows their potential in the field of TLS (Wang et al., 2023; Hu et al., 2024; Sojitra et al., 2024).



Self-Question Round 1
- Which bank was involved in the 2023 U.S. Banking Crisis?
- How would the bank failure impact the U.S. economy?

Self-Question Round 2
- What actions were suggested by the government?
- When did the Fed make report on the banking crisis?

1	2023-03-10: U.S. regulators took control of the Silicon Valley Bank Friday morning, shutting it down.
2	2023-03-12: Regulators shut down the Signature Bank.
3	2023-03-14: Government launched investigation of Silicon Valley Bank failure.
4	2023-03-22: Fed raised interest rates by 0.25 percent.
5	2023-03-28: Regulators conducted a hearing on the failures of SVB and Signature Bank.
6	2023-04-28: The Fed released a report faulting itself for not taking forceful enough action ahead of SVB's collapse.

Figure 1: TLS of the news *Banking Crisis*. Edges between event nodes can be established by iterative self-questioning, ultimately building an event graph around the target news for timeline generation.

The core of synthesizing a timeline is establishing temporal and causal relationships between events (Ansah et al., 2019; Li et al., 2021; Xiuying et al., 2022). As depicted in Figure 1, assuming

[†] This work was done during Weiqi Wu's internship at Tongyi Lab, Alibaba Group.

[*] Yong Jiang and Hai Zhao are corresponding authors.

[1] The code and dataset are released at https://github.com/Alibaba-NLP/CHRONOS.

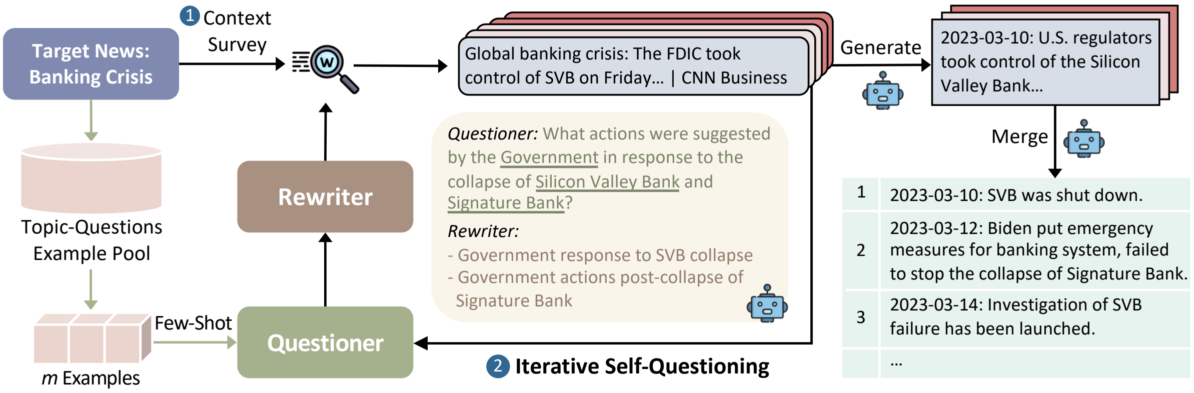

Target News: Banking Crisis
1 Context Survey
Global banking crisis: The FDIC took control of SVB on Friday... | CNN Business
Generate
2023-03-10: U.S. regulators took control of the Silicon Valley Bank...
Merge
Topic-Questions Example Pool
Rewriter
Questioner: What actions were suggested by the Government in response to the collapse of Silicon Valley Bank and Signature Bank?
Rewriter:
- Government response to SVB collapse
- Government actions post-collapse of Signature Bank
1 2023-03-10: SVB was shut down.
2 2023-03-12: Biden put emergency measures for banking system, failed to stop the collapse of Signature Bank.
3 2023-03-14: Investigation of SVB failure has been launched.
...
Few-Shot
Questioner
m Examples
2 Iterative Self-Questioning

Figure 2: Pipeline of CHRONOS. Giving a target news, it first searches for general context and iteratively poses questions to retrieve more relevant news, while employing a divide-and-conquer strategy to generate the timeline.

that each news event is represented as a distinct node, our goal is to establish edges between these nodes to present their correlation and ultimately form a heterogeneous graph, starting from the node of topic news. Establishing these edges can be effectively achieved through a search mechanism that retrieves relevant news articles. Thereby, an event node is linked to another if it can retrieve the other event through this search process.

Based on the sources of retrievable news, we categorize the TLS task into open-domain and closed-domain settings. Open-domain TLS refers to the process of generating timelines from news directly searched and retrieved from the Internet, while closed-domain TLS involves creating timelines from a predefined set of news articles related to a specific domain. Open-domain TLS faces additional challenges due to the vast and dynamic nature of online information. The information overload makes it difficult to retrieve relevant and comprehensive information from the Internet, introducing noisy data that complicates the task of filtering and assessing the quality of retrieved content. Hence, establishing relationships among events is more challenging in an open domain without access to a global view of relevant news.

