DTELS: Towards Dynamic Granularity of Timeline Summarization

Chenlong Zhang[1,2,*], Tong Zhou[1,2,*], Pengfei Cao[1,2], Zhuoran Jin[1,2],
Yubo Chen[1,2,†], Kang Liu[1,2], Jun Zhao [1,2]
[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China,
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China,
{zhangchenlong2023, tong.zhou}@ia.ac.cn
{pengfei.cao, zhuoran.jin, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

The rapid proliferation of online news has posed significant challenges in tracking the continuous development of news topics. Traditional timeline summarization constructs a chronological summary of the events but often lacks the flexibility to meet the diverse granularity needs. To overcome this limitation, we introduce a new paradigm, **D**ynamic-granularity **T**im**EL**ine **S**ummarization, (**DTELS**), which aims to construct adaptive timelines based on user instructions or requirements. This paper establishes a comprehensive benchmark for DTLES that includes: (1) an evaluation framework grounded in journalistic standards to assess the timeline quality across four dimensions: *Informativeness*, *Granular Consistency*, *Factuality*, and *Coherence*; (2) a large-scale, multi-source dataset with multiple granularity timeline annotations based on a consensus process to facilitate authority; (3) extensive experiments and analysis with two proposed solutions based on Large Language Models (LLMs) and existing state-of-the-art TLS methods. The experimental results demonstrate the effectiveness of the proposed solutions. However, even the most advanced LLMs struggle to consistently generate timelines that are both informative and granularly consistent, highlighting the challenges of the DTELS task.[1]

1 Introduction

With the surge in news production, the volume of news articles published on the internet is expanding rapidly, making it increasingly challenging to track the developments of news topics.

TimeLine Summarization (**TLS**) (Wang et al., 2016; Li et al., 2021; Chen et al., 2023b; Cao et al., 2023; Zhang et al., 2024) aims to construct a sequence of chronologically ordered summaries.

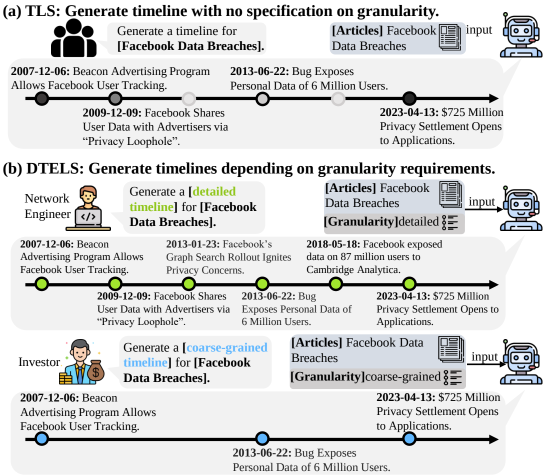

(a) TLS: Generate timeline with no specification on granularity.
Generate a timeline for **[Facebook Data Breaches].**
[Articles] Facebook Data Breaches
input
2007-12-06: Beacon Advertising Program Allows Facebook User Tracking.
2013-06-22: Bug Exposes Personal Data of 6 Million Users.
2009-12-09: Facebook Shares User Data with Advertisers via "Privacy Loophole".
2023-04-13: $725 Million Privacy Settlement Opens to Applications.
(b) DTELS: Generate timelines depending on granularity requirements.
Network Engineer
Generate a [detailed timeline] for **[Facebook Data Breaches].**
[Articles] Facebook Data Breaches
[Granularity]detailed
input
2007-12-06: Beacon Advertising Program Allows Facebook User Tracking.
2013-01-23: Facebook's Graph Search Rollout Ignites Privacy Concerns.
2018-05-18: Facebook exposed data on 87 million users to Cambridge Analytica.
2009-12-09: Facebook Shares User Data with Advertisers via "Privacy Loophole".
2013-06-22: Bug Exposes Personal Data of 6 Million Users.
2023-04-13: $725 Million Privacy Settlement Opens to Applications.
Investor
Generate a [coarse-grained timeline] for **[Facebook Data Breaches].**
[Articles] Facebook Data Breaches
[Granularity]coarse-grained
input
2007-12-06: Beacon Advertising Program Allows Facebook User Tracking.
2023-04-13: $725 Million Privacy Settlement Opens to Applications.
2013-06-22: Bug Exposes Personal Data of 6 Million Users.

Figure 1: (a) In traditional TLS, a timeline with a predefined number of node summaries is constructed. (b) DTELS provides timelines at different granular levels: network engineers require the technical causes and solutions to data breaches, therefore, a fine-grained granularity is preferred to track the technical details. For investors, a coarse-grained timeline showing the full picture of the breach's influence on investment may suffice.

These timelines provide traceable skeletons to support various applications including event modeling (He et al., 2024), policymaking (Chen et al., 2024b), crisis management, and temporal analysis (Ambe, 2023; Hu et al., 2022).

Traditional TLS typically constructs static timelines at a fixed granularity: in Figure 1a, for a specific news topic, the granularity is heuristically predefined by the number of "salient events". However, in practice, the granularity of the timeline should change dynamically, depending on user needs and the nature of news topics: **For readers**: Different readers have very different requirements on granularity for the same topic (see example in Figure 1b). **For news topics**: A reader's need for granularity varies across topics. One may require fine-grained timelines for trending news such as local disasters to follow the progression and imme-

[*] These authors contribute equally to this work.
[†] Corresponding Author.
[1] Codes are available at https://github.com/chenlong-clock/DTELS-Bench.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2682–2703
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

diate impacts. In contrast, for long-standing topics like the Russian-Ukrainian war, people may warrant coarse-grained timelines with wider intervals to capture broader developments.

Unfortunately, existing TLS ignores the importance of providing timelines at dynamic granularities. Existing evaluations also lack appropriate reference annotations and metrics to comprehensively evaluate timelines at dynamic granularities.

In this paper, we propose a new paradigm: **D**ynamic granularity **T**im**EL**ine **S**ummarization (**DTELS**). We define the granularity of a timeline by the degree of omission between the node summaries. Given a collection of news articles on the specific news topic and granularity requirements, our task aims to construct dynamic-granularity timelines tailored to various requirements.

Meanwhile, to take the study a step further, grounded in the criteria from journalism (Kunelius, 2006), an ideal timeline should: (1) convey information effectively, avoiding redundant events, and ensuring that no important events are missed. (2) maintain consistency with the granular requirements. (3) ensure the mentioned events in each summary are factually correct. (4) be self-contained, allowing the reader to clearly understand the context. By adhering to these criteria, we set the standard that not only meets the dynamic granularity needs but also upholds high quality.

We construct a benchmark including:

Evaluation Framework. To comprehensively measure a timeline, We propose metrics that address the aforementioned criteria:

Informativeness: This metric evaluates the effective volume of information in the node summaries. We propose a "*mount-then-measure*" paradigm to align predicted node summaries to those in the reference timeline based on the entailment score of the "*event atoms*", which represents the smallest unit of event information within a sentence.

Granular Consistency: The granularity is reflected by the amount of event information omitted between adjacent nodes. The more events omitted, the coarser the granularity is. We regard adjacent nodes as edges and calculate the ratio of mounts on the correct reference granularity edge.

Factuality: Considering the hallucinated contents and misinformation in the era of Large Language Models (LLMs) (Ji et al., 2023; Li et al., 2023; Zhang et al., 2023; Sun et al., 2024,?), it is crucial to ensure the information accuracy. We introduce a factuality metric that incorporates atoms-level entailment verification from reference news articles to measure the non-fabricated information in each summary.

Coherence: Coherence is pivotal in summarization tasks (Goyal et al., 2022; Steen and Markert, 2022; Jing et al., 2023). We adopt this metric for our task, ensuring that summaries are generated in a structurally, linguistically, and stylistically coherent manner. To facilitate this, we design a review form to guide the most advanced LLMs for coherence evaluation.

We verify the effectiveness of the metrics, showcasing high alignment with humans.

Dataset Construction. To ensure evaluation across varying granularities, we meticulously construct a dataset called *DTELS-Bench*. We initially collect diverse news topics and journalists' annotations on timelines from news events websites[2]. We then gather corresponding large-scale news articles from diverse sources, resulting in a large-scale, multi-source Chinese dataset. Subsequently, the reference timelines are annotated at three predefined granularities through a consensus-based automated annotation. Finally, the timelines are refined by specialists to ensure the authority.x

Comprehensive Evaluation. In the experiments, we present two LLM-based solutions for long-context and context-limited LLMs. We systematically evaluate our proposed solutions with multiple LLMs. In addition, we compare existing state-of-the-art extractive TLS approaches. Experiments show that our LLM-based solutions dominate in all dimensions, however, they fall short of providing high-quality information and aligning the required granularity. We then analyze the performance of these methods across various settings of DTELS. The results indicate that there is still substantial room for improvement in DTELS.

To sum up, our contributions are as follows:

- We propose a new task: **D**ynamic granularity **T**im**EL**ine **S**ummarization (**DTELS**). It aims to summarize timelines tailored to the unique needs of dynamic granularities.

- We build an event-centric evaluation framework. Extending from journalism, we propose metrics to evaluate timelines in four dimensions: informativeness, granular consistency, factuality, and coherence. Experiments with

[2] https://events.baidu.com

human annotators demonstrate the effectiveness of our metrics.

- We collect a large-scale, multi-source Chinese dataset, *DTELS-Bench*[3], which contains 543 news topics with 55,432 articles from 2,858 sources. It covers three predefined granularities annotated via a consensus-based mechanism. The expert's refinement enhances the annotation authority.

- We evaluate existing state-of-the-art TLS methods as well as LLMs with two proposed DTELS methods. Through extensive experiments, we find the proposed solutions outperform existing TLS methods, however, they are far from being an ideal solution to DTELS.

2 Related Works

2.1 Timeline Summarization Task

Timeline Summarization (TLS) has been a long-standing task in Natural Language Processing. The challenge of this task is to chronologically condense information from hundreds of articles. Existing work mainly focuses on generating and evaluating timelines at fixed granularity with sole evaluation metrics. The task is first proposed by Swan and Allan (2000). Kessler et al. (2012) presents an approach for detecting important dates to automatically construct timelines. Tran et al. (2013) provides a clear definition of TLS with fixed numbers of nodes for each news topic. Nguyen et al. (2014) introduces a system by selecting and ranking events from multiple documents. Martschat and Markert (2017) proposes an alignment-based ROUGE score and they proposed a submodularity Framework (Martschat and Markert, 2018) to construct timelines. La Quatra et al. (2021) propose a novel date selection method. However, these works disregard the varying granularity requirements. Besides, the ROUGE-based evaluation (Gholipour Ghalandari and Ifrim, 2020) results can be significantly affected by the narrative styles.

2.2 Timeline Summarization Dataset

Tran et al. (2013) proposes the "T17" dataset discussing famous topics. Tran et al. (2013) constructs the "Crisis" dataset focusing on long-span armed conflict topics. Wang et al. (2015). Gholipour Ghalandari and Ifrim (2020) builds "Entities" with longer time-ranges topics typed around 'people' and 'disasters'. Rajaby Faghihi et al. (2022) constructs a dataset called "CrisisTLS" focusing on the local crisis. Li et al. (2021) build a larger dataset TLS$_{100}$ covering various topics. Some recent works also propose LLM-based methods (Song et al., 2024; Hu et al., 2024; Chen et al., 2024a). However, they lack annotations across multiple levels. Besides, topics on existing datasets are likely to have been leaked in the pretraining corpus of LLMs, leading to potential unfair evaluations.

