CNTLS: A Benchmark Dataset for Abstractive and Extractive Chinese Timeline Summarization

Anonymous ACL submission

Abstract

Timeline summarization (TLS) involves creating summaries of long-running events by amalgamating dated summaries from multiple news articles. However, the scarcity of available data has considerably hindered the advancement of timeline summarization. In this paper, we introduce the CNTLS dataset, an open resource for Chinese timeline summarization. CNTLS comprises 77 real-life topics, each containing 2524 documents, and achieves an average compression of nearly 60% of the duration of all topics. We meticulously analyze the corpus using established metrics, focusing on the style of the summaries and the complexity of the summarization task. We rigorously assess the performance of various classic extraction TLS systems and substantiate the applicability of the large model approach for generative TLS systems on the CNTLS corpus, thereby furnishing benchmarks and fostering further research. To the best of our knowledge, CNTLS marks the inception of the first Chinese timeline summarization dataset. The dataset and source code are released [1].

1 Introduction

With the rapid growth of web services, there is a continuous surge in the daily publication of news articles, covering a wide range of events from around the world. This sheer volume of news articles can overwhelm readers, making it challenging to navigate through this deluge of information. To address this issue, it is necessary to develop techniques that help us tackle this huge amount of information. Timeline summarization (TLS) serves as a solution to the arduous manual summarization process, offering readers a faster and more comprehensive way to comprehend events from diverse viewpoints. TLS is a technique designed to automatically extract sentences that depict the chronological progression of a particular topic from a large collection of web articles. This approach has garnered considerable attention in recent years (Martschat et al., 2018; Ghalandari et al., 2020; Quatra et al., 2021; Liao et al., 2021; Yu et al., 2021; Mao et al., 2022; Faghihi et al., 2022; You et al., 2022).

As in most NLP tasks, the majority of timeline summarization datasets, including TL17 (Tran et al., 2013), Crisis (Tran et al., 2015b), and Entities (Ghalandari et al., 2020), are predominantly available in English. The absence of equivalent resources for other languages presents a challenge, constraining the potential influence of language technologies. Despite Chinese being the most spoken first language worldwide, used in 37 countries, it is often marginalized or omitted in timeline summarization corpora.

In this paper, our primary goal is to create a high-quality, large-scale corpus suitable for training automatic timeline summarization in the Chinese language. To achieve this, we systematically crawl Chinese timeline newspaper websites to compile annotation data. These websites present the information as HTML metadata alongside articles, serving as page descriptions for news media services and search engines. The resulting timeline newspaper summaries and articles offer a comprehensive depiction of timeline summarization practices across various news topics. Authored by professional writers and editors in Chinese newspapers, these summaries cover diverse subjects such as news, sports, entertainment, finance, and more.

The final CNTLS corpus comprises 77 topics, each with its respective timeline summaries, providing extensive coverage surpassing most other English datasets. This abundance of data supports the evaluation of diverse event timeline methods, reducing dependency on specific datasets and enhancing result robustness. The CNTLS corpus includes articles and summaries spanning politics, economics, sports, culture, and various journalistic subjects. CNTLS is a large-scale timeline summa-

[1] Code and data available at: *Accompanied ARR submission*.

rization dataset for the Chinese language.

To characterize the constructed corpus, we utilize four well-known metrics: extractive fragment coverage and density (Grusky et al., 2018), abstractivity$_p$ (Bommasani and Cardie, 2020), and novel n-grams (Kryscinski et al., 2018). For benchmarking purposes, we evaluate automatic timeline summarization systems on the CNTLS corpus, including two major classes of extractive systems (a clustering method of CLUST (Ghalandari et al., 2020) and a data-ranking method of DATEWISE (Ghalandari et al., 2020)). The ORACLE system is used to compute upper bounds for extractive timeline summarization performance. Given the current emergence of 'large models', we also validate the latest efficient generative pretrained large language models (ChatGLM (Du et al., 2022), Alpaca (Taori et al., 2023)) capable of handling long-text inputs during summarization.

In summary, our contributions in this work include: (i) Collecting a large and high-quality Chinese timeline summarization dataset from real-life news articles, namely CNTLS. (ii) Conducting an analysis of the corpus using well-known metrics in the timeline summarization field, focusing on the character of the summaries and the difficulty of the timeline summarization task. (iii) Evaluating the performance of extractive summarization systems on the CNTLS corpus for benchmarking purposes. Additionally, we explore the use of advanced large generative language models (fine-tuned from the original Meta's LLaMA or Stanford's Alpaca) for creating long timeline summaries.

2 Building the CNTLS Corpus

2.1 Timeline Summary Scraping

The CNTLS dataset is curated from web media metadata through a web-scale crawl spanning over 77 topics sourced from various online publishers. We utilize the HTML crawl tool `BeautifulSoup` [2] to extract HTML body content from specific timeline newspaper websites (houxu.app/, dsj365.cn/, etc.). These websites specialize in organizing news topics, titles, and timeline summaries, offering access to explicit metadata of timeline summaries. The Chinese newspaper summaries available on these platforms are authored by human writers, intended for general readership, and explicitly designed for summarization purposes.

Each topic, accompanied by its corresponding timeline summary texts, serves as annotation samples for creating timeline summaries. A topic is linked with a timeline sequence showcasing headlines with publication dates, with each headline having a corresponding summary. In short, we identify and leverage topics and their associated timeline summaries from the HTML metadata.

2.2 Source Article Extraction

Upon acquiring the timeline summary, we proceed to retrieve the original documents corresponding to the timeline summaries organized by dates, following the construction process outlined in the English timeline corpus (Tran et al., 2013; Yan et al., 2011). We manually define a set of keywords for each topic. Initially, we use the `HanNLP` tool [3] to extract keywords from news headlines in each topic after removing stop words. Subsequently, we use the obtained start and end times as search criteria, along with the keywords from the headlines, to search for relevant documents in the RING (Peng et al., 2021) news database, stored in HBase and indexed using Elastic Search. This approach aims to retrieve multiple news articles for each known publication time, forming a set of news documents corresponding to each timeline summary. These news documents serve as input for the summarization system, producing a time-ordered list along with its corresponding summary.