To address such challenges, we propose **CHRONOS**, **C**ausal **H**eadline **R**etrieval for **O**pen-domain **N**ews Timeline SummarizatiO**n** via Iterative **S**elf-Questioning, a new scheme for both settings of TLS based on the Retrieval-Augmented Generation (RAG) framework (Li et al., 2022; Zhang et al., 2023; Gao et al., 2023; Zhao et al., 2024), as shown in Figure 2. By simulating the way humans search for information, which involves learning about the topic by formulating well-defined questions or problems, scanning retrieval results and term suggestions, and further coming up with new subquestions (Bates, 1989; O'Day and Jeffries, 1993), we iteratively utilize LLMs to pose 5W1H questions — What, Who, Why, Where, When, How — related to the news topic to gather comprehensive information about related events. We then rewrite the questions to enable a more effective search of it. For each round of retrieved news, we employ an LLM to generate a timeline, which would be merged to produce the ultimate timeline.

Despite the possibility of evaluating TLS systems in an open-domain setting by not utilizing the corpus provided by current news datasets, these datasets are often limited in size and topic diversity. Therefore, we introduce a more up-to-date and comprehensive news timeline dataset called Open-TLS. It encompasses various topics, including politics, economy, society, sports, and technology, and is sourced from news articles authored by professional journalists.

Our contributions can be summarized as follows:

- We propose CHRONOS, a novel retrieval-based approach to TLS by iteratively posing questions about the topic and the retrieved documents to generate chronological summaries.
- We construct an up-to-date dataset for open-domain TLS, which surpasses existing public datasets in terms of both size and the duration of timelines.
- Experiments demonstrate that our method is effective on open-domain TLS and achieves comparable results with state-of-the-art methods of closed-domain TLS, with significant

improvements in efficiency and scalability.

2 Related Works

2.1 Timeline Summarization

Timeline summarization (TLS) synthesizes a chronological narrative of event progression (Allan et al., 2001; Chen et al., 2019; Gholipour Ghalandari and Ifrim, 2020; Yu et al., 2021). While it could be approached as an extension of multi-document summarization (Chieu and Lee, 2004; Martschat and Markert, 2018), common strategies include focusing pivotal dates (Tran et al., 2015a,b; Steen and Markert, 2019) or identifying milestone events (Li et al., 2021; Xiuying et al., 2022). LLMs have introduced advancements to the field of TLS (Wang et al., 2023; Sojitra et al., 2024). Specifically, Hu et al. (2024) leverage LLMs for the generation and clustering of event summaries.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances LLMs by incorporating external knowledge during inference, addressing issues such as hallucination and outdated information (Liu et al., 2022; Shi et al., 2022; Ram et al., 2023; Izacard et al., 2023; Li et al., 2023; Agrawal et al., 2023). The retrieval sources of RAG can range from local databases (Siriwardhana et al., 2023) to web searches (Nakano et al., 2021; Komeili et al., 2022). As an application, Shao et al. (2024) researches a topic via multi-perspective Question-Asking during writing. We focus on the task of TLS and expand it to an open-domain setting, introducing news retrieval using the Internet with new challenges.

3 Methodology

We present CHRONOS, a new framework for effective and efficient TLS. It iteratively self-questions about previously retrieved news to gather other related events from various perspectives and combines the timelines it creates from each round of search for a thorough summary.

3.1 Iterative News Self-Questioning

The initial step of constructing a timeline for a specific target involves gathering relevant news articles. A straightforward method is to search with the news headline as a keyword to obtain the most general and directly linked information to the target news, where we define the retrieved articles as *News Context*. To obtain more comprehensive information about the target, we ask the LLM to generate questions that cannot be answered based on the news context, and iteratively search for new reference articles according to these questions.

To enhance the quality of self-questioning, we leverage the In-Context Learning (ICL) ability of LLMs by employing a few-shot prompt (Brown et al., 2020; Dong et al., 2022; Qian et al., 2024; Yao et al., 2024) to instruct the LLM to generate questions about the target news based on the previously retrieved news articles. The few-shot method is known to be highly dependent on the quality of the demonstration examples (Liu et al., 2022; Yang et al., 2023b; Peng et al., 2024). Therefore, curating effective few-shot examples becomes a critical aspect of our self-questioning method.

To systematically evaluate the quality of the generated questions in the field of TLS, we introduce the concept of *Chrono-Informativeness* (CI). It is designed to assess the ability of the questions to retrieve relevant documents that align chronologically with a reference timeline produced by a professional journalist. The *Chrono-Informativeness* of a set of questions $Q = (q_1, \ldots, q_m)$ for a given news topic is calculated as:

$$\mathrm{CI}(Q, N) = Date\_F_1(T_{Q,N}, T_{ref})$$

where $T_{Q,N}$ is the timeline generated from the N documents retrieved through the rewritten version of Q (see Sec. 3.2), and T_{ref} is the reference timeline. The $Date\_F_1$ score is a widely accepted metric in the field of TLS that compares the dates contained in the generated timeline to those in the reference timeline (detailed in Sec. 5.2).

By generating an extensive set of questions for a given news topic, we can use the greedy algorithm to identify the top m questions that maximize $CI(Q, N)$, selecting the question that provides the greatest improvement in CI during each step. The topic-questions pairs are stored in an example pool. When generating questions for a new target news story, we utilize a BERT-base-uncased model[2] to embed the query keyword and apply cosine similarity to retrieve the s most similar topics and associated example pairs from the pool. These dynamically retrieved few-shot demonstrations ensure that the demonstrations are contextually relevant and chronologically informative, which enhances the overall quality of the self-questioning process.