3 Task Definition

3.1 Timeline Summarization

In traditional text summarization, previous works have explored controlling granularity levels but lack a clear definition of "granularity" and focus on varying annotations. For the first time in TLS, we define and measure timeline granularity from an event-centric perspective. Consider a news topic $\mathbf{q}$ spans over a time range $\mathcal{T} = \{t_1, \ldots, t_n\}$ and a corresponding set of news articles $\mathcal{A} = \{A_{t_1}, A_{t_2}, \ldots, A_{t_n}\}$ as inputs. Each date $t_i \in \mathcal{T}$ is accompanied by multiple articles $A_{t_i} = \{a_{t_i,1}, \ldots, a_{t_i,m}\}$. The task is to generate a temporal sequence of summaries by model Θ:

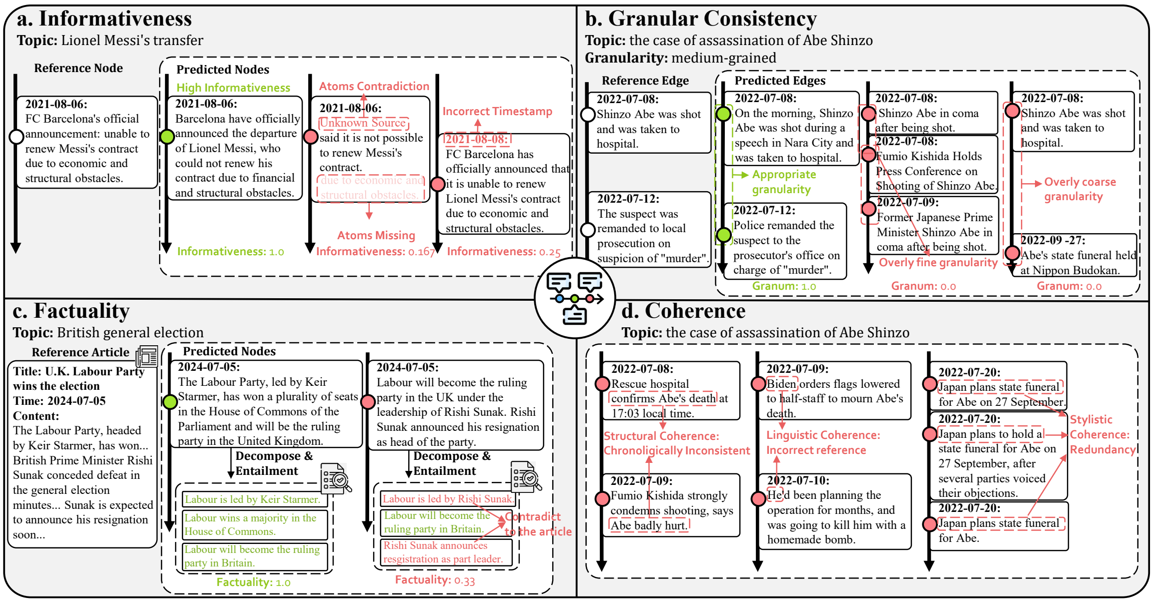
$$\mathcal{S} = \Theta(\mathbf{q}, \mathcal{A}), \quad (1)$$

where $\mathcal{S} = \{S_{t_1}, \ldots, S_{t_k}\}$ and k corresponds to the node numbers. S_{t_i} includes a timestamp $t_i \in T$ and summary s_i, i.e., s_i is a concise summary of the news event at time t_i. Typically, for a specific news topic, the amount of node summaries k is fixed based on the number of salient events.

3.2 Dynamic-granularity Timeline Summarization

In traditional text summarization, previous works (Zhong et al., 2022; Shen et al., 2022) have explored controlling granularity levels but lack a clear definition of "granularity" and focus on varying annotations. For the first time in TLS, we define and measure timeline granularity from an event-centric perspective. In traditional text summarization, previous works (Zhong et al., 2022; Shen et al., 2022) have explored controlling granularity levels but lack a clear definition of "granularity" and focus on varying annotations. For the first time in TLS, we define and measure timeline granularity from an

[3] All Chinese information in the paper is translated into English for ease of understanding.



2021-08-06: FC Barcelona's official announcement: unable to renew Messi's contract due to economic and structural obstacles.

2021-08-06: Barcelona have officially announced the departure of Lionel Messi, who could not renew his contract due to financial and structural obstacles.

2021-08-06: Unknown Source said it is not possible to renew Messi's contract. due to economic and structural obstacles.

2021-08-08: FC Barcelona has officially announced that it is unable to renew Lionel Messi's contract due to economic and structural obstacles.

2022-07-08: Shinzo Abe was shot and was taken to hospital.

2022-07-12: The suspect was remanded to local prosecution on suspicion of "murder".

2022-07-08: On the morning, Shinzo Abe was shot during a speech in Nara City and was taken to hospital.

2022-07-12: Police remanded the suspect to the prosecutor's office on charge of "murder".

2022-07-08: Shinzo Abe in coma after being shot.

2022-07-08: Fumio Kishida Holds Press Conference on Shooting of Shinzo Abe.

2022-07-09: Former Japanese Prime Minister Shinzo Abe in coma after being shot.

2022-07-08: Shinzo Abe was shot and was taken to hospital.

2022-09 -27: Abe's state funeral held at Nippon Budokan.

Title: U.K. Labour Party wins the election
Time: 2024-07-05
Content:
The Labour Party, headed by Keir Starmer, has won... British Prime Minister Rishi Sunak conceded defeat in the general election minutes... Sunak is expected to announce his resignation soon...

2024-07-05: The Labour Party, led by Keir Starmer, has won a plurality of seats in the House of Commons of the Parliament and will be the ruling party in the United Kingdom.

Labour is led by Keir Starmer.
Labour wins a majority in the House of Commons.
Labour will become the ruling party in Britain.

2024-07-05: Labour will become the ruling party in the UK under the leadership of Rishi Sunak. Rishi Sunak announced his resignation as head of the party.

Labour is led by Rishi Sunak.
Labour will become the ruling party in Britain.
Rishi Sunak announces resgistration as part leader.

2022-07-08: Rescue hospital confirms Abe's death at 17:03 local time.

2022-07-09: Fumio Kishida strongly condemns shooting, says Abe badly hurt.

2022-07-09: Biden orders flags lowered to half-staff to mourn Abe's death.

2022-07-10: He'd been planning the operation for months, and was going to kill him with a homemade bomb.

2022-07-20: Japan plans state funeral for Abe on 27 September.

2022-07-20: Japan plans to hold a state funeral for Abe on 27 September, after several parties voiced their objections.

2022-07-20: Japan plans state funeral for Abe.

Figure 2: Examples of metrics. Green nodes indicate positive examples and red nodes indicate negative examples.

event-centric perspective. Considering that event information passed through nodes is certain, we define granularity as the degree to which neighboring nodes are omitted: coarse-grained timelines with fewer nodes should omit less important events, while fine-grained timelines capture detailed chronological chains.

We introduce a granularity indicator "Granularity: [$\mathcal{G}_o$]" as an additional input to indicate the desired granularity. It can be either a specific number of nodes or a natural language instruction. Here, m denotes the chosen granularity level of the timeline. Based on this, the model Θ generates a timeline summarization at the specific granularity $\mathcal{G}_o$:

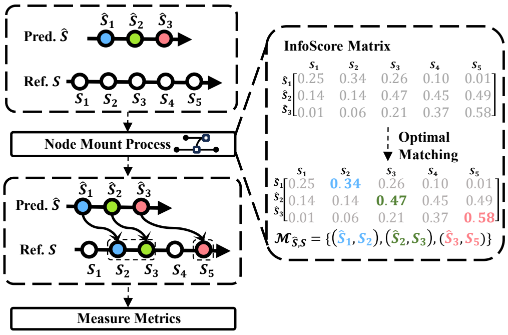
$$\mathcal{S}^{\mathcal{G}_o} = \Theta(\mathcal{G}_o, \mathbf{q}, \mathcal{A}), \tag{2}$$

where $\mathcal{S}^{\mathcal{G}_o} = \{(t_1^{\mathcal{G}_o}, s_1^{\mathcal{G}_o}), \ldots, (t_k^{\mathcal{G}_o}, s_k^{\mathcal{G}_o})\}$. the granularity of a timeline for a topic can vary: $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}$, where n ranges from coarse to fine granularity. This approach ensures that the summarization output matches the specified granularity requirements. The reference timelines are annotated at multiple granularity levels[4].

4 Evaluation Framework

4.1 Event Atoms

The references in narrative summarization are influenced by the annotator's preference and narrative style. Existing ROUGE-based evaluation (Lin, 2004) approaches evaluate the n-gram similarity, which can be inadequate in fairly reflecting the quality (Ng and Abrecht, 2015) of the outputs. For example, given reference and predicted summaries "*Barcelona announced the departure of Lionel Messi*" and "*King of Football, Messi, has left the club he served, Barcelona*.", the $Rouge_1 = 3/7$ varies with different narrative styles in predicted nodes. We hope to find a consistent measurement that is not affected by narrative style and granularity.


Pred. $\hat{S}$
$\hat{S}_1$ $\hat{S}_2$ $\hat{S}_3$
Ref. S
S_1 S_2 S_3 S_4 S_5
Node Mount Process
Pred. $\hat{S}$
$\hat{S}_1$ $\hat{S}_2$ $\hat{S}_3$
Ref. S
S_1 S_2 S_3 S_4 S_5
Measure Metrics
InfoScore Matrix
Optimal Matching
$\mathcal{M}_{\hat{S},S} = \{(\hat{S}_1, S_2), (\hat{S}_2, S_3), (\hat{S}_3, S_5)\}$

$$\begin{array}{c|ccccc} & S_1 & S_2 & S_3 & S_4 & S_5 \\ \hat{S}_1 & 0.25 & 0.34 & 0.26 & 0.10 & 0.01 \\ \hat{S}_2 & 0.14 & 0.14 & 0.47 & 0.45 & 0.49 \\ \hat{S}_3 & 0.01 & 0.06 & 0.21 & 0.37 & 0.58 \end{array}$$

$$\begin{array}{c|ccccc} & S_1 & S_2 & S_3 & S_4 & S_5 \\ \hat{S}_1 & 0.25 & \mathbf{0.34} & 0.26 & 0.10 & 0.01 \\ \hat{S}_2 & 0.14 & 0.14 & \mathbf{0.47} & 0.45 & 0.49 \\ \hat{S}_3 & 0.01 & 0.06 & 0.21 & 0.37 & \mathbf{0.58} \end{array}$$

Figure 3: The predicted timeline is mounted to the reference according to "Optimal Matching". The colored nodes denote mounted nodes.

Inspired by recent advances in atom-based evaluations (Min et al., 2023; Setty, 2024; Xu et al., 2024), we introduce the concept of "event atoms" as the fundamental units for evaluation, which remain consistent despite changes in narrative style and granularity. We define the "event atoms" as

[4] In the following section if $\mathcal{G}_o$ is not explicitly stated, we use the reference timeline with the same granularity $\mathcal{G}_o$ as the predicted timeline for evaluation.

the smallest distinguishable unit of events within a sentence. Each node summary s_i can be decomposed into a certain number of atoms: $\mathcal{E}_i = \{e_{i,1}, \ldots, e_{i,m}\} = Decompose(s_i)$, where m indicates the number of atoms. This function can be achieved by LLMs (detailed in Appendix A.2).

