OpenForecast: A Large-Scale Open-Ended Event Forecasting Dataset

Zhen Wang[1,2,3], Xi Zhou[1,2,3,*], Yating Yang[1,2,3,*], Bo Ma[1,2,3,*],
Lei Wang[1,2,3], Rui Dong[1,2,3], Azmat Anwar[1,2,3],
[1]Xinjiang Technical Institute of Physics & Chemistry,
Chinese Academy of Sciences, Urumqi, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi, China
{wang_zhen, zhouxi, yangyt, mabo, wanglei, dongrui, azmat}@ms.xjb.ac.cn

Abstract

Complex events generally exhibit unforeseen, multifaceted, and multi-step developments, and cannot be well handled by existing closed-ended event forecasting methods, which are constrained by a limited answer space. In order to accelerate the research on complex event forecasting, we introduce OpenForecast, a large-scale open-ended dataset with two features: (1) OpenForecast defines three open-ended event forecasting tasks, enabling unforeseen, multifaceted, and multi-step forecasting. (2) OpenForecast collects and annotates a large-scale dataset from Wikipedia and news, including 43,419 complex events spanning from 1950 to 2024. Particularly, this annotation can be completed automatically without any manual annotation cost. Meanwhile, we introduce an automatic LLM-based Retrieval-Augmented Evaluation method (LRAE) for complex events, enabling OpenForecast to evaluate the ability of complex event forecasting of large language models. Finally, we conduct comprehensive human evaluations to verify the quality and challenges of OpenForecast, and the consistency between LEAE metric and human evaluation. OpenForecast and related codes will be publicly released[1].

1 Introduction

Event forecasting (Granroth-Wilding and Clark, 2016; Zhou et al., 2022; Du et al., 2022; Zhang et al., 2023), a challenging and attractive task, aims to forecast future events based on the analysis of background and can be applied in various domains such as political event forecasting (Ma et al., 2023), disaster warning (Zhao, 2022), and financial market analysis (Ashtiani and Raahemi, 2023).

Existing event forecasting tasks can be categorized into script event prediction (Li et al., 2018; Wang et al., 2021; Zhu et al., 2023) and temporal knowledge graph completion (TKGC, Granroth-Wilding and Clark, 2016; Ma et al., 2023; Shi et al., 2023), which aim to select a subsequent event from a few options and to predict missing links for a temporal graph, respectively. These tasks and studies (Li et al., 2021b; Yuan et al., 2024) contribute significantly to the progression of event forecasting but **are constrained to a limited answer space**, thereby belonging to closed-ended event forecasting. However, as illustrated in Figure 1, complex events typically exhibit unforeseen developments such as the Ukraine's fierce resistance; multifaceted developments such as the military progress, political condemnation, economic sanctions; and multi-step developments such as the Russian attack, Russian retreat, and Ukraine's counterattack. These **unforeseen, multifaceted, and multi-step developments** cannot be well handled by existing closed-ended event forecasting due to its limited answer space, underscoring the necessity and urgency of open-ended event forecasting.

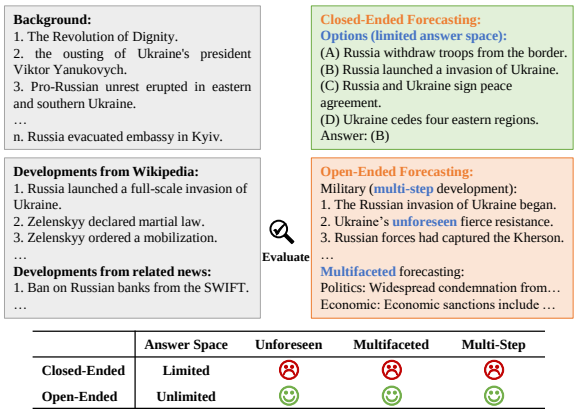

Figure 1: Comparison between open-ended and closed-ended event forecasting for complex events. Closed-ended forecasting is constrained to a limited answer space, while open-ended forecasting facilitates unforeseen, multifaceted, and multi-step predictions.