[2] https://huggingface.co/google-bert/bert-base-uncased

	T17	Crisis	OPEN-TLS					
			Overall	Politics	Society	Economy	Sports	Technology
# of topics	9	4	50	25	12	5	5	3
# of timelines	19	22	50	25	12	5	5	3
Avg. # of articles	508	2310	-	-	-	-	-	-
Avg. # of pub dates	124	307	-	-	-	-	-	-
Avg. duration (days)	212	343	4139	4624	1719	1297	8219	7694
Avg. l	36	29	23	25	19	22	20	20
Avg. k	2.9	1.3	1.8	1.8	2.1	1.7	1.8	1.6

Table 1: Statistics of closed-domain news TLS datasets and our proposed OPEN-TLS. A timeline contains l dates associated with k sentences describing the events that happened at each date.

3.2 Question Rewrite

However, the generated questions are usually quite complex to reach a certain level of depth and breadth, adding difficulty to searching. For instance, regarding the news of the Banking Crisis, the questioner posed a question *What actions were suggested by the government in response to the collapse of Silicon Valley Bank and Signature Bank*, and using this question directly as a query in a search engine yields poor retrieval performance. Hence, we apply a question rewriting mechanism (Ma et al., 2023) to improve the retrieval precision of our questions, achieved using a few-shot prompt design. Specifically, we employ the LLM to decompose each complex or under-performing query into 2-3 focused queries, such as *Government response to Silicon Valley Bank collapse* and *Government actions post-collapse of Signature Bank*. Such decomposition enhances the specificity and coverage of the retrieved documents, making the subsequent summarization tasks more effective.

3.3 Timeline Summarization

To create a coherent timeline containing l dates from the news articles retrieved using the questions, we utilize a divide-and-conquer strategy by first generating individual timelines from each round and merging them to produce the final timeline.

Generation We divide the problem of timeline generation into individual rounds of generation. At the end of each round of self-questioning, the LLM is instructed to extract the significant milestone events with clarified dates and write detailed summarizations of these events, using phrases directly from the news articles when possible to maintain authenticity and accuracy.

Merging After processing each round individually, the final step is to merge the generated timelines to ensure that only the most significant events are retained. The merging process involves aligning events from different rounds and resolving any conflicts of dates and descriptions. We instruct the LLM to select the top-l milestone events from the original timeline. Dates with more events happening are given precedence as they are likely to be more important since these events are consistently identified across multiple rounds of retrieval.

4 Open-TLS

Evaluating TLS systems commonly involves comparing system-generated timelines to those authored by professional journalists. While several benchmarks have been proposed for closed-domain news TLS along with the provided corpus for each topic, existing public datasets like T17 (Binh Tran et al., 2013) and Crisis (Tran et al., 2015b) remain constrained in terms of size and topical diversity. Furthermore, they often lack the timeliness and flexibility characterized by open-domain timeline generation. To bridge these gaps, we introduce *Open-TLS*, a novel dataset that collects timelines about recent news events, written by professional journalists from reputable news organizations such as the Associated Press[3], Public Broadcasting Service[4], and The Guardian[5].

As detailed in Table 1, Open-TLS comprises 50 timelines across various domains, including politics, economics, society, sports, and technology. The majority of the timelines are published post-2020. Each timeline is accompanied by a publication date and a query keyphrase that facilitates searching. In cases where the news is documented on Wikipedia, the title defined in Wikipedia is used as the query. Otherwise, we manually create a suit-

[3] https://apnews.com
[4] https://www.pbs.org
[5] https://www.theguardian.com

		Concat F1		Agree F1		Align F1		Date F1
		R-1	R-2	R-1	R-2	R-1	R-2	
GPT-3.5-Turbo	DIRECT	0.243	0.063	0.056	0.021	0.071	0.025	0.208
	REWRITE	0.233	0.067	0.054	0.022	0.070	0.026	0.205
	CHRONOS	0.328	0.086	0.092	0.078	0.092	0.034	0.283
GPT-4o	DIRECT	0.297	0.085	0.078	0.032	0.093	0.036	0.263
	REWRITE	0.283	0.080	0.079	0.034	0.093	0.038	0.272
	CHRONOS	0.351	0.103	0.105	0.047	0.121	**0.051**	**0.343**
Qwen2.5-72B	DIRECT	0.328	0.101	0.087	0.044	0.104	0.049	0.265
	REWRITE	0.337	0.106	0.091	0.046	0.107	0.050	0.291
	CHRONOS	**0.368**	**0.110**	**0.106**	**0.049**	**0.125**	0.050	0.324

Table 2: Experimental results on Open-TLS. We present the outcomes from the optimal self-questioning round.

able query based on its headline. All timelines are carefully curated to ensure high standards, providing exact dates and accurate narratives.

5 Experiments

5.1 Implementation Details

We construct experiments on CHRONOS based on three popular LLMs: GPT-3.5-Turbo[6], GPT-4o[7], and Qwen2.5-72B (Bai et al., 2023). We report the average results of 3 runs during evaluation.