To evaluate a predicted node summary, we measure the amount of valuable event information it provides compared to the reference node summary using an entailment score. For a predicted node summary $\hat{s}_i$ and a reference node summary s_j, their event atoms are $\hat{\mathcal{E}}_i$ and $\mathcal{E}_j$, respectively. The entailment precision ent_p can be measured by:

$$ent_p(\hat{s}_i, s_j) = \frac{1}{|\hat{\mathcal{E}}_i|} \sum_{\hat{\varepsilon}_{i,s} \in \hat{\mathcal{E}}_i} Entail(\mathcal{E}_j, \hat{\varepsilon}_{i,s}), \quad (3)$$

where event atoms $\hat{\varepsilon}_{i,s}$ derives from $\hat{\mathcal{E}}_i$. $Entail(Evidence, Claim)$ quantifies the entailment of event atoms: it returns 1 if the evidence entails the claim, and 0 if it contradicts or is unrelated to the claim. The function can be implemented by the widely used Natural Language Inference models. (Camburu et al., 2018; Klemen et al., 2024).

Similarly, we can get the entailment recall $ent_r(\hat{s}_i, s_j)$. The entailment F1 can be calculated:

$$ent_{f1}(\hat{s}_i, s_j) = \frac{2 * ent_p(\hat{s}_i, s_j) * ent_r(\hat{s}_i, s_j)}{ent_p(\hat{s}_i, s_j) + ent_r(\hat{s}_i, s_j)}. \quad (4)$$

By adopting the score, we can evaluate the coverage of the node summaries over the references.

We propose a "mount-then-measure" paradigm, illustrated in Figure 3, to find the optimal mapping from the predicted timeline to the reference timeline. For a predicted node $\hat{\mathcal{S}}_i = (\hat{t}_i, \hat{s}_i)$, we mount it to a specific reference node $\mathcal{S}_j = (t_j, s_j)$ by computing the information score $InfoScore(\hat{\mathcal{S}}_i, \mathcal{S}_j)$. Considering the matching of event information on the temporal dimension for timelines, we introduce a temporal interval penalty term δ:

$$\delta_{\hat{t}_i, t_j} = \frac{1}{|\hat{t}_i - t_j|^2 + 1}. \quad (5)$$

Then, we can define the information score:

$$InfoScore(\hat{\mathcal{S}}_i, \mathcal{S}_j) = \delta_{\hat{t}_i, t_j} * ent_{f1}(\hat{s}_i, s_j). \quad (6)$$

4.2 Mount-then-measure Paradigm

The *InfoScore()* provides an objective measurement of the predicted nodes' coverage from an event-centric perspective. We can get the mapping cost between predicted and reference nodes via:

$$map(\hat{\mathcal{S}}_i \rightarrow \mathcal{S}_j) = -InfoScore(\hat{\mathcal{S}}_i, \mathcal{S}_j). \quad (7)$$

The mount process for the entire timeline can be automatically completed by Hungarian algorithm (Kuhn, 1955) for a global optimal matching:

$$\mathcal{M}_{\hat{\mathcal{S}}, \mathcal{S}} = \arg\min_{\mathcal{M}} \sum_{(\hat{\mathcal{S}}_i, \mathcal{S}_j) \in \mathcal{M}} map(\hat{\mathcal{S}}_i \rightarrow \mathcal{S}_j). \quad (8)$$

This process determines a maximum coverage of the predicted timeline to the reference, enabling fine-grained evaluation that requires references.

4.3 Evaluation Metrics

To evaluate timelines from multiple perspectives, we adopt criteria in journalism (Kunelius, 2006) and categorize the quality of a timeline into four dimensions: *Informativeness*, *Granular Consistency*, *Factuality*, and *Coherence*. The subsequent section details the definition of these metrics.

Informativeness. Informativeness measures the extent to which the node summary captures the essential information of events (Cao et al., 2023). As illustrated in Figure 2a, it is important to ensure the timeline contains all key atoms at correct timestamps and is not overly verbose. We match references for each node summary by "mount-then-measure". We calculate the informativeness $Info()$ after mounting the predicted timeline $\hat{\mathcal{S}}$ to the reference timeline $\mathcal{S}$:

$$Info(\hat{\mathcal{S}}) = \frac{1}{|\hat{\mathcal{S}}|} \sum_{(\hat{\mathcal{S}}_i, \mathcal{S}_j) \in \mathcal{M}} InfoScore(\hat{\mathcal{S}}_i, \mathcal{S}_j). \quad (9)$$

$\hat{S}_i$ and S_j are predicted and reference nodes in Equation 8.

Granular Consistency. Granular consistency measures how well the timeline aligns with its reference in terms of granularity. As illustrated in Figure 2b, differences in granularity emerge not from individual node content but from relationships between adjacent nodes.

we extend "mount-then-measure" to edge views: For a predicted timeline at $\mathcal{G}_o$, its edges are $\hat{\mathbf{E}} = \{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_{k-1}\}$, where $\hat{e}_m = (\hat{\mathcal{S}}_m, \hat{\mathcal{S}}_{m+1})$. The reference edges across all granularities are $\mathbf{E}^{\mathcal{G}} = \{\mathbf{E}^{\mathcal{G}_1}, \mathbf{E}^{\mathcal{G}_2}, \ldots, \mathbf{E}^{\mathcal{G}_n}\}$. We calculate the mapping cost of aligning a predicted edge $\hat{e}_m$ to a reference edge $e_n = (\mathcal{S}n, \mathcal{S}n+1) \in \mathbf{E}^{\mathcal{G}}$ using the formula:

$$\begin{aligned} map(\hat{e}_m \to e_n) = -InfoScore(\hat{\mathcal{S}}_m, \mathcal{S}_n) \\ -InfoScore(\hat{\mathcal{S}}_{m+1}, \mathcal{S}_{n+1}). \end{aligned} \tag{10}$$

We then mount $\hat{e}_m$ to a minimum cost e_n:

$$\mathcal{M}_{\hat{\mathbf{E}},\mathbf{E}} = \arg\min_{\mathcal{M}} \sum_{(\hat{e}_m, e_n) \in \mathcal{M}} map(\hat{e}_m \to e_n). \tag{11}$$

Finally, granular consistency is measured by the number of edges that are aligned with the correct granularity level $\mathcal{G}_o$:

$$Granu_i(\hat{\mathcal{S}}) = \frac{1}{|\mathbf{E}|} \sum_{(\hat{e}_m, e_n) \in \mathcal{M}} [e_n \in \mathbf{E}^{\mathcal{G}_o}], \tag{12}$$

where [] is a binary function.

Factuality. Factuality measures the faithfulness of summaries, which is crucial given the potential for hallucinated and fabricated content in LLMs (Chen et al., 2023a; Gekhman et al., 2023). In DTELS, factuality assesses whether the information in the timeline can be traced back to support articles. We use a selection mechanism to choose reference articles as support for each predicted node: For a given timestamp $\hat{t}_i$ in the predicted node $\hat{\mathcal{S}}_i$, we select reference articles $\mathcal{A}_{\hat{t}_i}$ that are closest to the timestamp. The factuality score is then computed using entailment precision:

$$Fact(\hat{\mathcal{S}}) = \frac{1}{|\hat{\mathcal{S}}|} \sum_{(\hat{s}_i, \hat{t}_i) \in \hat{\mathcal{S}}} ent_p(\hat{s}_i, \mathcal{A}_{\hat{t}_i}). \tag{13}$$

The articles are decomposed into a set of event atoms $\mathcal{E}_A$ as reference event atoms in equation 3. If the node summary contains hallucinated or fabricated content, it won't be fully entailed by the reference articles (see Figure 2c).

Dataset	#Topics	#Topic types	#Articles	#Sources	#Granu
T17	9	1	4,650	2	1
Crisis	4	1	9,240	3	1
Entities	47	2	45,075	1	1
CrisisTL	**1,000**	1	10,610	1	1
TLS100	100	4	10,379	2	1
Ours	543	**7**	**55,432**	**2,858**	**3**

Table 1: Comparison with existing datasets.

Coherence. While coherence is crucial in document summarization tasks (Wu and Hu, 2018; Chang et al., 2024), directly applying it to timeline summarization is insufficient. Unlike standard summaries that emphasize narrative coherence, timeline summaries demand structural coherence, including linguistic and stylistic consistency.

Figure 2d shows common coherence issues. We introduce an evaluation process similar to the ACL Review Form[5], assessing Structural, Linguistic, and Style Coherence, with details in Appendix B.

The process involves: (1) Paraphrasing content to improve understanding and reduce bias; (2) Rating each aspect from 1 to 3 and explaining the rationale for fine-grained evaluation; (3) Giving an overall score from 1 to 5 for a holistic assessment.

To reduce reviewers' workload, we use GPT-4o API[6] for automatic coherence assessment. Domain experts provide annotated examples to guide the model in understanding the criteria.

5 Dataset Construction

To ensure comprehensive evaluation across timelines at different granularity levels, dataset construction must meet two key premises: (1) The dataset should include news topics of varying complexity, types, and scales, with articles from diverse sources to simulate different granularity needs, enabling robust evaluation. (2) During annotations, annotators should minimize personal biases and annotate nodes at multiple granularity levels to facilitate the evaluation of both fine-grained and coarse-grained timelines. Our solutions to these challenges will be discussed in the following sections.

5.1 Data Collection

For data collection, we aim to assemble a diverse and representative set of news topics. We begin by leveraging Baidu's event news websites, known for expert fine-grained timeline annotation, to obtain news topics and their corresponding reference

[5] https://aclrollingreview.org/reviewform

[6] https://platform.openai.com/docs/models/GPT-4o

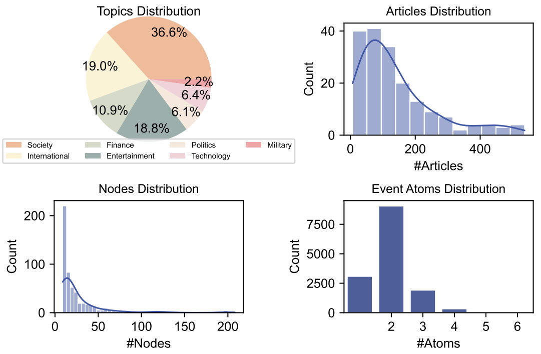


Figure 4: Dataset statistics.

timelines. We then manually filter these to ensure quality and diversity based on a standard.

The final dataset includes 543 news topics after October 2023, categorized into seven major types (Politics, Economy, Society, Science, Technology, Sports, and Entertainment), with reference timelines ranging from 9 to 200 nodes.

To gather reference articles, we use Baidu, Google, and Bing, employing multiple keywords to ensure each node is supported by at least 5 articles on average. The final dataset includes 55,432 articles, averaging 102 articles per topic. Table 1 compares our dataset with existing datasets.[7]

5.2 Consesus-based Annotation

DTELS requires multiple granularity levels, but annotating all is impractical. To aid evaluation, we define three levels: fine-grained (G_N), medium-grained (G_{10}), and coarse-grained (G_5), where N, 10, and 5 denote the number of nodes in the reference timelinee[8]. G_N reflects the original timeline with an unspecified node count, while medium and coarse timelines are annotated through consensus.