For each document, we utilize `PyLTP` (Che et al., 2020)[4] for Chinese text segmentation and we identify temporal expressions with `RecognizeTextDate`[5]. If a document contains a recognizable time expression, its date is assigned accordingly, prioritizing the first expression or setting it to the publication date if no specific expression is found.

3 Analysis of CNTLS Dataset

All dataset statistics are shown in Table 1. TL17 has longer (`A.L`=36) timelines than the other datasets, Crisis has a more long sequence of word tokens for each timeline (`A.DocL`=3,650,095), and Entities has longer date duration (`A.Duration`=4437).

[2] *https://www.crummy.com/software/BeautifulSoup/*

[3] *https://github.com/hankcs/HanLP*

[4] *https://github.com/HIT-SCIR/pyltp*

[5] *https://github.com/microsoft/Recognizers-Text*

Table 1: Dataset Statistics for the English datasets and our Chinese CNTLS dataset.

Dataset	Topics/	TLs/	A.L/	A.DocN/	A.Sent/	A.Token/	A.DocL/	A.DateT/	A.SumT/	A.K/	A.Duration/	A.DurComp/	A.DateComp/	A.DateCov/
TL17	9	19	36	467	20,409	945	441,346	990	4,561	2.9	212	41.35	0.45	0.81
Crisis	4	22	29	4,393	82,761	831	3,650,095	822	5,067	1.3	343	19.50	0.11	0.90
Entities	47	47	23	959	31,545	783	840,655	793	568	1.2	4,437	0.48	0.06	0.51
CNTLS	77	77	6	564	28,725	590	33,2361	1723	698	1.4	55	59.96	0.51	0.88

Table 2: Average values of the metrics in the Datasets.

Dataset	Coverage/	Density/	Compression/	Abstractivity (p=2)/	Novel 2-grams/	Novel 3-grams/	Novel 4-grams/
TL17	61.76	0.10	29.54	96.31	28.41	30.30	30.61
Crisis	41.57	0.08	16.00	97.26	19.71	20.08	20.15
Entities	87.00	0.72	17.90	99.97	59.48	74.23	90.84
CNTLS	30.96	2.56	41.49	94.88	77.63	82.60	83.53

In contrast to other English datasets, the CNTLS dataset distinguishes itself with a significantly larger number of topics (Topics/). The average number of word tokens in input articles for each date (A.DateT/) is also notably larger than the other three English datasets. Additionally, despite CNTLS having a relatively short timeline length (A.L/), its significant compression ratio of time duration (A.DurComp/) aligns closely with real-world scenarios. Besides, this distinction sets it apart by offering a wealth of events and timelines, providing ample data for evaluating the generalizability of various timeline strategies and reducing the potential for results to be influenced by specific data.

Besides, as shown in Table 2, CNTLS exhibits the highest compression value (41.99), implying that the summarization system must compress a greater amount of original information to generate the summary during the timeline summarization process. Moreover, the number of novel tokens in the CNTLS dataset is also among the highest (novel 2-grams: 77.63 and novel 3-grams: 82.60) across several English datasets.

We also use density, coverage, and compression to understand the data distribution of the constructed corpus, as shown in Figure 1. We showcase the distribution of the samples by combining the values of 'abstractivity$_p$' (p=2) and novel 2-grams. These graphical representations visually convey the degree of abstractivity in the summaries within the CNTLS corpus: (i) The plots for 'Density and coverage distributions' show a positive correlation with extractivity, with higher extractivity concentrated in the partition positioned around the upper right corner. (ii) The plots illustrating 'abstractivity$_p$ and novel 2-grams distributions' demonstrate a positive correlation with abstractivity, with distributions centred closer to the upper right corner, indicating highly abstractive summaries. (iii) A higher compression ratio c increases the difficulty of the summarization task, requiring the model to accurately capture crucial aspects or events from the original text to condense into a concise and informative summary.

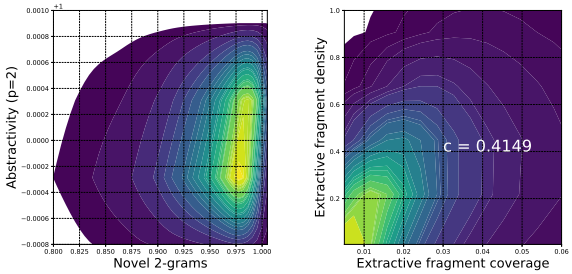


Figure 1: Abstractivity$_p$ (p=2) and novel 2-gram distributions on CNTLS datasets (left subfigure). Density and coverage distributions of extractive compression scores on four datasets (right subfigure). Each box represents a normalized bivariate density plot and the plot shows the median compression ratio c between summaries and source text.

Based on the distribution analysis of density, coverage, and compression, it can be concluded that CNTLS encompasses a broader spectrum of summarization styles, demonstrating significant summary diversity. The analysis of English datasets is available in the Appendix 11.1.

4 Timeline Summarization Systems

We assess several extractive summarization systems to comprehend the challenges posed by the CNTLS dataset. This includes evaluating conventional extractive models and an extractive ORA-

Table 3: Results of the extractive and generative TLS methods on the constructed CNTLS datasets.

Methods	Concat		Agree		Align+ m:1		Date F1
	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2	
Oracle Date	.460	.190	.330	.159	.039	.153	1.0
Oracle Text	.454	.240	.309	.145	.323	.157	.993
Oracle Full	.440	.230	.320	.160	.318	.160	1.0
CLUST	.224	.073	.031	.004	.035	.006	.326
DATAWISE$^{\texttt{Textrank}}$	.359	.136	.154	.069	.174	.095	.605
DATAWISE$^{\texttt{Centro-rank}}$	.383	.152	.190	.090	.196	.102	.605
DATAWISE$^{\texttt{Centro-opt}}$	.396	.157	.184	.089	.206	.104	.605
Chinese-Alpaca-2-7B-4K	.020	.006	.015	.004	.016	.004	.605
ChatGLM-6B-2K	.262	.052	.093	.022	.103	.024	.605
ChatGLM2-6B-8K	.248	.052	.103	.024	.112	.025	.605
ChatGLM2-6B-32K	.268	.067	.111	.034	.120	.035	.605

CLE, providing an upper bound for extractive performance within the corpus.