Example Pool To build the example pool for the few-shot self-questioning prompt, we utilize GPT-4o to generate 50 questions for topics in the Crisis, T17, and Open-TLS datasets. Each topic is self-questioned based on the directly searched news context. When selecting the most similar demonstrations from the example pool, we exclude the topic-questions pair of the target news.

Search Engine For open-domain TLS, we use the Bing Web Search API[8] and set the query parameter *freshness* to the publish date of reference timeline to retrieve news articles only before it. We additionally use JINA[9] to read the content of the web pages. In the closed-domain setting, we employ Elasticsearch (Gormley and Tong, 2015), a well-established text search engine. Each document from the news corpus provided by the dataset is chunked into segments of approximately 500 words for retrieval.

5.2 Evaluation Metrics

We adopt the Tilse framework (Martschat and Markert, 2017, 2018) to evaluate the generated timeline with reference timelines, which includes the following metrics:

ROUGE-N Derived from the original ROUGE-N metrics, these metrics measure the overlap of N-grams in generated and reference timelines: (1) *Concat F1* computes ROUGE by concatenating all date summaries; (2) *Agree F1* computes ROUGE using only summaries of matching dates. (3) *Align F1* initially aligns predicted summaries with reference summaries based on similarity and date proximity, then calculates ROUGE between the aligned summaries, penalizing distant alignments.

Date F1 It is the F1 score of dates in the generated timelines compared with the ground truth.

5.3 Open-Domain TLS

5.3.1 Baselines

We propose two baselines for Open-Domain TLS. The number of retrieved news by baselines equals the total number of news retrieved by CHRONOS.

- **DIRECT** Directly search for the target news and output a timeline with the retrieved news.
- **REWRITE** Rewrite the target news to create 2-3 queries, search with these rewritten queries, and output a timeline with the retrieved news.

5.3.2 Results

The results in Table 2 demonstrate a consistent improvement across all metrics when using the CHRONOS approach compared to the baselines for each evaluated model. This indicates that

[6] https://platform.openai.com/docs/models/gpt-3-5-turbo
[7] https://platform.openai.com/docs/models/gpt-4o
[8] https://www.microsoft.com/en-us/bing/apis/bing-web-search-api
[9] https://jina.ai/reader/

Dataset	Model	AR-1	AR-2	Date F1
Crisis	CLUST	0.061	0.013	0.226
	EGC	0.079	0.015	0.291
	LLM-TLS▲	**0.112**	0.032	**0.329**
	LLM-TLS★	0.111	0.036	0.326
	DIRECT	0.094	0.031	0.182
	REWRITE	0.093	0.040	0.215
	CHRONOS	0.108	**0.045**	0.323
T17	CLUST	0.082	0.020	0.407
	EGC	0.103	0.024	**0.550**
	LLM-TLS▲	**0.118**	0.036	0.528
	LLM-TLS★	0.114	0.040	0.543
	DIRECT	0.077	0.028	0.418
	REWRITE	0.079	0.029	0.443
	CHRONOS	0.116	**0.042**	0.522

Table 3: Comparison of CHRONOS with previous works on closed-domain TLS benchmarks, reporting results of the top model. The best F1 scores are **bolded**, and the second bests are underlined.

CHRONOS enhances both the quality of event summarization and the alignment of dates with the reference timelines. The higher *Date F1* scores show that CHRONOS is more effective at accurately predicting the correct dates for significant events, with GPT-4o outperforming other models in extracting milestone events. Additionally, the improvements in ROUGE-N metrics suggest that the model excels at producing summaries of news events. Moreover, the general improvement by REWRITE compared with DIRECT shows the advantage of query writing preliminarily.

5.4 Closed-Domain TLS

5.4.1 Baselines

We evaluate CHRONOS on the closed-domain TLS task with several prior event-based approaches:

- **CLUST** Gholipour Ghalandari and Ifrim (2020) uses Markov clustering for event aggregation and determines the cluster significance by its date frequency in the news corpus.
- **EGC** Li et al. (2021) utilizes an event graph modelling method, integrating time-aware optimal transport to compress the whole graph into a salient sub-graph for event selection.
- **LLM-TLS** Hu et al. (2024) leverages LLMs as pseudo-oracles for incremental event clustering to construct timelines from a streaming context. We utilize LLaMA2-13B and Qwen2.5-72B for implementation and denote the resulted systems as LLM-TLS▲ and LLM-TLS★ respectively.

	Dataset	AR-1	AR-2	Date F1
CHRONOS	OPEN	**0.125**	**0.051**	**0.343**
	Crisis	**0.108**	**0.045**	**0.323**
	T17	**0.116**	**0.042**	**0.522**
Self-Questioning				
Random Exemplar	OPEN	0.113	0.042	0.312
	Crisis	0.079	0.038	0.314
	T17	0.112	0.036	0.498
Zero-Shot	OPEN	0.106	0.035	0.286
	Crisis	0.059	0.023	0.306
	T17	0.102	0.037	0.471
Question Rewrite				
w/o Rewrite	OPEN	0.095	0.038	0.262
	Crisis	0.078	0.047	0.286
	T17	0.072	0.026	0.446

Table 4: Ablation study of the topic-questions exemplars and question rewriter. OPEN is short for Open-TLS.