Even experienced journalists may differ in selecting events for coarse-grained timelines from fine-grained ones. To ensure uniformity, maximizing consensus among annotators on salient events and granularity is essential. However, this process can be costly and time-consuming in DTELS, especially with numerous articles per topic. We utilize GPT-4o to facilitate consensus through role-playing (He et al., 2023; Tao et al., 2024). The consensus-based annotation process involves three stages: (1) *Salient Events Decomposition*: For medium-grained timelines $\mathcal{S}^{\mathcal{G}_{10}}$, we decompose the fine-grained timeline $\mathcal{S}^{\mathcal{G}_N}$ into event atoms and group them by timestamp. (2) *Consensus-based Selection*: For each news topic, we prompt GPT-4o in different roles to select the 10 most important event groups from the atom groups, based on consensus among three roles. (3) *Expert Refinement*: Domain experts refine the selected groups to ensure quality, summarizing them into a 10-node timeline. The fine-grained timelines are annotated similarly. We show details and agreement in Appendix C.2. The results with high inter-annotator agreement show the effectiveness of our annotation.

We list the statistics of the dataset in Figure 4. A more detailed description of dataset construction and annotation can be found in Appendix C.

6 Experiments

6.1 Experimental Settings

For **extractive methods**, We implement two state-of-the-art methods (Gholipour Ghalandari and Ifrim, 2020) as baselines:

Datewise: This method selects key dates in a regression-based manner and then applies centroid-opt (Gholipour Ghalandari, 2017) to extract summaries for each date.

Clustering: This method clusters articles using TF-IDF vectors and then converts the clusters into a temporal graph. Dates are assigned to each cluster through a regression model. For DTELS, we constrain the number of nodes in the timeline according to the specified granularity level.

For **generative methods**, we select LLMs with Chinese ability. We propose two solutions:

Long-context Prompting (**LP**): For the long-context model, we directly prompt the model by providing the news topic, the entire articles with timestamps, and the granularity instruction.

Hierarchical Merging (**HM**): For models with limited context length, they generate summaries for each date according to the input articles's timestamps. Subsequently, these summaries are hierarchically merged following the merging prompts.

We also establish two distinct experimental settings to evaluate the task's characteristics:

Gold Timestamps (**GT**): We instruct models with correct timestamps to guide content generation, ensuring the focus is on content quality rather than timestamp accuracy. This setting can be used for both LP (LP_{GT}) and HM (HM_{GT}).

[7] All documents in our dataset are sourced under fair use for research purposes, and we adhere strictly to China's laws on copyright protection and the guidelines of data use in academic research. We will explicitly state this in the paper to avoid confusion.

[8] For simplicity, granularity is defined by node quantity. In application, the node numbers do not affect evaluation.

Methods	Models	Granularity $\mathcal{G}_N$				Granularity $\mathcal{G}_{10}$				Granularity $\mathcal{G}_5$			
		Info	Granu_N	Fact	Coherence	Info	Granu_{10}	Fact	Coherence	Info	Granu_5	Fact	Coherence
Datewise	-	**17.35**	**79.15**	**76.7**	**56.27**	**5.2**	16.13	**74.37**	**55.21**	**4.46**	8.06	**72.24**	**57.99**
Cluster	-	4.14	72.01	69.6	55.33	2.65	**16.90**	66.27	52.56	2.32	**10.73**	64.79	54.10
TO	GPT-3.5-Turbo	1.45	60.21	41.19	**91.20**	0.83	20.54	41.22	**94.76**	0.65	11.58	38.99	**97.46**
LP	GPT-4o	6.55	61.94	**65.78**	**69.21**	0.92	8.80	**86.82**	**77.15**	0.74	3.99	**88.11**	**87.55**
	GLM-3-Turbo	1.51	56.16	45.04	60.20	4.45	20.64	71.84	62.99	4.69	11.51	70.58	70.95
	Yi-medium	**9.87**	**66.45**	65.39	63.10	**4.91**	17.91	77.48	65.49	**8.69**	**23.36**	51.32	71.88
LP_GT	GPT-4o	1.91	59.69	48.24	55.89	2.17	**26.74**	46.56	56.28	1.76	14.78	47.94	56.30
HM	GPT-3.5-Turbo	24.24	**81.72**	91.95	65.87	0.82	7.96	91.96	68.38	0.72	4.74	91.96	76.92
	GLM-3-Turbo	21.07	72.43	87.39	67.40	1.37	**15.65**	87.61	68.01	0.91	**8.74**	88.33	71.34
	Yi-medium	17.46	75.32	82.26	64.28	**2.36**	14.34	86.02	65.56	1.75	6.91	85.60	73.41
	Qwen1.5-110b	28.00	76.51	83.99	**78.36**	2.24	10.75	83.27	79.77	**1.78**	6.81	81.09	**86.69**
	Qwen1.5-72b	24.69	80.25	85.14	74.82	0.92	10.31	85.4	**80.86**	0.74	5.48	84.56	85.57
	Qwen1.5-32b	23.37	73.97	86.29	68.64	0.61	10.14	86.32	75.47	0.55	5.54	88.04	82.08
	Qwen1.5-14b	25.26	67.78	85.76	69.69	0.71	13.06	86.31	69.98	0.56	6.98	85.58	78.64
HM_GT	GPT-3.5-Turbo	**36.82**	78.59	**94.63**	64.20	1.21	9.41	**93.59**	70.00	1.02	6.07	**93.48**	68.60

Table 2: Main results of different methods on DTELS task. The best results for different methods are in **bold**. The best results across all methods are underlined.

Topic Only (**TO**): Only providing the news topics and granularity requirements to generate a fabricated timeline.

The full implementations of the methods and model are detailed in Appendix D.

6.2 Main Results

We conduct experiments with the proposed metrics. The main results are shown in Table 2. From the results, we can observe the following conclusions:

LLMs dominate in DTELS. LLM-based methods outperform state-of-the-art models across all metrics. The HM excels at $\mathcal{G}_N$ in *Info* and *Granu*. while LP performs robustly at coarse and medium granularities, indicating the hierarchical method's strength in capturing details and LP's capability in managing timelines with long-context windows.

Context window matters for long-context prompting. With 200k context windows, Yi-medium-200k outperforms models with 128k windows, particularly at coarse granularities, demonstrating the effectiveness in broad event overviews.

Model capacity influences fine-grained metrics. Results from Qwen at different scales show that as model size decreases, performance in informativeness and granular consistency declines, suggesting that larger models are better at capturing and conveying detailed information.

Observation from the variants. With gold timestamps in "HM", the factuality is enhanced with temporal guidance. However, timestamps provide minimal benefits to LP and may reduce factuality and coherence. The "Topic Only" approach achieves the highest coherence scores but significantly lags in factuality, indicating that it maintains narrative continuity at the cost of factual accuracy.

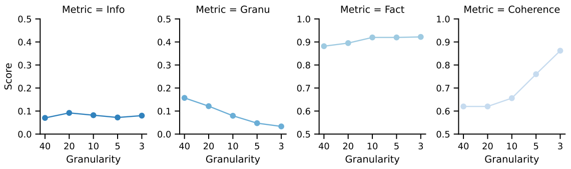


Figure 5: Extended evaluation on granularity levels.

7 Analysis

To further assess performance across different granularities, we conduct an extended evaluation using hierarchical merging with GPT-3.5-Turbo.

7.1 Extended Evaluation on Granularities

More Detailed Granularities. We define five distinct granularity levels: $\mathcal{G}_{40}$, $\mathcal{G}_{20}$, $\mathcal{G}_{10}$, $\mathcal{G}_5$, and $\mathcal{G}_3$. We collect a subset of the dataset with each topic containing over 50 nodes at $\mathcal{G}_N$. Reference timelines are annotated at these levels, and performance is assessed using Hierarchical Merging with GPT-3.5-Turbo. Results, shown in Figure 5, reveal that while coarser summaries generally offer better informativeness, factuality, and coherence, they may struggle with granular consistency, highlighting a trade-off between detail and summary quality.

Natural Language Granularity Instructions We evaluate natural language granularity instructions, defining fine- and coarse-grained instructions (see Table 3). Reference timelines include the both fine- ($\mathcal{S}^{\mathcal{G}_N}$) and coarse-grained ($\mathcal{S}^{G_5}$). Results in Table 3 show that the one-shot method performs competitively with the #Node method, indicating

Granularity	Granular Instruction	Info	Granu_i	Fact	Coherence
$\mathcal{G}_N$	Prompt	10.10	75.61	91.67	68.42
	One-shot	22.65	81.10	**93.88**	**70.86**
	#Node	**24.24**	**81.72**	91.95	65.87
$\mathcal{G}_5$	Prompt	0.74	7.26	91.44	69.2
	One-shot	**0.75**	**7.41**	**92.37**	71.79
	#Node	0.72	4.74	91.96	**76.92**

Table 3: Results of natural language granularity instructions, where #Node represents the HM method with the number of nodes as the granularity instruction.

that models learn to generate accurate timelines with natural language granularity instructions.

7.2 Influential Factors on Metrics

We analyze the influence of topics and the article numbers. We conclude that the two aspects greatly influence the performance. Results in Appendix E suggest improvement in stability is necessary.

7.3 Metrics Alignment with Human

Agreement Score. Annotators are asked to rate timelines on a scale from 1 to 5 based on the metrics. Pearson correlations are: informativeness (78.74%), granular consistency (76.66%), factuality (95.87%), and coherence (99.14%).

Consistency Score. Given pairs of timelines for a topic generated by two models, annotators rate the better one for each metric. We calculate the consistent score between annotators and metrics. Each metric's consistency exceeds 90%, showing high consistency between humans and the metrics.

We also conduct empirical comparisons on correlation alignment with existing metrics (e.g., Alignment ROUGE-L (Martschat and Markert, 2017), BERTScore (Zhang et al., 2020), and QAEval (Deutsch et al., 2021)). Details can be found in the Appendix F. (the annotation process is to let evaluators independently rate the timelines on a scale from 1 to 5 for the four aspects. Then after a group discussion, they revise their scores. Details are described in Appendix F):

7.4 Robustness of the Automatic Coherence Scoring

Since coherence can be subjective due to model biases. We conduct a robustness test by comparing the automatic coherence scores produced by different models: we use three evaluators: Qwen1.5 110B (Team, 2023), GPT-3.5-turbo, and GPT-4o-mini to score the timelines generated by the datewise method. We then assess the mean absolute error (MAE): $MAE(m_1, m_2) = \frac{1}{N}\sum_{i=1}^{N} |score_{m_1} - score_{m_2}|$ in Table 4. The results show that the automatic coherence scores are robust across different models, indicating the reliability of the automatic scoring.

m_1	m_2	$mae(m_1, m_2)$
Qwen1.5	gpt-3.5-turbo	0.2596
Qwen1.5	gpt-4o-mini	0.2615
gpt-3.5-turbo	gpt-4o-mini	0.1114

Table 4: Robustness of automatic coherence scoring across different models.

8 Conclusions

In this paper, we introduce a **D**ynamic-granularity **TimEL**ine **S**ummarization (**DTELS**) task, which aims to construct timeline summaries at dynamic granularity levels following the granularity requirements. We build a comprehensive benchmark including: (1) **Evaluation Framework**: We propose an event-centric evaluation along with metrics: informativeness, granular consistency, factuality, and coherence. Evaluation of alignment with the human annotator proves the rationality of the proposed metrics. (2) **Dataset Construction**: We construct a large-scale Chinese dataset for DTELS with consensus-based annotation for multi-granularity references. We apply expert refinement to ensure the authority of the annotation. (3) **Comprehensive Evaluation**: We present two solutions for large language models. Through experiments on existing state-of-the-art timeline summarization methods as well as LLM-based solutions on multiple models, we find that the DTELS task remains challenging. Further research is required to improve the informativeness granularity consistency. In the future, we plan to diversify the language sources and improve LLM-based methods to better capture information and enhance granular consistency.