To advance the research on complex event fore-

* Corresponding author
[1]https://github.com/miaomiao1215/Openforecast

Proceedings of the 31st International Conference on Computational Linguistics, pages 5273–5294
January 19–24, 2025.

casting, we introduce OpenForecast, a large-scale open-ended dataset with two features. (1) **OpenForecast defines three open-ended event forecasting tasks**, including argument-level, short-term, and long-term forecasting, which predict fine-grained arguments, events on a specified date, and long-term event evolution, respectively. (2) **OpenForecast collects and annotates a large-scale dataset from Wikipedia and news**, including 43,417 complex events spanning from 1950 to 2024. Each complex event is annotated with multi-step event evolution, including the background, multifaceted development, and aftermath. Particularly, this annotation can be completed automatically without any manual annotation cost. To prevent knowledge leakage during the evaluation, the dataset is partitioned according to occurrence time. Additionally, **we introduce an automatic LLM-based Retrieval-Augmented Evaluation method (LRAE) for complex events.** As illustrated in Figure 1, in open-ended tasks, true developments and predictions contain multiple fine-grained (atomic) events, and some true predictions are only recorded in related news. Inspired by the fact-confirmation pipeline of human and Retrieval-Augmented Generation (RAG, Lewis et al., 2020), LRAE segments original prediction into atomic predictions, retrieves relevant contents from the web, and performs many-to-many semantic matching.

In experiments, we conduct comprehensive human evaluations and demonstrate the high quality of OpenForecast, achieving average scores of 98.0%, 94.2%, and 95.7% on dataset collection, timeline annotation, and question annotation, respectively. Additionally, evaluations across eight LLMs highlight the challenges of OpenForecast, revealing that LLMs exhibit strong potential in open-ended forecasting but show a pessimistic tendency. Using human evaluations as the gold standard, our LRAE achieves the highest consistency across the three open-ended tasks, significantly outperforming other automatic evaluation methods. We summarize our contributions as follows:

- We define three open-ended event forecasting tasks, including argument-level, short-term, and long-term forecasting.
- Using automatic methods, we propose a large-scale dataset, including 43,417 high-quality complex events spanning from 1950 to 2024.
- We introduce an open-ended evaluation method, LRAE, demonstrating the highest consistency.

2 Related Works

Benchmarks There are mainly two kinds of benchmarks corresponding to script event prediction and TKGC. For script event prediction, Li et al. (2018) employed an extraction pipeline (Granroth-Wilding and Clark, 2016) to extract structured event chains and released the multi-choice narrative cloze (MCNC) dataset, which requires models to select the answer from candidates. Additionally, Jin et al., 2021 proposed an unstructured QA dataset ForecastQA, Autocast(Zou et al., 2022) and Halawi et al. (2024) proposed binary event prediction (True/False) and numerical event prediction. For TKGC, ICEWS (García-Durán et al., 2018) and GDELT (Qiao et al., 2015) are two open-source projects to monitor global events and are widely used. These datasets include numerous atomic events but lack event relation linking, with each event annotated with predefined entities and types according to the CAMEO taxonomy. To capture the complex relations among atomic events, IED (Li et al., 2021a) and SCTc-TE (Ma et al., 2023) employed automatic approaches to construct complex events. However, these datasets are constrained to a limited answer space, hindering the forecasting of unforeseen, multifaceted, and multi-step events.