5.4.2 Results

We select the well-established benchmarks Crisis and T17 for evaluating closed-domain TLS and focus on representative performance metrics including Align F1 (short for AR-1 and AR-2) and Date F1. Table 3 presents a comprehensive overview of the performance of CHRONOS alongside previous representative works and two fundamental document retrieval baselines defined in the open-TLS task, i.e., DIRECT and REWRITE. We select the best-performing model to report its performance for presentation. Experiments show that CHRONOS matches and even exceeds the performance of previous models in terms of Alignment-based ROUGE-2 scores on both datasets. For the other lagging indicators, CHRONOS ranks second only to LLM-TLS on the Crisis dataset, as well as its Alignment-based ROUGE-1 score of T17. Regarding Date F1, its performance is less than 0.03 behind the state-of-the-art model, which however suffers from the other two metrics.

5.5 Ablation Study

5.5.1 Effects of Question Examples

CHRONOS selects the top-s most similar examples to the target news from the topic-questions example pool to construct few-shot self-questioning prompts. However, when these examples are selected randomly, i.e., *Random Exemplar* in Table 4, an evident drop in all metrics is witnessed across the three datasets, demonstrating the effectiveness of strategically selecting examples. This suggests

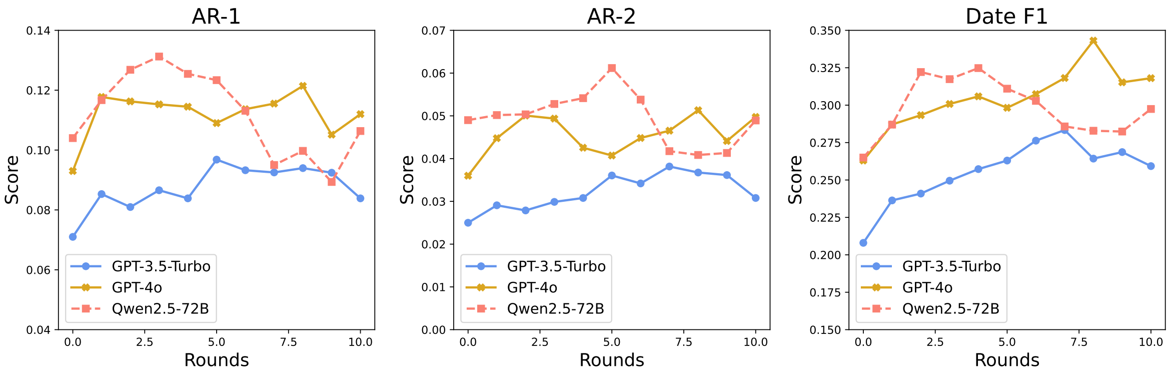


Figure 3: Impact of rounds of Self-questioning on model performance within the Open-TLS dataset.

that simply relying on providing examples and neglecting their relevance to the target is suboptimal, as random examples fail to provide contextual guidance for the model. Additionally, using a zero-shot prompt, which bypasses the use of examples entirely, leads to worse performance in most cases.

5.5.2 Necessity of Rewriting

To validate the importance of question rewriting, we compare the performance of our framework with and without this component. As shown in Table 4, the removal of the rewriting step leads to a significant decline in the overall performance of TLS, despite a slight improvement (+0.02) in the Alignment-based ROUGE-2 score for the Crisis dataset. This minor increase could be due to cases where the original questions closely resemble the phrasing of news articles, which enhances surface-level n-gram overlap. However, the overall decrease in Date F1 and other ROUGE metrics indicates that, without the rewriter, the model encounters difficulties in generating a complete and coherent timeline.

5.5.3 Rounds of Self-Questioning

The CHRONOS framework thrives on iterative self-questioning, a process that iteratively expands the news timeline. By increasing the number of questioning rounds, CHRONOS can retrieve a greater volume of news articles, thereby enhancing the comprehensiveness of its news database. However, as depicted in Figure 3, a pattern emerges across all three models on the Open-TLS dataset that their performance initially improves with additional rounds of questioning, but eventually declines. This trend can be attributed to the challenge of merging an excessive number of retrieved news articles into a coherent timeline.

5.5.4 Number of Retrieved news

To determine the impact of retrieved news in each round, we experiment with retrieving 20, 30, 40 documents using Qwen2.5-72B on the Open-TLS dataset. Table 5 indicates that increasing the number from 20 to 30 documents significantly improves the results, with marginal improvements when increasing to 40 documents. Intuitively, retrieving more documents provides the model with a richer context. However, due to the potential of introducing noise when integrating less relevant news, the marginal improvements observed when further increasing the number of retrieved news suggest a threshold beyond which the benefits plateau.