Limitations

Though our DTELS approach has shown promising results, there are several limitations that need to be addressed in future work: Our approach relies heavily on the availability of a large-scale, annotated dataset. The creation of such datasets is time-consuming, which may limit the scalability and applicability of our approach to other domains or languages where such resources are not available. To evaluate the generated timelines, we rely on large language models' APIs, which are costly and may not be accessible to all researchers. Besides, The language of our dataset is Chinese, which may limit the generalizability of our approach to other languages. Further research is needed to develop more efficient data collection and evaluation methods that can be applied to a wider range of languages and domains.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.U24A20335, No. 62176257, No, 62406321). This work is also supported by the Youth Innovation Promotion Association CAS and the China Postdoctoral Science Foundation under Grant Number 2024M753500.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rex Ambe. 2023. Classification and quantification of timestamp data quality issues and its impact on data quality outcome. *Data Intelligence*, pages 1–39.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Pengfei Cao, Yupu Hao, Yubo Chen, Kang Liu, Jiexin Xu, Huaijun Li, Xiaojian Jiang, and Jun Zhao. 2023. Event ontology completion with hierarchical structure evolution networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 306–320, Singapore. Association for Computational Linguistics.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Jianhao Chen, Haoyuan Ouyang, Junyang Ren, Wentao Ding, Wei Hu, and Yuzhong Qu. 2024a. Timeline-based sentence decomposition with in context learning for temporal fact extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3415–3432, Bangkok, Thailand. Association for Computational Linguistics.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023a. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6325–6341, Singapore. Association for Computational Linguistics.

Tianyu Chen, Yiming Zhang, Guoxin Yu, Dapeng Zhang, Li Zeng, Qing He, and Xiang Ao. 2024b. EFSA: Towards event-level financial sentiment analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7455–7467, Bangkok, Thailand. Association for Computational Linguistics.

Xiuying Chen, Mingzhe Li, Shen Gao, Zhangming Chan, Dongyan Zhao, Xin Gao, Xiangliang Zhang, and Rui Yan. 2023b. Follow the timeline! generating an abstractive and extractive timeline summary in chronological order. *ACM Transactions on Information Systems*, 41(1):1–30.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Preprint*, arXiv:2010.00490.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Demian Gholipour Ghalandari. 2017. Revisiting the centroid-based method: A strong baseline for multi-document summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 85–90, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. SNaC: Coherence error detection for narrative summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 444–463, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhitao He, Pengfei Cao, Yubo Chen, Kang Liu, Ruopeng Li, Mengshu Sun, and Jun Zhao. 2023. LEGO: A multi-agent collaborative framework with role-playing and iterative feedback for causality explanation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9142–9163, Singapore. Association for Computational Linguistics.

Zhitao He, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, Zhiqiang Zhang, Mengshu Sun, and Jun Zhao. 2024. Zero-shot cross-lingual document-level event causality identification with heterogeneous graph contrastive transfer learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17833–17850, Torino, Italia. ELRA and ICCL.

Danyang Hu, Meng Wang, Feng Gao, Fangfang Xu, and Jinguang Gu. 2022. Knowledge representation and reasoning for complex time expression in clinical text. *Data Intelligence*, 4(3):573–598.

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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Liqiang Jing, Yiren Li, Junhao Xu, Yongcan Yu, Pei Shen, and Xuemeng Song. 2023. Vision enhanced generative pre-trained language model for multimodal sentence summarization. *Machine Intelligence Research*, 20(2):289–298.

Rémy Kessler, Xavier Tannier, Caroline Hagège, Véronique Moriceau, and André Bittar. 2012. Finding salient dates for building thematic timelines. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 730–739, Jeju Island, Korea. Association for Computational Linguistics.

Matej Klemen, Aleš Žagar, Jaka Čibej, and Marko Robnik-Šikonja. 2024. SI-NLI: A Slovene natural language inference dataset and its evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14859–14870, Torino, Italia. ELRA and ICCL.

Harold W. Kuhn. 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Risto Kunelius. 2006. Good journalism: On the evaluation criteria of some interested and experienced actors. *Journalism studies*, 7(5):671–690.

Moreno La Quatra, Luca Cagliero, Elena Baralis, Alberto Messina, and Maurizio Montagnuolo. 2021. Summarize dates first: A paradigm shift in timeline summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 418–427, New York, NY, USA. Association for Computing Machinery.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Kiem-Hieu Nguyen, Xavier Tannier, and Veronique Moriceau. 2014. Ranking multidocument event descriptions for building thematic timelines. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1208–1217, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Hossein Rajaby Faghihi, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2022. CrisisLTLSum: A benchmark for local crisis event timeline extraction and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5455–5477, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinay Setty. 2024. Factcheck editor: Multilingual text editor with end-to-end fact-checking. *arXiv preprint arXiv:2404.19482*.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024. Combining hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14651–14672, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2022. How to find strong summary coherence measures? a toolbox and a comparative study for summary coherence measure evaluation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6035–6049, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. 2024. Moss: An open conversational large language model. *Machine Intelligence Research*, 21(5):888–905.

Russell Swan and James Allan. 2000. Automatic generation of overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56.

Yufei Tao, Ameeta Agrawal, Judit Dombi, Tetyana Sydorenko, and Jung In Lee. 2024. ChatGPT roleplay dataset: Analysis of user motives and model naturalness. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3133–3145, Torino, Italia. ELRA and ICCL.

Qwen Team. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Giang Binh Tran, Tuan A Tran, Nam-Khanh Tran, Mohammad Alrifai, and Nattiya Kanhabua. 2013. Leveraging learning to rank in an optimization framework for timeline summarization. In *SIGIR 2013 Workshop on Time-aware Information Access (TAIA*.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Lu Wang, Claire Cardie, and Galen Marchetti. 2015. Socially-informed timeline generation for complex events. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065, Denver, Colorado. Association for Computational Linguistics.

William Yang Wang, Yashar Mehdad, Dragomir R. Radev, and Amanda Stent. 2016. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 58–68, San Diego, California. Association for Computational Linguistics.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. *Preprint*, arXiv:2406.02472.

Ming Zhong, Yang Liu, Suyu Ge, Yuning Mao, Yizhu Jiao, Xingxing Zhang, Yichong Xu, Chenguang Zhu, Michael Zeng, and Jiawei Han. 2022. Unsupervised multi-granularity summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4980–4995, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Event Atoms Decomposition

Evaluation metrics based on event atoms require decomposition on both the reference and generated timelines. The event atoms in reference timelines are annotated in advance by human annotators. For the generated timelines, the decomposition process should be completed in real-time. Similar to the decomposition for atomic facts (Min et al., 2023), we adopt GPT-3.5 to complete the automatic annotation.

A.1 Manual Decomposition Protocol

To maintain consistency and accuracy in event atom decomposition, human annotators should follow this protocol:

(1) Understanding the Context: Read the entire node summary carefully to understand the overarching event or narrative described and then identify the primary subject(s) and action(s) within the sentence.

(2) Segmentation of Events: Down each sentence into smaller units by identifying distinct actions or states that involve a subject and an object. Then, consider each clause within a complex sentence as a potential event atom if it represents a unique action or state. For instance, for the sentence “John arrived at the station and met his friend.”, two event atoms can be identified:

- Event Atom 1: “John arrived at the station.”
- Event Atom 2: “John met his friend.”

A.2 Automatic Decomposition

The automatic decomposition process using GPT-3.5 is implemented by prompts in Table 5.

B Coherence Review Form

In this section, we introduce the details of the Coherence Review Form and the sub-metrics.

Coherence is assessed through a review process similar to the ACL Review Form. As illustrated in Figure 6, we decompose the review form into three steps. We ask experts to provide a comprehensive categorization of the main coherence errors that occur in timeline summarization. Based on these errors, we propose three sub-metrics along with their evaluation criteria for the scores (detailed in Figure 7). We provide annotators with the task definition, detailed descriptions of the sub-metrics, and examples, both positive and negative, annotated by domain experts. For automatic evaluation, we apply GPT-4o API and apply the *Task Descriptions* as system prompts.

C Details of Dataset Construction

In this section, we provide a detailed description of the dataset construction.

C.1 Data Collection.

The data collection encompasses news topics of varying complexity, types, and scales. The original dataset contains 1,012 news topics. We apply a filtering standard for the news topics detailed as follows:

Diversity of Sources. To ensure a broad representation of perspectives and avoid bias in data collection, we include timelines and articles sourced from a diverse set of reputable news sources. For the news topics: the dataset encompasses diverse news topics from various geographical regions across domestic and international events (a total of 32 countries/regions are included). For the timeline, the timeline nodes should come from multiple sources. Only timelines with events sourced from at least three news sources are included. For the news articles, we select articles from 2,858 sources including global press, forums, and social media,

Timeliness of News Topics. The pretraining data for the models used in our evaluation are all prior to October 2023. To minimize the risk of using contaminated data from LLMs pretraining corpus, we exclude older or stale news topics that no longer reflect current events or have lost relevance over time. We select news topics after Oct. 2023 for dataset creation.

Event-centric and Complete News Topics. To ensure that each news topic revolves around a specific, well-defined event or series of related events. We retain topics that provide comprehensive coverage of the development and conclusion of events, capturing key milestones and outcomes. We let annotators evaluate each selected topic to confirm it narrates a coherent storyline from beginning to end, avoiding fragmented or ambiguous narratives, and verify that the reference timelines associated

Atoms Decomposition Prompts

System Prompt

You are a Fact Decomposer.
## Your task is:
As a specialized journalist, you will be provided with a sentence that may describe multiple events. Your task is to decompose the sentence into atomic propositions. An atomic proposition consists of, and only of, a subject, a predicate, and an object.
## Output format:
Please use the following format for your output:
[“Atom_1”, “Atom_2”, . . .]
## Example:
Here is an example for you to better understand the task:
Input: “Myanmar military: one-year state of emergency imposed”
Output: [“Myanmar military imposes state of emergency”, “State of emergency lasts for one year”]

Input

{Node Summary of a timeline}

Table 5: Prompts used for Atoms Decomposition Process. We show examples for the model to better comprehend the task.

with each topic adequately depict the chronology and significance of events.

By applying the standards, our dataset filter to 543 high-quality, multiple-sources news topics. Then, news articles for each topic are collected with the following steps: (1) Retrieve on search engines (Baidu, Google, and Bing) for articles with news topics as keywords and time limits to the beginning and end of the corresponding timeline to get the most relevant 5 articles. (2) For each reference node summaries annotated in Baidu event websites, we directly get one source article from the website. Then we apply the summary as keywords on previously mentioned search engines to get 4 articles. (3) Filter low-quality articles. By thresholding the article titles against the news topics, we filter out low-quality articles. Articles with a BERT embedding similarity score of less than 0.3 between the title and the news topic are filtered out.

We list the statistics of the dataset in Figure 4.

C.2 Consensus-based Annotation

To facilitate consensus, we prompt GPT-4o to play different roles as annotators, including news editor, journalist, and NLP researcher. These annotators focus on different aspects. The prompts are listed in Table 6.