Additionally, we explore the use of Large Language Models (LLMs) in timeline summarization. For the generative timeline summarization model, we choose LLMs capable of handling lengthy text sequences. The average token sequence length for CNTLS is 554k, whereas most large models have a maximum sequence length of approximately 32K. This limitation renders the direct input of the complete token sequence unfeasible. Therefore, we employ regression methods for date prediction, feeding each time point's respective document into the large model for summary generation.

4.1 Experimental Results

According to our experimental results in Table 3, notably fluctuates across extractive and abstractive methods within datasets. Notably, the ORACLE consistently outperforms other systems, highlighting the significant impact of accurate time prediction on timeline summarization performance.

In both extractive and generative methods, we employ the optimal regression time prediction method to ensure a fair comparison between extraction and generative approaches. As observed in Tables 3, the efficacy of partial extractive methods surpasses that of generative Large Language Models (LLMs), such as ChatGLM2-7B. Among the generative LLMs used in our experiments, ChatGLM2-6B-32K exhibits slightly higher ROUGE-1 and ROUGE-2 scores across all prediction methods compared to other generative LLMs. This suggests that enlarging the total number of input tokens can be more beneficial for our timeline summarization task (from ChatGLM2-6B-2K to ChatGLM2-6B-32K). However, none of the models achieve particularly high ROUGE scores compared to extractive methods. In essence, the enhancements in extractive methods are still notable, surpassing the performance of large and long-context generative models.

Possible reasons include CNTLS having a high compression rate (41.49, as shown in Table 2), which generally makes summarization more challenging. Additionally, longer topic durations (`A.Duration`=4,437, as shown in Table 1) may impact the performance of generative methods. As previously mentioned, our efforts to create a large model supporting longer contextual inputs result in marginal improvements when increasing the sequence length from 2k to 32k. However, there is no significant difference. The current method for generating large models demonstrates inherent limitations when dealing with lengthy input sequences, leading to a failure to comprehend events across extensive document sequences. This implies the potential for improving generative LLMs in summarizing lengthy timelines more effectively.

5 Conclusions

We introduce CNTLS, a dataset comprising articles and their timeline summaries authored by online publication editors. In comparison to existing datasets, ours offers a greater number of topics and a diverse set of summaries. We conduct benchmark evaluations using prominent extractive frameworks. Additionally, we explore the integration of LLMs in timeline summarization. Our proposed timeline systems, along with the corresponding analysis of LLMs-based abstractive strategies, open new avenues for assessing the challenges of timeline summarization tasks and for advancing future summarization models.

6 Broader Impacts & Limitations

This study contributes a timeline summary corpus for the Chinese research community, offering practical applications in news content organization, topic detection, and event tracking.

In selecting large models, we have prioritized those capable of handling long sequences. However, this does not fully resolve the challenge of incorporating all documents input related to a topic. While there are currently limitations, the emergence of new large models supporting longer sequence inputs, such as Baichuan2-192K (Baichuan, 2023) [6] supporting 192K tokens and Claude 2.1 (Bedrock, 2023) [7] supporting 200K tokens, could be explored for their suitability in TLS tasks. Our exploration is constrained by the limitations of a single NVIDIA Tesla V100 32GB GPU, which undoubtedly restricts the length of input text and the cost of fine-tuning large language models.

Additionally, more sophisticated processes could be beneficial as been proved by Pratapa et al. (2023). For example, incorporating topic relevance planning (Wang et al., 2023) or introducing event background (Pratapa et al., 2023) or preference (Ye and Simpson, 2023) can help eliminate redundant information. Additionally, topic relevance instructions (Koike et al., 2023) can help narrow down the search scope of documents.

Besides, LLMs are known for having hallucination issues (Ji et al., 2023), which are prone to generating incorrect facts and leading to wrong summaries. New metrics or tools capable of concurrently assessing both factual accuracy and temporal accuracy, leveraging existing hallucination evaluation tools (Chern et al., 2023; Peng et al., 2023; Manakul et al., 2023; Gekhman et al., 2023) for large language models, are necessary for generating TLS tasks.

As our primary focus is the creation of a Chinese TLS dataset, these avenues for improvement could be elaborated upon in future works.

7 Ethical Statement

Data Availability and Safety. The summarization data analyzed in this paper are primarily publicly accessible; otherwise, we will provide links upon request for access. While filtering has been implemented in compiling the original datasets, some content may contain sensitive descriptions, such as news coverage of violent crimes and events. Furthermore, certain news articles may divulge details such as the identities of individuals involved, which are publicly accessible information shared by the news outlets. This aspect can be valuable for assessing the factual accuracy of generative methods. Hence, we have refrained from anonymizing this information. Our dataset does not include any protected information (e.g., sexual orientation or political views under GDPR).

Usage of Large PLM. The GPT-3.5 model or its variants are used to generate text (summaries) from input documents in summarization tasks. This generated text is exclusively employed for experiments and analysis, as outlined in the corresponding sections. The paper does not engage in any additional utilization, such as generating content for manuscripts, using GPT-3.5 or its derivatives.

Human Evaluation. We perform human evaluation with the assistance of a single judge, who holds a postgraduate degree in AI or Computer Science from China and possesses extensive experience in evaluating summarization tasks.

Authors Conflicts. The authors declare that we have no conflicts of interest. Informed consent is obtained from all individual participants.

[6] https://top.aibase.com/tool/baichuandamoxing

[7] https://www.anthropic.com/news/claude-2-1

References

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of news topics. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 10–18. ACM.

Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2009. Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM*, pages 97–106. ACM.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Amazon Bedrock. 2023. Introducing claude 2.1.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 8075–8096. Association for Computational Linguistics.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2020. N-LTP: A open-source neural chinese language technology platform with pretrained models. *CoRR*, abs/2009.11616.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.

Hai Leong Chieu and Yoong Keok Lee. 2004. Query based event extraction along a timeline. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pages 425–432.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *CoRR*, abs/2304.08177.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 320–335. Association for Computational Linguistics.

Yijun Duan, Adam Jatowt, and Masatoshi Yoshikawa. 2020. Comparative timeline summarization via dynamic affinity-preserving random walk. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1778–1785. IOS Press.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Hossein Rajaby Faghihi, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel R. 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*, pages 5455–5477. Association for Computational Linguistics.