5.6 Inference Time

We further compare the running time of the LLM-based methods on the closed-domain datasets, CHRONOS and LLM-TLS. LLM-TLS, which processes each article individually, experiences substantial time delays due to the extensive news corpus of the Crisis dataset. On the other hand, CHRONOS employs a retrieval-based mechanism to focus on highly relevant news articles. Therefore, as shown in Table 6, CHRONOS spends only 5.6% of the total time required by LLM-TLS to reach a comparable performance. Even on the T17 dataset with fewer articles per topic, CHRONOS is almost twice as fast while producing similar or improved results. In conclusion, CHRONOS is more practical for real-world applications where efficiency and scalability are critical factors.

5.7 Discussions

5.7.1 Topic Analysis

We analyze the impact of different topics on the performance of CHRONOS, as shown in Figure 4.

N	**Concat-R1**	**Concat-R2**	**Agree-R1**	**Agree-R2**	**Align-R1**	**Align-R2**	**Date F1**
20	0.321	0.082	0.078	0.041	0.098	0.042	0.287
30	**0.368**	0.110	**0.106**	**0.049**	**0.125**	0.050	**0.324**
40	0.354	**0.121**	0.092	0.049	0.118	**0.051**	0.321

Table 5: Performance on Open-TLS with different numbers of news retrieved in each round.

	Crisis	**T17**
LLM-TLS	7 hr 12 min	2 hr 12 min
CHRONOS	**24 min**	**1 hr 9 min**

Table 6: Inference time for LLM-based methods.

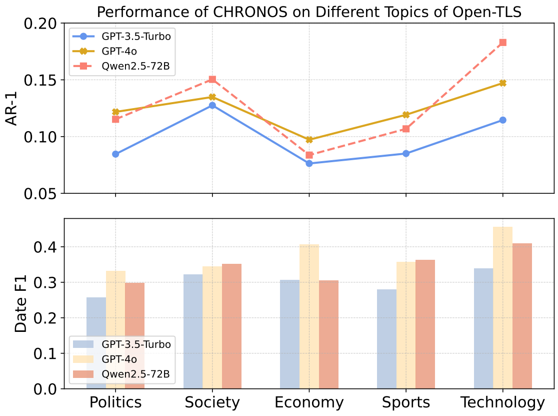


Figure 4: Topic analysis of CHRONOS on Open-TLS.

Upon examining the AR-1 metric, we observe that the Economy and Politics topics tend to challenge the LLMs, likely due to the significant amount of domain knowledge and entities required within these areas. The complexity and specificity of content in these domains make it harder for models to summarize event narratives effectively, resulting in relatively lower scores. Especially for the Economy topic, while the Date F1 scores remain relatively high, indicating that the models are generally successful in extracting dates, the lower ROUGE scores highlight the difficulty of summarizing economic events. Despite the variations in performance across different topics, the three models perform similarly on the Society topic. This convergence in performance could be attributed to the more general and less specialized nature of societal issues, which are easier for the models to handle equally well.

5.7.2 Case Study

Table 7 demonstrates how CHRONOS summarizes a timeline of *Greatest Apple Announcements*, constrained by the news publication date of June 30, 2024. CHRONOS generates two rounds of questions to gradually refine its knowledge of the news from a broad overview to more detailed insights. In Round 1, questions like *How has Apple's corporate strategy evolved?* guide the model to explore Apple's historical milestones and capture key events in it. In Round 2, the questioning shifts toward more specific topics to enrich the timeline with finer details. Comparing the generated timelines from both rounds to the reference timeline, CHRONOS accurately extracts major events with high precision. However, the omission of the Apple Vision Pro announcement and an incorrect date for the iPad unveiling indicate improvement in extracting milestone events with the correct dates.

6 Conclusion

In conclusion, this paper presents CHRONOS, a novel framework for TLS that leverages LLMs through an iterative self-questioning and retrieval-based process. Our method addresses the challenge of constructing coherent timelines by systematically retrieving event-related documents, reflecting the causal relationships between events. Experiments demonstrate its effectiveness in both open-domain and closed-domain TLS, as we propose a newly curated Open-TLS dataset for up-to-date open-domain news TLS. Moreover, CHRONOS demonstrates significant improvements in scalability and efficiency, making it a valuable tool for news TLS from vast and unstructured information.

Limitations

While our work presents several innovative contributions to the field of TLS, we acknowledge certain limitations that may affect its performance: (1) Our method is heavily dependent on the logical correlation between events for effective retrieval. However, if the causal links between events are not strong enough that they only happened chronologically, the system may struggle to retrieve relevant news articles efficiently. (2) The stability and consistency of our outputs are influenced by the volatility of LLMs and Search Engine Results