GPT-4o annotators are instructed with the decomposed “event atoms” from the fine-grained timeline. These event atoms are grouped based on their timestamps, which could later be used to construct medium-grained and coarse-grained timelines. To determine the consensus among GPT-4o agents, we employed the following approach: (1) Each of the three GPT-4o agents independently selected their top 10 groups. (2) Groups selected by all three agents are automatically included in the final selection. (3) Groups selected by two agents are reviewed for inclusion based on their relevance and importance. After the independent selections are made, the selected event atom groups from all three annotators are compared. The primary focus here is to identify the level of consensus among the annotators. We list the agreement degree of the annotated nodes as shown in Table 7. Once GPT-4 has selected the initial set of events, domain experts review and refine these selections to ensure accuracy and completeness. The refinement process includes: (1) Fact-Checking: Ensuring that each selected event was factually accurate and well-supported by credible sources. (2) Composing Atoms: Composing Atomic Facts into node summaries. (3) Coherence Refinement: Refine the summary in total to ensure that the timeline as a whole presents a coherent narrative. (4) Detail Adjustment: Adding or removing details as necessary to meet the target granularity.

D Methods Implementation

The extractive baselines are implemented based on the original code provided by the authors. For LLMs, we use the official API. For Qwen (Team,

Role: News editor

System Prompt

You are a specialized news editor. Your response should be in JSON format, start with "{" and end with "}".

Input

Given the event atom groups derived from the original timeline, your task is to select the most critical event groups that should be included in a condensed timeline. As a News Editor, focus on the following metrics:

- Inclusive: The selected event groups should maximize the coverage of key events, ensuring that the most newsworthy and impactful events are included.
- Accurate: The selected event groups must accurately reflect the essential developments without adding any ambiguity or misinformation.
- Traceable: Ensure each selected event group can be directly traced back to the original timeline, maintaining the integrity and source of information.

Please select the top {N} event atom groups according to the [**Input**]. Your response must follow the [**Template**].

[**Example**]

{Example Annotation Results}

[**Template**]

["Group$_1$", "Group$_2$", . . . , "Group$_{\{N\}}$"] # Selected Event Atom Groups.

[**Input**]

Topic: {Input Topic}

Event Atom Groups: {Event Atom Groups of the timeline}

Role: Journalist

System Prompt

You are a specialized journalist. Your response should be in JSON format, start with "{" and end with "}".

Input

Given the event atom groups derived from the original timeline, your task is to select the most newsworthy event groups that should be included in a condensed timeline. As a News Editor, focus on the following metrics:

- Insightfulness: Focus on selecting event atom groups that offer deep insights into the topic, providing the audience with a comprehensive understanding of the events' context and implications.
- Objectivity: Ensure the selected groups are presented in an unbiased manner, maintaining journalistic integrity by avoiding sensationalism or subjective interpretation.
- Relevance: Select event atom groups that are most relevant to the central theme or story, ensuring that the timeline remains focused and cohesive.

Please select the top {N} event atom groups according to the [**Input**]. Your response must follow the [**Template**].

[**Example**]

{Example Annotation Results}

[**Template**]

["Group$_1$", "Group$_2$", . . . , "Group$_{\{N\}}$"] # Selected Event Atom Groups.

Role: NLP Researcher

System Prompt

You are a specialized NLP researcher. Your response should be in JSON format, start with "{" and end with "}".

Input

Given the event atom groups derived from the original timeline, your task is to select the most comprehensive event groups that should be included in a condensed timeline. As an NLP researcher, focus on the following metrics:

- Comprehensiveness: Ensure that the selected event atom groups provide broad and detailed coverage of the original timeline, capturing all significant events and nuances.
- Accuracy: Focus on selecting event atom groups that are factually correct, with a high level of precision in how the events are described, avoiding any distortion of the original data.
- Reproducibility: Prioritize event atom groups that can be easily traced back to the original data, ensuring that the selections are well-documented and can be verified by others.

Please select the top {N} event atom groups according to the [**Input**]. Your response must follow the [**Template**].

[**Example**]

{Example Annotation Results}

[**Template**]

["Group$_1$", "Group$_2$", . . . , "Group$_{\{N\}}$"] # Selected Event Atom Groups.

Table 6: Prompts for event consensus-based annotation, where N denotes the number of reference timeline nodes, in our case 10 and 5 for medium- and coarse-grained annotations.

Agreement Type	Count	Percentage
Full Agreement	3118	45.09%
Partial (1, 2)	2316	33.49%
Partial (1, 3)	573	8.28%
Partial (2, 3)	380	5.50%
No Agreement	525	7.59%

Table 7: Agreement among annotators, where 1, 2, and 3 correspond to the news editor, journalist, and NLP researcher, respectively. The agreement can be categorized as: (1) Full Agreement: The annotators selected the same event atom group. (2) Partial Agreement: Two out of three annotators selected the same event atom group. (3) No Agreement: No common event atom groups were selected by the annotators.

2023), we build upon their open-sourced weights[9]. The NLI model is implemented by BERT-based models (Devlin et al., 2018) fine-tuned on Chinese NLI datasets (Wang et al., 2022). The evaluation metrics include informativeness (Info), granular consistency (Granu), factuality (Fact), and coherence.

D.1 Details of LLM-based Methods

We choose LLMs with advanced Chinese capability, including both open-source and closed-source models for our analysis. For closed-source LLMs, we select the most representative GPT series and widely known Chinese model Yi-medium. For open-source LLMs, we select GLM-3 and Qwen series with multiple model sizes. Particularly, for LP, we evaluate models including GPT-4o (128k) (Achiam et al., 2023), Yi-medium (200k) (Young et al., 2024) and GLM-3-Turbo (128k) (Zeng et al., 2022). For hierarchical merging, we evaluate GPT-3.5-Turbo[10], GLM-3-Turbo (Zeng et al., 2022), Yi-medium (Young et al., 2024), and Qwen1.5 (Team, 2023). The temperature is set to 0 for greedy sampling.

D.2 Long-congtext Prompting

The prompts used for long-context prompting are illustrated in Table 8. To handle topics with hundreds of articles that may exceed the maximum token length, we truncate the last paragraph of each article recursively until the total content falls within the token limit. Figure 8 shows the distribution of token consumption for long-context prompting.

D.3 Hierarchical Merging

The hierarchical merging method first generates a day summary for the news topic based on prompts in Table 9. Then, all nodes are hierarchically merged to form a complete timeline. Similarly, if the input exceeds the token length, we do the same operation as in long-context prompting.

D.4 Natural Language Granular Instruction

We list the natural language granularity instruction in Table 10.

E Influential Factors on Metrics

We analyze the influence of topic types on the performance of hierarchical merging with GPT-3.5-Turbo. The results are shown in Figure 9. We observe that the performance on different topic types varies significantly. “Military” topics consistently achieve the highest scores in all metrics, suggesting that the model handles structured and well-defined content more effectively. Conversely, “Politics” and “Technology” topics present the greatest challenges, particularly in informativeness and coherence, likely due to the complexity and variability of information required in these domains. This suggests that the model’s performance is closely tied to the nature of the topic.

We also assess the influence of the number of news articles. The results are shown in Figure 10. We find that the model performs better with fewer news articles, as the model can better capture the key information and generate more coherent summaries. However, the factuality of the summaries decreases with fewer news articles, as the model may lack sufficient information to generate faithful summaries.

F Details of Alignment Evaluation

To measure how well human evaluators’ assessments of timelines align with the proposed metrics.

F.1 Evaluation Process

We choose three evaluators with a background in journalism and experience in summarization or timeline construction. Then, we prepare a set of 50 timelines generated by our DTELS system. Include a mix of high and low scores across different dimensions. We have each evaluator independently

[9] https://huggingface.co/Qwen

[10] https://platform.openai.com/docs/models/GPT-3-5

Long-context Prompts

System Prompt

You are a News Event Timeline Generator.
## Your task is:
As a specialized journalist, you will be provided with a news [topic] and related news [articles].
Based on this information, construct a chronologically ordered timeline summarizing the key events of the [topic].
Each event summary should be accompanied by an accurate timestamp.
## Output format:
Please use the following format for your output:
1. yyyy-mm-dd: Event summary 1
2. yyyy-mm-dd: Event summary 2
. . .
{N}. yyyy-mm-dd: Event summary {N}
## Note:
- The timeline should contain at least {N} event summaries.
- The summary content must match the timestamp.
- It's important to select key events to build the timeline, as not all [articles] are worth summarizing.

Input

[Topic]
{Topic of the Timeline}
[Article 0]
Title: {Title of Article 0}
Release-time: {Release-time of Article 0}
Content: {Content of Article 0}
. . .

Table 8: Prompts used in Long-context Prompting (LP) for long-context large language models. Here, N denotes the required node amounts for the timeline.

assess the timelines using the scoring sheets in Figure 11. After the initial round, we facilitate a group discussion where evaluators can compare their scores and discuss discrepancies. This can help in understanding different perspectives and potentially refining the evaluation criteria. Then, we allow evaluators to revise their scores based on insights gained from the discussion.

F.2 Correlation Evaluation

we calculate the correlation coefficient between existing metrics and the human assessment results. The results in Table 11 indicate that existing candidates-references alignment-based metrics (Alignment ROUGE-L (Martschat and Markert, 2017) and BERTScore (Zhang et al., 2020)) focus more on the amount of effective information transferred from the reference. QAEval (Deutsch et al., 2021), which is similar to recalling factual information from references, achieves a competitive correlation with factuality. However, when it comes to other aspects like granularity, factuality, and coherence, existing metrics failed to align with all aspects. In contrast, we can observe that our proposed metrics align closely with the specific need of the DTELS task, demonstrating the effectiveness of our metrics.

F.3 Metrics Definitions

F.3.1 Informativeness

Informativeness measures how much useful information is provided by each node in the timeline. A high score indicates that the node adds significant and relevant detail to the overall understanding of the event.

F.3.2 Granular Consistency

Granular Consistency assesses how well the timeline maintains a coherent level of detail across different nodes. A high score reflects that the granularity of events is consistent and appropriate throughout the timeline.

F.3.3 Factuality

Factuality evaluates the accuracy and truthfulness of the information presented in each node. A high score indicates that the node contains verified and accurate facts.

F.3.4 Coherence

Coherence measures how logically and smoothly the nodes are connected to form a comprehensible narrative. A high score suggests that the timeline is well-organized and the events are presented in a logical order.

Day Summary Prompts

System Prompt

You are a News Event Timeline Generator.
## Your task is:
As a specialized journalist, you will be provided with a news [topic] and related news [articles]. Based on this information, construct a chronologically ordered timeline summarizing the key events of the [topic]. Each event summary should be accompanied by an accurate timestamp.
## Output format:
Please use the following format for your output:
1. yyyy-mm-dd: Event summary 1
2. yyyy-mm-dd: Event summary 2
...
{N}. yyyy-mm-dd: Event summary {N}
## Note:
- There can only be ONE event summary per day.
- It's important to select key events to build the timeline, as not all [articles] are worth summarizing.

Input

[Topic]
{Topic of the Timeline}
[Article 0]
Title: {Title of Article 0}
Release-time: {Release-time of Article 0}
Content: {Content of Article 0}
...

Timeline Merging Prompts

System Prompt

You are a News Event Timeline Generator.
## Your task is:
As a specialized journalist, you will be provided with a news [topic], multiple partially completed timelines. Based on this information, merge the timelines to create a chronologically ordered timeline summarizing the key events of the [topic].
## Output format:
Please use the following format for your output:
1. yyyy-mm-dd: Event summary 1
2. yyyy-mm-dd: Event summary 2
...
N. yyyy-mm-dd: Event summary N
## Note:
- There can only be ONE event summary per day.
- It's important to select key events to build the timeline, as not all events are worth summarizing.