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, EMNLP*, pages 2053–2070. 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, NFiS@EMNLP 2017*, pages 85–90. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Georgiana Ifrim, 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, ACL, Volume 1 (Long Papers)*, pages 1322–1334. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2023. How you prompt matters! even task-oriented constraints in instructions affect llm-generated text detection. *CoRR*, abs/2311.08369.

Wojciech Kryscinski, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1808–1817. Association for Computational Linguistics.

Jiwei Li and Sujian Li. 2013. Evolutionary hierarchical dirichlet process for timeline summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL, Volume 2: Short Papers*, pages 556–560. The Association for Computer Linguistics.

Yiming Liao, Shuguang Wang, and Dongwon Lee. 2021. WILSON: A divide and conquer approach for fast and effective news timeline summarization. In *Proceedings of the 24th International Conference on Extending Database Technology, EDBT*, pages 635–645. OpenProceedings.org.

Hui Lin and Jeff A. Bilmes. 2011. A class of submodular functions for document summarization. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, ACL, Volume 1 (Long Papers)*, pages 510–520. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 9004–9017. Association for Computational Linguistics.

Qianren Mao, Jianxin Li, JiaZheng Wang, Xi Li, Peng Hao, Lihong Wang, and Zheng Wang. 2022. Explicitly modeling importance and coherence for timeline summarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 8062–8066. IEEE.

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, EACL, Volume 2: Short Papers*, pages 285–290. Association for Computational Linguistics.

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

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

OpenAI. 2022. Introducing chatgpt.

Arian Pasquali, Ricardo Campos, Alexandre Ribeiro, Brenda Salenave Santana, Alípio Jorge, and Adam Jatowt. 2021. Tls-covid19: A new annotated corpus for timeline summarization. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR*, volume 12656 of *Lecture Notes in Computer Science*, pages 497–512. Springer.

Arian Pasquali, Vítor Mangaravite, Ricardo Campos, Alípio Mário Jorge, and Adam Jatowt. 2019. Interactive system for automatically generating temporal narratives. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR*, volume 11438 of *Lecture Notes in Computer Science*, pages 251–255. Springer.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813.

Hao Peng, Jianxin Li, Yangqiu Song, Renyu Yang, Rajiv Ranjan, Philip S. Yu, and Lifang He. 2021. Streaming social event detection and evolution discovery in heterogeneous information networks. *ACM Trans. Knowl. Discov. Data*, 15(5):89:1–89:33.

Adithya Pratapa, Kevin Small, and Markus Dreyer. 2023. Background summarization of event timelines. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 8111–8136. Association for Computational Linguistics.

Moreno La Quatra, Luca Cagliero, Elena Baralis, Alberto Messina, and Maurizio Montagnuolo. 2021. Summarize dates first: A paradigm shift in timeline summarization. In *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval,SIGIR*, pages 418–427. ACM.

Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manag.*, 40(6):919–938.

Encarnación Segarra Soriano, Vicent Ahuir, Lluís-F. Hurtado, and José González. 2022. DACSA: A large-scale dataset for automatic summarization of catalan and spanish newspaper articles. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, pages 5931–5943. Association for Computational Linguistics.

Russell C. Swan and James Allan. 2000. Automatic generation of overview timelines. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56. ACM.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Giang Binh Tran, Mohammad Alrifai, and Eelco Herder. 2015a. Timeline summarization from relevant headlines. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR*, volume 9022 of *Lecture Notes in Computer Science*, pages 245–256.

Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. 2013. Predicting relevant news events for timeline summaries. In *22nd International World Wide Web Conference, WWW*, pages 91–92. International World Wide Web Conferences Steering Committee / ACM.

Giang Binh Tran, Eelco Herder, and Katja Markert. 2015b. 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 of the Asian Federation of Natural Language Processing, ACL, Volume 1 (Long Papers)*, pages 1598–1607. Association for Computational Linguistics.

Tuan Tran, Claudia Niederée, Nattiya Kanhabua, Ujwal Gadiraju, and Avishek Anand. 2015c. Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM*, pages 1201–1210. ACM.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2609–2634. 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 *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*, pages 58–68. The Association for Computational Linguistics.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. 2011. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 433–443. ACL.

Yuxuan Ye and Edwin Simpson. 2023. Towards abstractive timeline summarisation using preference-based reinforcement learning. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 2882–2889. IOS Press.

Jingyi You, Dongyuan Li, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2022. Joint learning-based heterogeneous graph attention network for timeline summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, pages 4091–4104. 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 *ACL/IJCNLP, (Volume 1: Long Papers)*, pages 377–387. Association for Computational Linguistics.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR*, pages 1949–1952. ACM.

Wayne Xin Zhao, Yanwei Guo, Rui Yan, Yulan He, and Xiaoming Li. 2013. Timeline generation with social attention. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR*, pages 1061–1064. ACM.

8 Background

8.1 Related Work of TLS

Since the inception of timeline summarization (Swan and Allan, 2000; Allan et al., 2001), this field has garnered considerable attention over the years (Alonso et al., 2009; Yan et al., 2011; Zhao et al., 2013; Li and Li, 2013; Tran et al., 2015a; Wang et al., 2016; Pasquali et al., 2021).

To put it succinctly, the evolution of representative methods involves transitioning to either event clustering (Alonso et al., 2009; Tran et al., 2015c; Pasquali et al., 2019; Zhao et al., 2020; Duan et al., 2020; Ghalandari et al., 2020) or sentence ranking (Radev et al., 2004; Lin and Bilmes, 2011; Nguyen et al., 2014; Ghalandari, 2017; Ghalandari et al., 2020; Mao et al., 2022) to select the optimal sentence. Chieu and Lee (2004) construct timelines by directly selecting the top-ranked sentences based on similarities within sentences and Li and Li (2013) select dates then extract sentences corresponding to the dates. Nguyen et al. (2014) propose a pipeline for generating timelines, involving date selection, sentence clustering, and sentence ranking. More recently, Martschat et al. (2018) have adapted a submodular function model for the TLS task, originally used for multi-document summarization (MDS). Furthermore, Ghalandari et al. (2020) examine various TLS strategies and categorize TLS frameworks into three types: direct summarization approaches, date-wise approaches, and event detection approaches.