Open-Ended Evaluation Different from closed-ended tasks, open-ended tasks such as open-ended QA lack absolute labels and thus cannot be evaluated using exact matching. There are mainly two kinds of evaluation methods: human evaluation and automatic evaluation. Human evaluations show better alignment with human preferences in interactive dialogue (Liu et al., 2023a; Ruan et al., 2024) and summarization (Pu et al., 2024; Liu et al., 2023c). However, they suffer from inconsistent quality (Chiang and Lee, 2023), reproducibility crisis (Belz et al. (2023)), and nonnegligible annotation costs. In contrast, automatic evaluations benefit from standardized, objective, and human-free property. These methods can be categorized into three groups: (1) lexical matching methods such as ROUGE and BLEU; (2) semantic matching methods such as BertScore (Zhang et al., 2020) and BEM (Bulian et al., 2022); (3) LLM-based evaluations such as PandaLM (Wang et al., 2023), GPTScore (Fu et al., 2023), GEMBA (Kocmi and Federmann, 2023), and G-EVAL (Liu et al., 2023b). Kamalloo et al. (2023) and Min et al. (2023) con-

duct comprehensive experiments and demonstrate the superior performance of LLM-based evaluation in open-ended evaluation. Interestingly, the evolution of automatic evaluation methods mirrors the advancements in NLP, characterized by increasingly enhanced language processing capabilities.

3 Task Definition

Given a complex event CE, we define the input as the background X before a specified time T and question Q, with the subsequent multifaceted developments as the gold answer Y. Depending on the question, the gold answer Y may be a single response or a list-style response.

Based on previous studies (Ma et al., 2023), we define a complex event CE as a chronologically ordered event chain $CE = \{e_1, e_2...e_n\}$ on the same topic, where e_i is i-th atomic event in CE. Each atomic event (Li et al., 2021a) is annotated with a standardized timestamp if explicitly mentioned in the original articles. A timestamp T_k then divides CE into background events $X = \{e_1, e_2...e_{k-1}\}$ and target events $Y = \{e_k, e_{k+1}...e_n\}$.

To facilitate the unforeseen, multifaceted, and multi-step event forecasting, we design a short-term and a long-term forecasting tasks, which predict events on a specific date and long-term event evolution, respectively, with unforeseen events inside. Note that multiple atomic events could happen in one day, resulting in a list-style Y for the short-term forecasting task. While atomic events in these tasks contain multiple arguments, such as *Subject: Russia, Event type: launched a full-scale invasion, Object: Ukraine*, we design an argument-level open-ended forecasting task to further examine the forecasting ability on specific fine-grained event arguments. The detailed descriptions to three tasks are listed below.

Short-Term Forecasting (STF). This task examines the short-term event forecasting ability on a given timestamp T_k. With event background X as input, all multifaceted atomic events occurring at T_k from Y form the gold answer.

Long-Term Forecasting (LTF). This task examines the ability to forecast long-term event evolution after a given timestamp T_k. With event background X as input, models are required to forecast the multi-step event chain Y.

Argument-level QA (AQA). This task examines the forecasting ability on fine-grained event arguments, including event type, subject, object, time, and location. In this task, models are provided with event background X and a question Q such as *"Who will", "What will", "When will"*, with corresponding argument such as *Ukraine* as the answer.

Based on existing closed-ended tasks, we also propose three closed-ended tasks:

Multi-Choice Narrative Cloze (MCNC). Similar to script event prediction, models are provided with event background X and four subsequent candidate events, with one gold answer inside.

Multi-Choice Argument-level Cloze (MCAC). Similar to AQA, four candidate answers are additionally provided, with one gold answer inside.

Verify QA (VQA). In this task, given the event background and one candidate event, models need to predict whether the candidate event will occur.

4 Dataset Construction Pipeline

We review current event forecasting datasets and identify a lack of datasets for the open-ended tasks. To support these tasks above, we present OpenForecast, a large-scale dataset. As illustrated in Figure 2, the dataset construction pipeline includes three steps: (1) dataset collection for complex events; (2) event timeline annotation; (3) question generation.

4.1 Dataset Collection

To enable forecasting on unforeseen, multifaceted, and multi-step events, it is essential to collect complex events with unforeseen changes and complete multifaceted evolutions, including backgrounds, developments, and aftermaths.