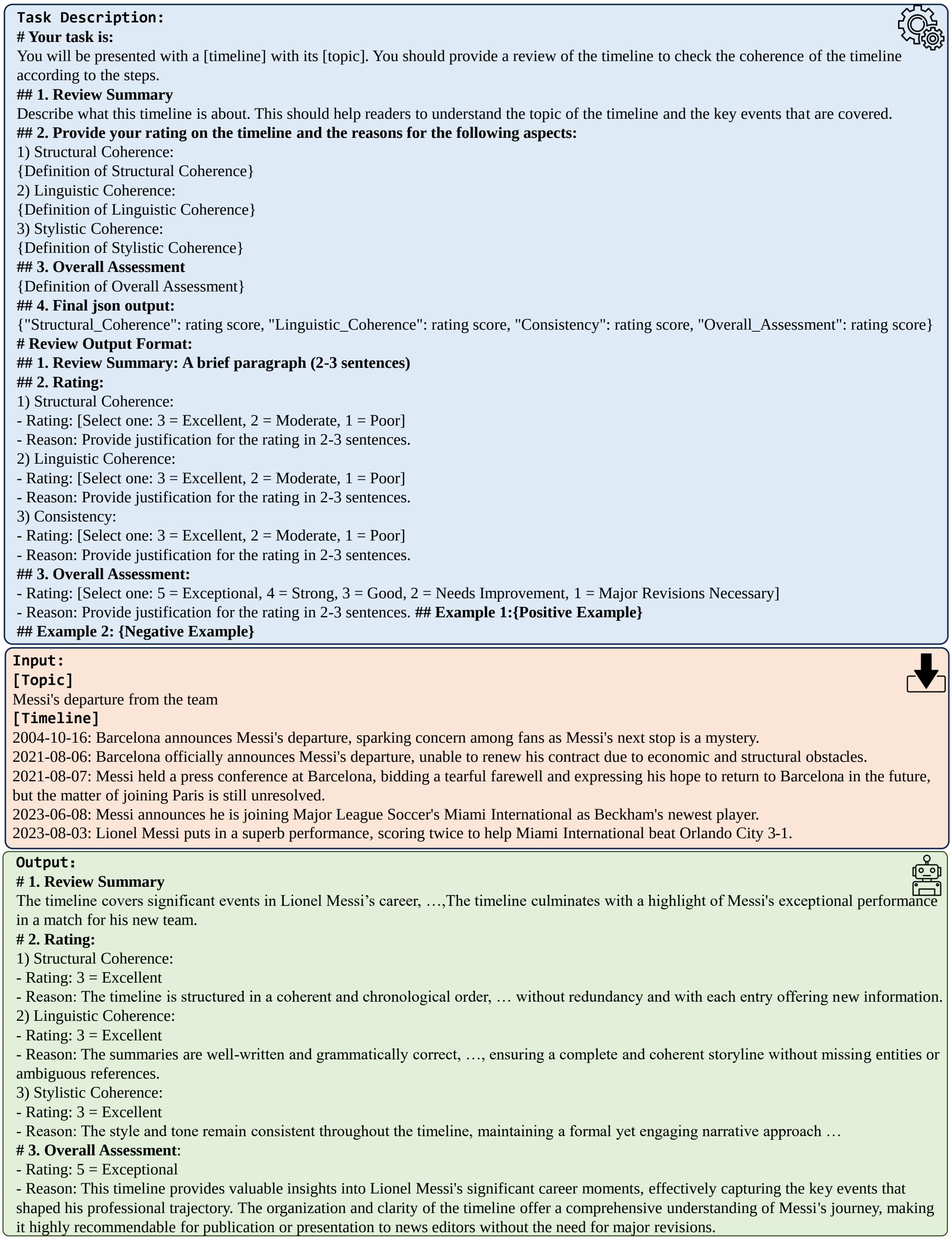
Input

[Timeline 0]
Timeline 0
...

Table 9: Prompts used in Hierarchical Merging (HM) for context length-limited large language models, where N denotes the required node amounts for the timeline.

Type	#Node	$\mathcal{G}_o$
Prompt	5	Please generate a coarse-grained timeline.[Task Prompt*]
	N	Please generate a fine-grained timeline. [Task Prompt*]
One-shot	5	Please generate a timeline like: {Timeline with 5 nodes}
	N	Please generate a timeline like: {Timeline with {N} nodes}

Table 10: Natural language granularity instructions used in the experiments. [Task Prompt*] denotes the prompts used for timeline summarization (LP and HM prompts in our paper). The "*" indicates that the node amounts {N} in the [Task Prompt] is replaced with "N" to represent an arbitrary number.



Task Description:

# Your task is:

You will be presented with a [timeline] with its [topic]. You should provide a review of the timeline to check the coherence of the timeline according to the steps.

## 1. Review Summary

Describe what this timeline is about. This should help readers to understand the topic of the timeline and the key events that are covered.

## 2. Provide your rating on the timeline and the reasons for the following aspects:

1) Structural Coherence:
{Definition of Structural Coherence}
2) Linguistic Coherence:
{Definition of Linguistic Coherence}
3) Stylistic Coherence:
{Definition of Stylistic Coherence}

## 3. Overall Assessment

{Definition of Overall Assessment}

## 4. Final json output:

{"Structural_Coherence": rating score, "Linguistic_Coherence": rating score, "Consistency": rating score, "Overall_Assessment": rating score}

# Review Output Format:

## 1. Review Summary: A brief paragraph (2-3 sentences)

## 2. Rating:

1) Structural Coherence:
- Rating: [Select one: 3 = Excellent, 2 = Moderate, 1 = Poor]
- Reason: Provide justification for the rating in 2-3 sentences.
2) Linguistic Coherence:
- Rating: [Select one: 3 = Excellent, 2 = Moderate, 1 = Poor]
- Reason: Provide justification for the rating in 2-3 sentences.
3) Consistency:
- Rating: [Select one: 3 = Excellent, 2 = Moderate, 1 = Poor]
- Reason: Provide justification for the rating in 2-3 sentences.

## 3. Overall Assessment:

- Rating: [Select one: 5 = Exceptional, 4 = Strong, 3 = Good, 2 = Needs Improvement, 1 = Major Revisions Necessary]
- Reason: Provide justification for the rating in 2-3 sentences. **## Example 1:{Positive Example}**

## Example 2: {Negative Example}

Input:

[Topic]

Messi's departure from the team

[Timeline]

2004-10-16: Barcelona announces Messi's departure, sparking concern among fans as Messi's next stop is a mystery.
2021-08-06: Barcelona officially announces Messi's departure, unable to renew his contract due to economic and structural obstacles.
2021-08-07: Messi held a press conference at Barcelona, bidding a tearful farewell and expressing his hope to return to Barcelona in the future, but the matter of joining Paris is still unresolved.
2023-06-08: Messi announces he is joining Major League Soccer's Miami International as Beckham's newest player.
2023-08-03: Lionel Messi puts in a superb performance, scoring twice to help Miami International beat Orlando City 3-1.

Output:

# 1. Review Summary

The timeline covers significant events in Lionel Messi's career, …,The timeline culminates with a highlight of Messi's exceptional performance in a match for his new team.

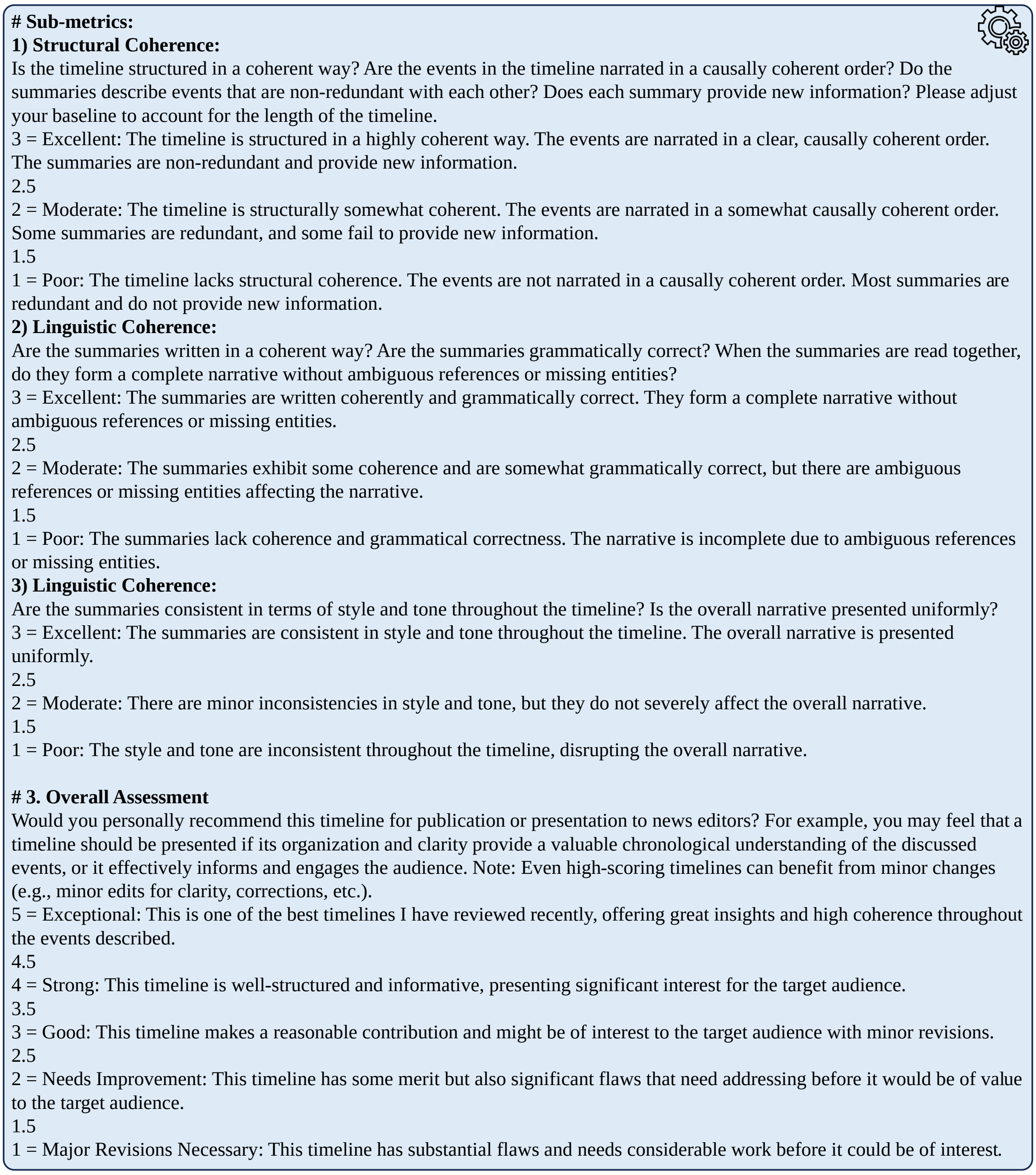
# 2. Rating:

1) Structural Coherence:
- Rating: 3 = Excellent
- Reason: The timeline is structured in a coherent and chronological order, … without redundancy and with each entry offering new information.
2) Linguistic Coherence:
- Rating: 3 = Excellent
- Reason: The summaries are well-written and grammatically correct, …, ensuring a complete and coherent storyline without missing entities or ambiguous references.
3) Stylistic Coherence:
- Rating: 3 = Excellent
- Reason: The style and tone remain consistent throughout the timeline, maintaining a formal yet engaging narrative approach …

# 3. Overall Assessment:

- Rating: 5 = Exceptional
- Reason: This timeline provides valuable insights into Lionel Messi's significant career moments, effectively capturing the key events that shaped his professional trajectory. The organization and clarity of the timeline offer a comprehensive understanding of Messi's journey, making it highly recommendable for publication or presentation to news editors without the need for major revisions.