8.2 Existing TLS Datasets

There are several frequently used timeline summarization datasets: TL17 (Tran et al., 2013), Crisis (Tran et al., 2015b), and Entities (Ghalandari et al., 2020). These datasets contain human-written timelines on specific topics, with source news articles retrieved from the web at a given point in time. Each dataset comprises journalist-generated timelines from major news media such as CNN, BBC, and Reuters, along with a corresponding corpus of articles per topic (e.g., H1N1 flu, Enron bankruptcy, and Egypt war).

Specifically, the number of topics and their time spans varies. TL17 contains 19 timelines from 9 topics, while Crisis involves 22 timelines from 4 topics. An overview of the existing English datasets is shown in Table 4.

Table 4: The Meaning of Abbreviated Symbols.

Abbr	Implication
Topics/	*the number of topics of the dataset.*
TLs/	*the number of ground truth timelines of* Topics.
A.L/	*the average number of daily sentences of* Topics.
A.DocN/	*the average number of files for the single one of* Topics.
A.Sent/	*the average number of sentences in source articles for the single one of* Topics.
A.Token/	*the average number of word tokens for each source articles*
A.DocL/	*the average number of word tokens of all dates mentioned sentences for the single one of* Topics.
A.DateT/	*the average number of word tokens of all source articles for each date*
A.SumT/	*the average number of word tokens for the single one of* Topics.
A.K/	*the average number of sentences for the single one of* TLs.
A.Duration/	*the average number of days experienced from the beginning to the end of* TLs.
A.DurComp/	*the compression ratio w.r.t. timeline length is divided by duration.*
A.DateComp/	*the compression ratio w.r.t. dates is divided by the total number of dates mentioned in articles.*
A.DateCov/	*the average coverage of dates in the ground truth timeline by the news in articles collection.*

Table 5: Human evaluation criteria, adapted from Fabbri et al. (2021).

Informativeness	*Q: How well does the summary capture the main points of the meeting segment?* *NOTA: A good summary should contain all and only the important information of the source.*
Factuality	*Q: Are the facts provided by the summary consistent with facts in the meeting segment?* *NOTA: A good summary should reproduce all facts accurately and not make up untrue information.*
Fluency	*Q: Consider the individual sentences of the summary, are they well-written and grammatical?* *NOTA: A good summary should have proper grammar, punctuation, and sentence structure.*
Coherence	*Q: Consider the summary as a whole, does the content fit together and sound natural?* *NOTA: A good summary should not just be a collection of related information, but should build from sentence to sentence to a coherent body of information about a topic.*
Redundancy	*Q: Does the summary contain redundant content?* *NOTA: A good summary should not have unnecessary word or phrase repetitions in a sentence or semantically similar sentences.*

9 Evaluation Metrics

9.1 Metrics of Characterizing TLS Dataset

We analyze the abstractivity of the timeline summarization corpus using established metrics from prior works (Grusky et al., 2018; Soriano et al., 2022) in dataset construction. These metrics gauge abstractivity by measuring the extent of text overlap between the summary and the article. Specifically, we employ the following metrics: Extractive Fragment Coverage and Density, Abstractivity$_p$ and novel n-grams.

Extractive Fragment Coverage (Grusky et al., 2018): This metric quantifies the extent to which a summary is derived from the text, indicating the percentage of summary words belonging to extractive fragments of the article.

Extractive Fragment Density (Grusky et al., 2018): In contrast to coverage, density considers the length of extractive fragments. While high coverage may result from numerous individual words in the summary, low density suggests short extractive fragments.

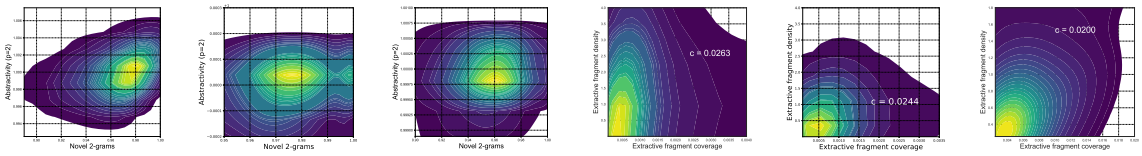
Compression Ratio (Grusky et al., 2018): This ratio measures the word ratio between the article and the summary. Summarizing with higher compression presents challenges in capturing critical aspects more precisely.

Abstractivity$_p$ (Bommasani and Cardie, 2020): This metric quantifies abstractivity by assessing the overlap between the summary and the original text. Higher values indicate reduced overlap, and the parameter p assigns weight to the length of each extractive fragment.

Novel n-grams (Kryscinski et al., 2018): This metric quantifies n-grams introduced in the summary but absent in the original text. We explore novel n-grams without considering the generated timeline summarization, illustrating the intrinsic novel properties of datasets. The metric's value is expressed as a percentage of the total number of n-grams in the summary.

9.2 Metrics of Evaluating TLS Systems

Access to extensive and high-quality data is a fundamental prerequisite for significant advancements in the field of summarization. Particularly, timeline summarization presents even greater challenges due to the inclusion of informal and spoken expressions, frequent topic shifts, multiple participants, and extended context. This complexity makes the creation of large-scale, high-quality datasets for



(a) TL17 (b) Crisis (c) Entities (d) TL17 (e) Crisis (f) Entities

Figure 2: Abstractivity$_p$ (p=2) and novel 2-grams distributions on English datasets (subfigures (a), (b), (c)). Density and coverage distributions of extractive compression scores on four datasets (subfigures (d), (e), (f)). Each box represents a normalized bivariate density plot and each plot shows the median compression ratio c between summaries and source text.

training neural summarization models a formidable task. The essence of a timeline summary task lies in consolidating information from a vast array of documents or topics. Consequently, the creation of an authentic timeline dataset necessitates incorporating a diverse spectrum of noteworthy news events. Generating thousands of manual summaries entails considerable human effort and ingenuity.

We assess these extractive and generative summarization systems using four proprietary summarization metrics (Martschat and Markert, 2017), widely recognized for evaluating timeline summary performance. These metrics include concatenation-based Rouge F1 (Concat R1 or R2), date-agreement Rouge F1 (Agree R1 or R2), alignment-based Rouge F1 (align R1 or R2), and Date F1 score. These metrics assess the concatenation of daily summaries, consideration of matching days, and alignment based on date and content similarity. Date selection is evaluated using the F1 score.