<table>
<tr><th colspan="2">Target News: Greatest Apple Announcements (2024.06.30)</th></tr>
<tr><th>Round 1</th><td rowspan="6"><b>Reference Timeline:</b>
1984-01-24: The Macintosh computer was unveiled.
2001-10-23: The iPod was unveiled, changing people's view about digital music players.
2007-01-09: The iPhone was unveiled, introducing the convenience of touchscreens.
2010-01-27: The iPad was unveiled, creating a middle ground between laptops and smartphones.
2014-09-09: The Apple Watch was unveiled, creating a device that made it possible to wear something akin to a smartphone on your wrist.
2016-09-07: AirPods was unveiled, helping popularize wireless headphones with an Apple chip that provided more reliable and stable connections while making it easy to shift from one gadget to another.
2024-06-05: The Apple Vision Pro was unveiled, which is a mixed-reality headset that seamlessly blends digital content with the physical world.</td></tr>
<tr><td><b>Self-Question:</b>
1. How did Apple transition from early computers into mobile tech?
2. How has Apple's corporate strategy evolved?
3. What were the key challenges during the initial years of Apple?
4. What were the critical factors that led to the innovation of the iPhone?
5. Who were the individuals instrumental in Apple's major products?</td></tr>
<tr><td><b>Generated Timeline:</b>
1984-01-24: Apple introduced the Macintosh computer.
2007-01-09: Apple introduced the first iPhone, marking a revolution in the smartphone industry with its multi-touch interface and sleek design.
...
2014-09-09: Apple announced the iPhone 6 and 6 Plus, as well as the Apple Watch, expanding its product lineup beyond phones.</td></tr>
<tr><th>Round 2</th></tr>
<tr><td><b>Self-Question:</b>
1. What were the key technological breakthroughs during Apple's development?
2. How did the launch of the iPhone impact Apple's strategy?
3. How has the user interface of Apple products evolved?
4. In what ways has Apple's design philosophy changed?
5. How did Apple navigate the transition from Steve Jobs to Tim Cook?</td></tr>
<tr><td><b>Generated Timeline:</b>
2001-10-23: The iPod was launched by Apple, revolutionizing how people listened to music and laying the groundwork for mobile devices.
2010-04-03: Apple released the iPad, a tablet that combined features from both the iPhone and the Mac.
2016-09-07: Apple launched the iPhone 7 and 7 Plus, the first models to eliminate the headphone jack, emphasizing wireless audio.
...</td></tr>
</table>

Table 7: Case study of the timeline generated by CHRONOS. The correctly predicted dates are in green, the missing dates are in red, and incorrect dates with correct events are in yellow. Overlapping words between the predicted and reference timeline are also highlighted.

Pages (SERPs). These fluctuations can lead to variations in the quality and reliability of the summaries generated by CHRONOS in real time.

Ethics Statement

A strong commitment to ethical standards and responsible research practices has guided the development and utilization of the Open-TLS dataset. We respect intellectual property rights and the guidelines established by content creators. Hence, we have strictly followed the terms of use set forth by the news organizations and websites from which we sourced the timelines. We have additionally made efforts to construct and present our dataset in a manner that preserves the integrity and accuracy of the original journalistic work. We are dedicated to ensuring that our dataset does not infringe upon the rights or privacy of individuals or organizations. Furthermore, all other datasets and models utilized in this work are publicly accessible and distributed under permissive licenses.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of new topics. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jeffery Ansah, Lin Liu, Wei Kang, Selasi Kwashie, Jixue Li, and Jiuyong Li. 2019. A graph is worth a thousand words: Telling event stories using timeline summarization graphs. *The World Wide Web Conference*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Marcia J. Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424.

Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. 2013. Predicting relevant news events for timeline summaries. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 91–92, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Xiuying Chen, Zhangming Chan, Shen Gao, Meng-Hsuan Yu, Dongyan Zhao, and Rui Yan. 2019. Learning towards abstractive timeline summarization. In *International Joint Conference on Artificial Intelligence*.

Hai Leong Chieu and Yoong Keok Lee. 2004. Query based event extraction along a timeline. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

Demian Gholipour Ghalandari and Georgiana Ifrim. 2020. Examining the state-of-the-art in news timeline summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1322–1334, Online. Association for Computational Linguistics.

Clinton Gormley and Zachary J. Tong. 2015. Elasticsearch: The definitive guide.

Qisheng Hu, Geonsik Moon, and Hwee Tou Ng. 2024. From moments to milestones: Incremental timeline summarization leveraging large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7232–7246, Bangkok, Thailand. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *ArXiv*, abs/2202.01110.

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6443–6456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *ArXiv*, abs/2305.14283.

Sebastian Martschat and Katja Markert. 2017. Improving ROUGE for timeline summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 285–290, Valencia, Spain. Association for Computational Linguistics.

Sebastian Martschat and Katja Markert. 2018. A temporally sensitive submodularity framework for timeline summarization. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 230–240, Brussels, Belgium. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Vicki L. O'Day and Robin Jeffries. 1993. Orienteering in an information landscape: how information seekers get from here to there. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

Kun Qian, Yisi Sang, Farima Bayat†, Anton Belyi, Xianqi Chu, Yash Govind, Samira Khorshidi, Rahul Khot, Katherine Luna, Azadeh Nikfarjam, Xiaoguang Qi, Fei Wu, Xianhan Zhang, and Yunyao Li. 2024. APE: Active learning-based tooling for finding informative few-shot examples for LLM-based entity matching. In *Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (DaSH 2024)*, pages 1–3, Mexico City, Mexico. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Daivik Sojitra, Raghav Jain, Sriparna Saha, Adam Jatowt, and Manish Gupta. 2024. Timeline summarization in the era of llms. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Julius Steen and Katja Markert. 2019. Abstractive timeline summarization. *Proceedings of the 2nd Workshop on New Frontiers in Summarization*.