Figure 6: Example of the Coherence Review Form.



# Sub-metrics:
1) Structural Coherence:
Is the timeline structured in a coherent way? Are the events in the timeline narrated in a causally coherent order? Do the summaries describe events that are non-redundant with each other? Does each summary provide new information? Please adjust your baseline to account for the length of the timeline.
3 = Excellent: The timeline is structured in a highly coherent way. The events are narrated in a clear, causally coherent order. The summaries are non-redundant and provide new information.
2.5
2 = Moderate: The timeline is structurally somewhat coherent. The events are narrated in a somewhat causally coherent order. Some summaries are redundant, and some fail to provide new information.
1.5
1 = Poor: The timeline lacks structural coherence. The events are not narrated in a causally coherent order. Most summaries are redundant and do not provide new information.
2) Linguistic Coherence:
Are the summaries written in a coherent way? Are the summaries grammatically correct? When the summaries are read together, do they form a complete narrative without ambiguous references or missing entities?
3 = Excellent: The summaries are written coherently and grammatically correct. They form a complete narrative without ambiguous references or missing entities.
2.5
2 = Moderate: The summaries exhibit some coherence and are somewhat grammatically correct, but there are ambiguous references or missing entities affecting the narrative.
1.5
1 = Poor: The summaries lack coherence and grammatical correctness. The narrative is incomplete due to ambiguous references or missing entities.
3) Linguistic Coherence:
Are the summaries consistent in terms of style and tone throughout the timeline? Is the overall narrative presented uniformly?
3 = Excellent: The summaries are consistent in style and tone throughout the timeline. The overall narrative is presented uniformly.
2.5
2 = Moderate: There are minor inconsistencies in style and tone, but they do not severely affect the overall narrative.
1.5
1 = Poor: The style and tone are inconsistent throughout the timeline, disrupting the overall narrative.

# 3. Overall Assessment
Would you personally recommend this timeline for publication or presentation to news editors? For example, you may feel that a timeline should be presented if its organization and clarity provide a valuable chronological understanding of the discussed events, or it effectively informs and engages the audience. Note: Even high-scoring timelines can benefit from minor changes (e.g., minor edits for clarity, corrections, etc.).
5 = Exceptional: This is one of the best timelines I have reviewed recently, offering great insights and high coherence throughout the events described.
4.5
4 = Strong: This timeline is well-structured and informative, presenting significant interest for the target audience.
3.5
3 = Good: This timeline makes a reasonable contribution and might be of interest to the target audience with minor revisions.
2.5
2 = Needs Improvement: This timeline has some merit but also significant flaws that need addressing before it would be of value to the target audience.
1.5
1 = Major Revisions Necessary: This timeline has substantial flaws and needs considerable work before it could be of interest.

Figure 7: Sub-metrics and overall assessment definition with their corresponding score criteria.

Metrics	Info	Granu	Fact	Coherence
Informativeness	**0.7874**	0.6128	0.6823	0.7004
Granular Consistency	0.6128	**0.7666**	0.6389	0.6795
Factuality	0.6823	0.6389	**0.9587**	0.7512
Coherence	0.7004	0.6795	0.7512	**0.9914**
AR-1	0.6213	0.5432	0.6589	0.7031
BERTScore	0.7032	0.5214	0.6317	0.6894
QAEval	0.7125	0.4987	0.9274	0.6342

Table 11: Correlation coefficients between existing metrics and human assessment results.

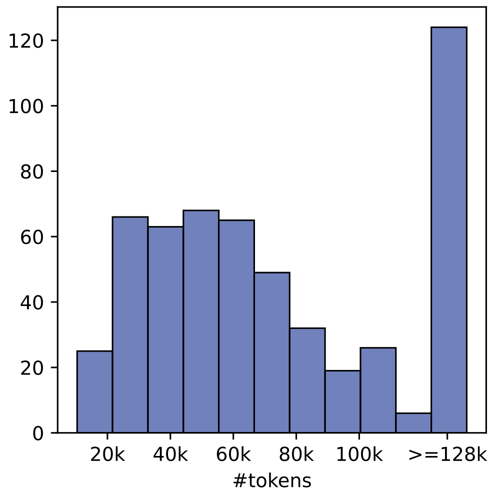


Figure 8: Token consumption histograms distribution for Long-context Prompting.

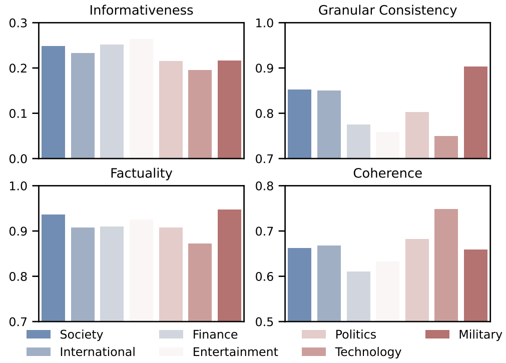


Figure 9: Topic types' influence on hierarchical merging GPT-3.5-Turbo.

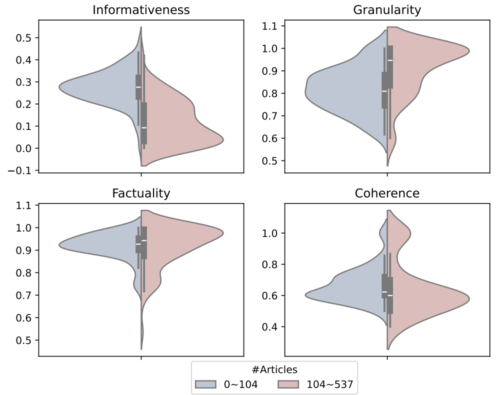
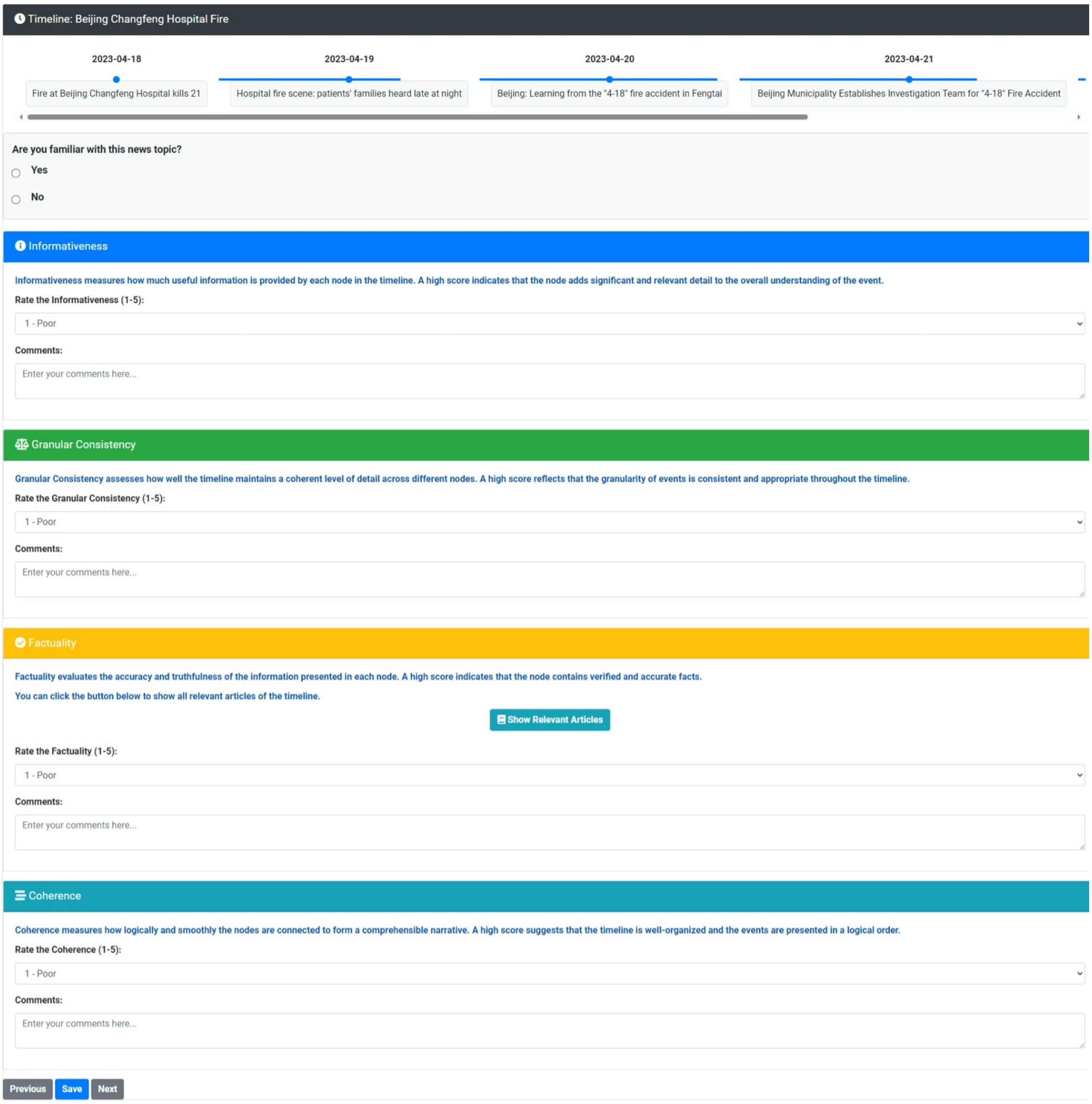


Figure 10: The influence of the number of news articles on evaluation metrics.



Timeline: Beijing Changfeng Hospital Fire

2023-04-18 — Fire at Beijing Changfeng Hospital kills 21

2023-04-19 — Hospital fire scene: patients' families heard late at night

2023-04-20 — Beijing: Learning from the "4-18" fire accident in Fengtai

2023-04-21 — Beijing Municipality Establishes Investigation Team for "4-18" Fire Accident

Are you familiar with this news topic?

○ Yes

○ No

Informativeness

Informativeness measures how much useful information is provided by each node in the timeline. A high score indicates that the node adds significant and relevant detail to the overall understanding of the event.

Rate the Informativeness (1-5):

1 - Poor

Comments:

Enter your comments here...

Granular Consistency

Granular Consistency assesses how well the timeline maintains a coherent level of detail across different nodes. A high score reflects that the granularity of events is consistent and appropriate throughout the timeline.

Rate the Granular Consistency (1-5):

1 - Poor

Comments:

Enter your comments here...

Factuality

Factuality evaluates the accuracy and truthfulness of the information presented in each node. A high score indicates that the node contains verified and accurate facts.

You can click the button below to show all relevant articles of the timeline.

Rate the Factuality (1-5):

1 - Poor

Comments:

Enter your comments here...

Coherence

Coherence measures how logically and smoothly the nodes are connected to form a comprehensible narrative. A high score suggests that the timeline is well-organized and the events are presented in a logical order.

Rate the Coherence (1-5):

1 - Poor

Comments:

Enter your comments here...

Figure 11: Human annotation scoring sheets of the proposed metrics.