The extractive ORACLE is obtained through a greedy approach. We assess Date-ORACLE, Text-ORACLE, and Full-ORACLE for extractive summarization systems. Date-ORACLE selects the correct (Ground-truth) dates and employs CENTROID-OPT for daily summarization. Text-ORACLE uses regression to select dates and constructs a summary for each date by optimizing the ROUGE score with the ground-truth summaries. Full-ORACLE selects the correct dates and generates a summary for each date by optimizing the ROUGE score with the ground-truth summaries.

10 TLS Methods

10.1 Extractive systems

Event-clust summarization approaches: Clust (Ghalandari et al., 2020) uses DATEMENTIONCOUNT (Ghalandari et al., 2020) [8] to rank clusters. It then employs CENTROID-OPT (Radev et al., 2004) to rank sentences based on their similarity to the centroid of all sentences for timeline summarization.

Date-wise summarization approaches: Date-wise (Ghalandari et al., 2020) employs supervised date selection, PM-MEAN (Ghalandari et al., 2020) for candidate sentence selection, and CENTROID-OPT (Radev et al., 2004) to rank sentences based on their similarity to the centroid of all sentences for timeline summarization.

10.2 Generative systems

Alpaca for long text summarization: The Chinese-Alpaca-2-7B-4K (Cui et al., 2023) is expanded on the original Alpaca (Taori et al., 2023) incorporating Chinese vocabulary and undergoing additional pre-training with Chinese data to enhance its foundational semantic understanding in Chinese. The model has a context length of 4K with position interpolation, making it suitable for text summarization applications. The implementation of Chinese-Alpaca for summarization inference is done using the Transformer [9].

ChatGLM for long text summarization: ChatGLM-6B-2K[10] and ChatGLM2-6B-2K[11] are open bilingual language models based on General Language Model (GLM) (Du et al., 2022) and use LLMs' technology similar to ChatGPT (OpenAI, 2022),. They are optimized for Chinese Q&A, dialogue and summarization. The implementation of ChatGLM for summarization inference is achieved using the `HuggingFaceHub`[12]. ChatGLM2-6B-32K further enhances its ability to understand long texts compared to ChatGLM2-6B,

[8] DATEMENTIONCOUNT: Rank by how often the cluster date is mentioned throughout the input collection.

[9] https://github.com/huggingface/transformers

[10] https://github.com/THUDM/ChatGLM-6B

[11] https://github.com/THUDM/ChatGLM2-6B

[12] https://huggingface.co/THUDM/chatglm-6b

Table 6: Human evaluation results of software extractive and generative TLS systems.

Criterion	Extractive	Generative
	DATAWISE$^{\texttt{Centro-opt}}$	ChatGLM2-6B-32K
Informativeness	3.22±1.10	3.31±1.13
Factuality	4.41±1.23	3.19±1.15
Fluency	3.41±1.09	3.33±1.25
Coherence	3.16±1.17	3.12±1.09
Redundancy	3.30±1.41	4.21±1.31
Average Score	3.50±1.22	3.43±1.28

allowing it to handle up to a 32K context length.

11 TLS Performances

11.1 Distribution Analysis

Figure 2 illustrates that Crisis and TL17 datasets predominantly prefer abstractive summaries, with the distribution tending to decrease and shift left.

In subplots (a-c) in Figure 2 and the comparison of our CNTLS's Abstractivity and novel 2-gram distributions in Figure 1, it is evident that TL17 exhibits a relatively high density (y-axis), suggesting that TL17 summaries are prone to containing long extractive fragments. In subplots (d-f) of Figure 2 and comparing our CNTLS's compression score in Figure 1, it is apparent that our dataset obtains a stronger compression ratio, indicating higher semantic abstraction of the corpus. This presents greater challenges for both extractive and generative methods. Extractive methods must identify more critical daily summary sentences, while generative methods need to excel in condensing the core essence of the article and generating content more relevant to each sub-event within the topic.

11.2 Human Evaluation

We evaluate the performance of state-of-the-art extractive and generative TLS systems, including DATAWISE$^{\texttt{Centro-opt}}$ and ChatGLM2-6B-32K. A total of three workers from the China colleges and universities were in our evaluations, including pilot annotations. A 5-point Likert scale is used to evaluate each criterion.

Workers read all documents for three randomly selected topics, typically within 30 minutes, and then evaluated the quality of each system-generated timeline summary based on five criteria: informativeness, factuality, fluency, coherence, and redundancy. These criteria are outlined in Table 5. Importantly, the summaries are presented in the order of daily summaries.

In Table 6, we present the performance of summarization systems. The scores are then averaged, and the standard deviation is also reported. We observed that the results of the extractive systems indeed outperformed those of the generative systems. However, compared to the disparity presented in Table 3, the results of the generative systems evaluated by human judgment did not show explicit differences from the extractive summarization systems, especially in terms of informativeness. It is noteworthy that the results of the generative systems were superior to those of the extractive systems in terms of redundancy, reflecting the advantage of large-model generative methods.

11.3 Case Studies

The case study findings, illustrated in Figure 7 & 8, reveal that large models generally maintain readability and factual accuracy (marked with ☑). The generative model demonstrates superior performance in capturing very fine-grained facts.

However, the generative results of ChatGLM-6B-2K reveal missed facets, suggesting a limitation in capturing key content of the topics. Conversely, the other two large models exhibit a higher success rate in generating relevant facets. Nevertheless, it is noteworthy that large models also encounter issues with factual errors (marked with ☒). In the second case involves two facts: 'Police found the main suspect facing charges in up to 12 cases' and 'Prosecutors detained the main suspect with 14 offences', where '12' and '14' can easily be confused.

We have observed that the generation results of ChatGLM2-6B, are comparable to those of the extractive ORACLE in terms of the informativeness of the generated content, particularly in regard to readability. Particularly, large-model generative methods exhibit superior coherence among sentences, underscoring their advantage. Combining the generation cases with the evaluation in Table 6, we posit that traditional TLS evaluation metrics may not fully reflect the real performance of generative large models (with low values presented in Table 3). Developing tailored evaluation methods for generative approaches in TLS tasks is crucial. This highlights the essence of devising evaluation metrics suited for assessing generative LLM methods in timeline summarization.