In this paper, we collect data from two projects: **Wikipedia** and Wikipedia Current Events Portal (**WCEP**). Wikipedia offers numerous articles on historical events, providing detailed backgrounds, developments, and aftermaths. The WCEP continuously documents current events and organizes events on the same topic with the same subheaders, each with an event summary and at least one external link. These projects encompass extensive influencing and dramatic complex events across various domains, satisfying our needs. After data scraping, we propose a multi-step filtration to remove duplicate, non-events, and non-contemporary data. Subsequently, we group articles on the same topic together, resulting in a large-scale high-quality collection of complex events. The detailed procedures are illustrated in appendix A.1.

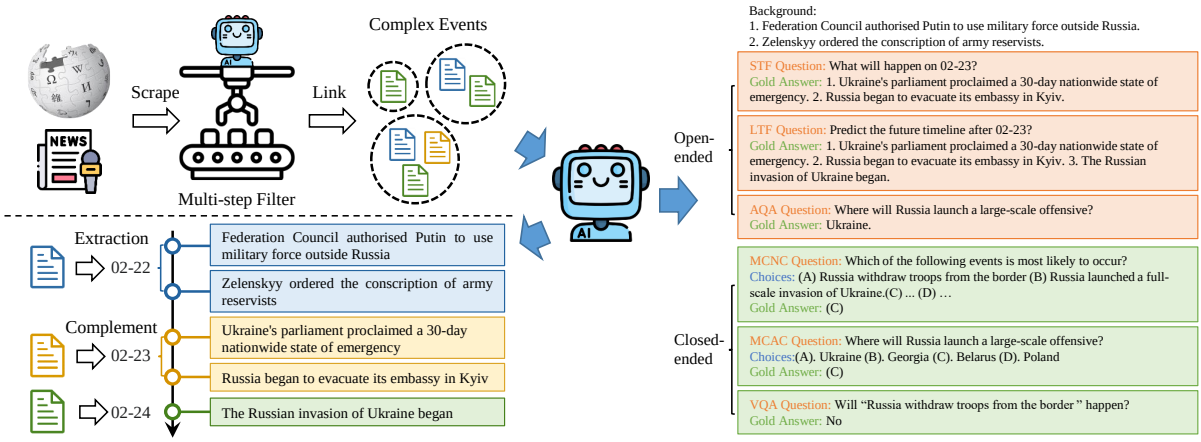

Figure 2: Illustration of the construction pipeline for OpenForecast, including three steps: (1) dataset collection for complex events; (2) event timeline annotation using the extraction-then-complement approach; and (3) question generation for six tasks.

4.2 Event Timeline Annotation

Different from GDELT and ICEWS, which extract atomic events but overlook their complex relations, our event timeline annotation aims to extract chronologically ordered event chains CEs from multiple articles. Each complex event from Wikipedia contains multiple sections in one article, whereas those from WCEP contain at least one article. Leveraging LLMs, we propose a two-stage pipeline named **extraction-then-complement**. Specifically, for each complex event, we sort articles by time and perform event timeline extraction on the first article, requiring that the atomic events objectively occur. For Wikipedia articles, an additional preprocessing step is applied to retain only sections related to event evolution, such as introduction, background, development, and aftermath, thereby reducing the input length. Then, for the remaining articles, we sequentially conduct the event timeline complement, requiring the LLMs to perform event extraction, coreference resolution, and event filling simultaneously. To ensure chronological coherence, we perform an additional reranking on the event timeline using LLMs. In our experiments, we observe that due to the high complexity of event timeline completion, the performance heavily depends on the capabilities of LLMs. When employing the stronger Llama3-70b, the performance of extraction is significantly improved compared to Llama2-70b.

4.3 Question Annotation

For the tasks **STF** and **LTF**, we randomly choose one timestamp T_k specified in a day, partition CE into background and target events, and form the sample as introduced in the task definition above.

For the **AQA** task, we randomly select one event from Y as the target event and an argument type from event category, time, location, subject, and object as the target argument. Then, we leverage LLMs to design an argument-level question, such as "When will", "What will", "Who will", and "Where will", and extract its gold answer.