Giang Tran, Eelco Herder, and Katja Markert. 2015a. Joint graphical models for date selection in timeline summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1598–1607, Beijing, China. Association for Computational Linguistics.

Giang Binh Tran, Mohammad Alrifai, and Eelco Herder. 2015b. Timeline summarization from relevant headlines. In *European Conference on Information Retrieval*.

Sha Wang, Yuchen Li, Hanhua Xiao, Lambert Deng, and Yanfei Dong. 2023. Web news timeline generation with extended task prompting.

Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2015. On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering*, 27:1301–1315.

Chen Xiuying, Li Mingzhe, Gao Shen, Chan Zhangming, Dongyan Zhao, Gao Xin, Zhang Xiangliang, and Rui Yan. 2022. Follow the timeline! generating abstractive and extractive timeline summary in chronological order. In *Transactions on Information Systems (TOIS '22)*. Association for Computing Machinery.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. 2011. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Zhe Yang, Damai Dai, Peiyi Wang, and Zhifang Sui. 2023b. Not all demonstration examples are equally beneficial: Reweighting demonstration examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13209–13221, Singapore. Association for Computational Linguistics.

Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. More samples or more prompts? exploring effective few-shot in-context learning for LLMs with in-context sampling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1772–1790, Mexico City, Mexico. Association for Computational Linguistics.

Yi Yu, Adam Jatowt, Antoine Doucet, Kazunari Sugiyama, and Masatoshi Yoshikawa. 2021. Multi-timeline summarization (mtls): Improving timeline summarization by generating multiple summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *ArXiv*, abs/2310.07554.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *ArXiv*, abs/2402.19473.

A Hyperparameters

In our experimental configuration, we have set the parameter m to 5, which represents the number of questions that the LLM generates in each round. The parameter N is set to 30, defining the maximum number of retrieved documents in each round. Furthermore, we have designated s as 3, indicating the number of few-shot examples included in the self-questioning prompt.

B Prompt Demonstration

We present the prompts used for three main modules within our system: self-questioning, question rewriting, and timeline generation.

B.1 Self-Questioning Prompt

Table 8 shows the prompt for news self-questioning. With the dynamically selected examples for each target news, the prompt is designed to guide LLMs in formulating a series of questions that expand the scope of the news database for generating the timeline.

Instruction for News Self-Questioning

You are an experienced journalist building a timeline for the target news. You need to propose at least 5 questions related to the Target News that the current news database cannot answer.
These questions should help continue organizing the timeline of news developments or the life history of individuals, focusing on the origins, development processes, and key figures of related events, emphasizing factual news knowledge rather than subjective evaluative content.
These 5 questions must be independent and non-overlapping. The overall potential information volume of all questions should be as large as possible, and the time span covered should also be as extensive as possible. Avoid asking questions similar to those already searched. Directly output your questions in the specified format.
Output format: ["Question_1", "Question_2", ...]
 	
{Retrieved Examples}
 	
Current News Database: {docs}
Target News: {news}
Questions Already Searched: {questions}

Table 8: Prompt for the questioner.

B.2 Rewrite Prompt

Table 9 presents the few-shot prompt used for question rewriting. The examples provided in the prompt demonstrate how to decompose complex questions while preserving their original intent.

Instruction for Question Rewriting

Generate 2-3 rewrite queries of the question as a python list, directly output it as ["..", "..", ..]

# Examples:
Question: When did the initial protests that led to the Egyptian Crisis begin?
Rewrite: ["Egyptian Crisis initial protests", "Time of protests lead to Egyptian Crisis"]

Question: When and where did Robert Jasmiden die?
Rewrite: ["Robert Jasmiden's death time", "Robert Jasmiden's death place"]

Question: What profession do Nicholas Ray and Elia Kazan have in common?
Rewrite: ["Nicholas Ray profession", "Elia Kazan profession"]

Question: {question}
Rewrite:

Table 9: Prompt for the rewriter.

B.3 Timeline Generation Prompts

Table 10 and Table 11 illustrate the prompts for timeline generation with detailed instructions.

Instruction for Timeline Generation

You are an experienced journalist building a timeline for the target news.

Instructions:
Step 1: Read each background news item and extract all significant milestone events related to the target news from your news database, along with their dates.
Step 2: Write a description for each event, including key detail information about the event, using the phrasing from the news database as much as possible. Save all events as a list. The format should be: [{"start": <date|format as "2023-02-02", cannot be empty, must include specific year, month, and day>, "summary": "<event description|no quotes allowed>"}, ...]

Target News: {news}
Current news database: {docs}

Table 10: Prompt for the timeline generator.

Instruction for Timeline Merging

You are an experienced journalist building a timeline for the target news.
Merge the existing news summaries and timelines in chronological order. When merging the news summaries, select the top-{1} significant news from the original timeline, and strictly follow the chronological order from past to present without changing the original date, using "\n" to separate events that occurred on different dates. Directly output your answer in the following format: [{"start": <date|format as "2023-02-02", cannot be empty, must include specific year, month, and day>, "summary": "<event description|no quotes allowed>"}, ...]
Target News: {news}
Original Timeline: {timelines}

Table 11: Prompt for merging the timelines from each round.