Table 7: A case study is conducted on two topics using generative LLMs, showing daily summaries from the first and last dates. The generation of each method is annotated, with '[Not. m]' indicating no involvement, '☒' indicating involvement with incorrect facts generated, and '☑' indicating successful generation.

Topic of 'Fengchao Express Container Charging Event'			
ORACLE (From CNTLS)	**ChatGLM-6B-2K**	**ChatGLM2-6B-8K**	**ChatGLM2-6B-32K**
2020.04.30 丰巢快递柜自4月30日开始推出会员制。据媒体报道丰巢科技于2020年4月30日上线会员服务其中，针对消费者滞留快件现象，明确非会员用户将享受12小时的免费保管，超时后收费0.5元/12小时，3元封顶。丰巢方面表示，对相关服务约定，消费者可在线上进行选择，未经消费者同意，将不产生任何费用.....自丰巢快递柜4月30日推出会员制后，浙江杭州东新园小区业委会因此发布通知暂时停用。	2020.04.30 丰巢快递柜自4月30日起开始对非会员用户收取超时费用，每超时1小时收取0.5元，最高封顶3元。此前，丰巢曾宣布自4月30日起推出会员制服务，但业界认为其超时收费和提前收费的问题仍然存在。丰巢方面表示，其已经调整了超时收费的政策，但需要消费者自行购买会员卡。	2020.04.30 丰巢快递柜自4月30日开始推出会员制，针对消费者滞留快件现象，非会员用户将享受12小时的免费保管，超时后收费0.5元/12小时，3元封顶。丰巢方面表示，消费者可在线上进行选择，未经消费者同意，将不产生任何费用。	2020.04.30 丰巢快递柜自4月30日开始推出会员制，针对非会员用户，将享受12小时的免费保管，超时后收费0.5元/12小时，最高3元封顶。南都此前报道，自丰巢快递柜4月30日推出会员制后，浙江杭州东新园小区业委会因此发布通知暂时停用。
Key Facets Analysis of Daily Summary in 2020.04.30			
丰巢快递柜推出会员制 Fengchao express lockers now offer membership 非会员享受12小时免费保管 Non-members enjoy 12 hours of free storage. 收费0.5元/12小时，3元封顶 Charge 0.5 yuan/12 hours, capped at 3 yuan 小区业委会暂停使用 Residential committee suspends use	丰巢快递柜推出会员制 [Not.m] Fengchao express lockers now offer membership 非会员享受12小时免费保管 [Not.m] Non-members enjoy 12 hours of free storage. 收费0.5元/12小时，3元封顶 ☑ Charge 0.5 yuan/12 hours, capped at 3 yuan 小区业委会暂停使用 [Not.m] Residential committee suspends use	丰巢快递柜推出会员制 ☑ Fengchao express lockers now offer membership 非会员享受12小时免费保管 ☑ Non-members enjoy 12 hours of free storage. 收费0.5元/12小时，3元封顶 ☑ Charge 0.5 yuan/12 hours, capped at 3 yuan 小区业委会暂停使用 [Not.m] Residential committee suspends use	丰巢快递柜推出会员制 ☑ Fengchao express lockers now offer membership 非会员享受12小时免费保管 ☑ Non-members enjoy 12 hours of free storage. 收费0.5元/12小时，3元封顶 ☑ Charge 0.5 yuan/12 hours, capped at 3 yuan 小区业委会暂停使用 ☑ Residential committee suspends use
...	...	...	...
2020.05.09 对于杭州东新园小区“断电”丰巢快递柜一事，5月9日下午，丰巢公司正式反击：业委会“停机”是严重的违约行为，已造成巨大的经济和商誉损失，公司将依法追责.....有媒体统计，继杭州东新园小区，5月7日起暂停使用丰巢智能快递柜后，上海多个小区也开始联合抵制，截至5月9日，已有78个小区业委会向丰巢超时收费说“不”。	2020.05.09 丰巢公司针对杭州东新园小区“断电”事件进行了反击，并表示将追责。丰巢公司发布了一封致用户的公开信，解释了停用快递柜的原因，并推出鼓励尽早取件的红包政策。此外，丰巢公司还回应了关于收取超时保管费的问题，并表示将完善会员制度。各地监管部门也陆续发声，回应快递柜收费相关话题。丰巢公司表示，业委会单方面断电已构成违约，已造成巨大经济损失。	2020.05.09 丰巢公司针对杭州东新园小区“断电”事件进行了反击，并表示将追责。丰巢公司发布了一封致用户的公开信，解释了停用快递柜的原因，并推出鼓励尽早取件的红包政策。此外，丰巢公司还回应了关于收取超时保管费的问题，并表示将完善会员制度。各地监管部门也陆续发声，回应快递柜收费相关话题。丰巢公司表示，业委会单方面断电已构成违约，已造成巨大经济损失。	2020.05.09 丰巢公司针对杭州东新园小区“断电”丰巢快递柜事件的反击行动，引发社会关注。公司表示已造成巨大的经济和商誉损失，将依法追责。同时，公司推出鼓励尽早取件的红包政策。此外，上海多个小区也开始联合抵制丰巢快递柜，并有多个监管部门发声回应快递柜收费问题。丰巢公司也发布了一封致用户公开信回应，并解释了原因。
Key Facets Analysis of Daily Summary in 2020.05.09			
丰巢公司正式反击 Fengchao company counterattacks officially 已造成巨大的经济和商誉损失 facing substantial economic and reputational losses. 公司将依法追责 Legal actions will be pursued 上海多个小区也开始联合抵制 Several communities in Shanghai joining the boycott	丰巢公司正式反击 ☑ Fengchao company counterattacks officially 已造成巨大的经济和商誉损失 ☑ facing substantial economic and reputational losses. 公司将依法追责 [Not.m] Legal actions will be pursued 上海多个小区也开始联合抵制 [Not.m] Several communities in Shanghai joining the boycott	丰巢公司正式反击 ☑ Fengchao company counterattacks officially 已造成巨大的经济和商誉损失 ☑ facing substantial economic and reputational losses. 公司将依法追责 [Not.m] Legal actions will be pursued 上海多个小区也开始联合抵制 [Not.m] Several communities in Shanghai joining the boycott	丰巢公司正式反击 ☑ Fengchao company counterattacks officially 已造成巨大的经济和商誉损失 ☑ facing substantial economic and reputational losses. 公司将依法追责 [Not.m] Legal actions will be pursued 上海多个小区也开始联合抵制 ☑ Several communities in Shanghai joining the boycott
...	...	...	...

Table 8: A case study is conducted on two topics using generative LLMs, showing daily summaries from the first and last dates. The generation of each method is annotated, with '[Not. m]' indicating no involvement, '☒' indicating involvement with incorrect facts generated, and '☑' indicating successful generation.

Topic of 'South Korean Nth Room Case'			
ORACLE (From CNTLS)	**ChatGLM-6B-2K**	**ChatGLM2-6B-8K**	**ChatGLM2-6B-32K**
2020.03.23 执政党共同民主党23日表示将争取在本届国会任期内通过防止再次发生"N号房"事件的三部法律，依据刑法严惩使用性影像威胁他人的行为，并对下载、复制、拍摄、散布等行为加强处罚。据韩联社报道，韩国总统文在寅23日就"N号房"事件表示，此案是践踏受害者的犯罪行为，指示警方彻查案件，让所有犯罪分子得到应有的惩罚。警方共调查发现74名女性受害者，包括16名未成年人。	2020.03.23 执政党共同民主党表示，将寻求通过立法来惩罚犯罪者，并希望调查所有参与者。总统文在寅也加入了这一行动，表示对事件感到"愤怒"和"震惊"，并指示调查所有参与者，并对加害者给予严厉的处罚。	2020.03.23 执政党共同民主党表示，将争取在本届国会任期内通过防止再次发生"N号房"事件的三部法律，依据刑法严惩使用性影像威胁他人的行为，并对下载、复制、拍摄、散布等行为加强处罚。韩国总统文在寅就"N号房"事件表示，此案是践踏受害者的犯罪行为，指示警方彻查案件，让所有加害者得到应有的惩罚。警方共调查发现74名女性受害者，包括16名未成年人。	2020.03.23 执政党共同民主党23日表示，将争取在本届国会任期内通过防止再次发生"N号房"事件的三部法律。韩国总统文在寅23日就"N号房"事件表示，此案是践踏受害者的犯罪行为，指示警方彻查案件，让所有加害者得到应有的惩罚。警方共调查发现74名女性受害者，包括16名未成年人。此外，韩国总统文在寅还表示，政府将删除所有涉案视频，并为受害者提供法律、医疗等所需支援。
Key Facets Analysis of Daily Summary in 2020.03.23			
通过三部法律 Enacting three laws 得到应有的惩罚 Receiving due punishment 发现74名女性受害者 Discovering 74 female victims	通过三部法律 [Not.m] Enacting three laws 得到应有的惩罚 ☑ Receiving due punishment 发现74名女性受害者 [Not.m] Discovering 74 female victims	通过三部法律 ☑ Enacting three laws 得到应有的惩罚 ☑ Receiving due punishment 发现74名女性受害者 ☑ Discovering 74 female victims	通过三部法律 ☑ Enacting three laws 得到应有的惩罚 ☑ Receiving due punishment 发现74名女性受害者 ☑ Discovering 74 female victims
...	...	...	...
2022.04.13 4月13日转引韩媒KBS新闻报道，韩国警察调查发现，适用于赵主彬的罪名高达12项，其中包括强制猥亵、强迫、欺诈、违反个人信息保护法等.....韩国检方4月13日以涉嫌14项罪名为由，对赵主彬进行拘留起诉。	2022.04.13 赵主彬被指控制作并传播非法视频，涉嫌14项罪名，包括强制猥亵、强迫、欺诈、违反个人信息保护法等。韩国首尔中央地方检察厅网络性犯罪特别调查组将对赵主彬进行起诉，并将公布中期调查报告。警方最早将于13日举行个人资料公开委员会，以决定是否公开文某的个人资料。赵主彬的拘留时间将被延长至13日。	2022.04.13 警方发现赵主彬涉嫌14项罪名，包括强制猥亵、强迫、欺诈、违反个人信息保护法等。此外，韩国检方以涉嫌14项罪名为由对赵主彬进行拘留起诉。截至4月8日，韩国警方共抓获涉嫌网络性犯罪的犯罪嫌疑人221人，其中32人被刑事拘留。经核实，受害女性中儿童和青少年为8人，还有17名成年人。	2022.04.13 根据韩媒KBS的报道，赵主彬被指控的罪名高达12项，包括强制猥亵、强迫、欺诈、违反个人信息保护法等。警方在调查过程中抓获了赵主彬方面主张的3名共犯中的2人，并对其进行调查。截至4月8日，韩国警方共抓获涉嫌网络性犯罪的犯罪嫌疑人221人，其中32人被刑事拘留。最高人民检察院、公安部派出联合督导组赶赴山东，对该案办理工作进行督导。
Key Facets Analysis of Daily Summary in 2022.04.13			
韩国警察调查发现，主犯罪名高达12项 The South Korean police found the main suspect facing charges in up to 12 cases. 韩国检查方以涉嫌14项罪名为理由，对主犯进行拘留起诉 The South Korean prosecutors detained and charged the main suspect with 14 offences.	韩国警察调查发现，主犯罪名高达12项 ☒ The South Korean police found the main suspect facing charges in up to 12 cases. 韩国检查方以涉嫌14项罪名为理由，对主犯进行拘留起诉 ☒ The South Korean prosecutors detained and charged the main suspect with 14 offences.	韩国警察调查发现，主犯罪名高达12项 ☒ The South Korean police found the main suspect facing charges in up to 12 cases. 韩国检查方以涉嫌14项罪名为理由，对主犯进行拘留起诉 ☑ The South Korean prosecutors detained and charged the main suspect with 14 offences.	韩国警察调查发现，主犯罪名高达12项 ☑ The South Korean police found the main suspect facing charges in up to 12 cases. 韩国检查方以涉嫌14项罪名为理由，对主犯进行拘留起诉 [Not.m] The South Korean prosecutors detained and charged the main suspect with 14 offences.
...	...	...	...