For the **MCNC** task, we randomly select one event from Y as the gold answer. Then, we prompt LLMs to generate three challenging negative candidates by replacing event arguments (Jin et al., 2021) or generating opposite events. Additionally, rules such as ensuring negative candidates explicitly not occur according to the given article are added. To eliminate negative candidates that actually occurred, we filter out those with the same arguments as true events in the timeline (approximately 7.3% are discarded). For the **MCAC** task, similar to the AQA task, three negative candidates are generated by LLMs. For the **VQA** task, we randomly select the gold answer or one negative candidate from samples in the MCNC task, assigning labels with "Yes" and "No" respectively.

4.4 Dataset Statistics

Table 2 presents the comparison of OpenForecast with other event forecasting datasets. Notably, OpenForecast exhibits the **largest scale** in complex events, surpassing SCTc-TE by approximately tenfold. The dataset includes 43,417 complex events (25,975 from Wikipedia, 17,442 from WCEP) and 473,155 atomic events, average 10.6 atomic events per complex event. Following the WCEP, we cat-

Datasets	Time	CEs	AEs
ICEWS18	2018	0	468,558
ESC	~2017	258	7,275
General	2020	617	8,295
IED	~2021	430	51,422
SCTc-TE	2015-2022	4,397	45,587
Ours	1950-2024	43,417	473,155

Table 1: Statistic of OpenForecast in comparison to existing event forecasting datasets. CEs denotes the number of complex events and AEs denotes the number of atomic events. The General dataset is curated by LDC (LDC2020E25) and the IED, ESC, and are constructed by Caselli and Vossen (2017), Li et al. (2021a), and Ma et al. (2023) respectively.

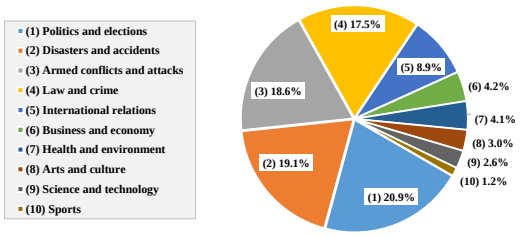

Figure 3: The category distributions of OpenForecast.

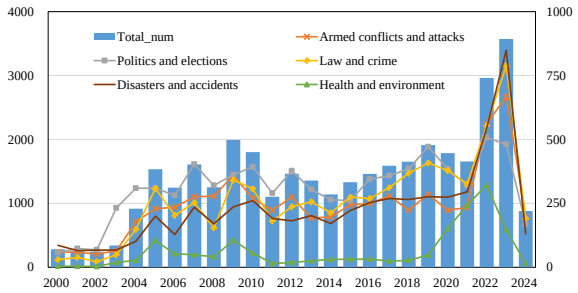

Figure 4: The temporal evolution of total event counts and five distinct event categories from 2000 to 2024.

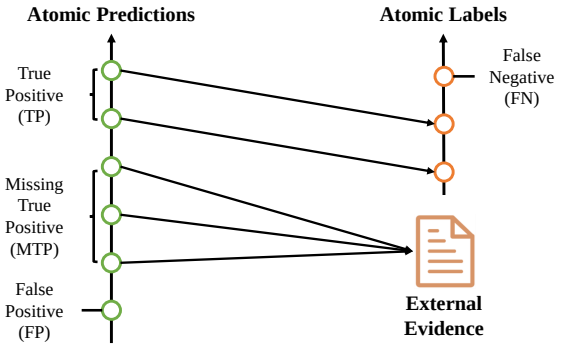

Figure 5: Illustration for the LLM-based retrieval-augmented evaluation.

egorize the complex events into ten types. The type distribution is shown in Figure 3, ensuring the training and evaluation for cross-domain forecasting models.

Moreover, OpenForecast encompasses major events from 1950 to 2024, allowing in-depth research into the long-term event evolution. We further analyze the changes in the total number of events and five specific categories from 2000 to 2024. There is a notable rise in *armed conflicts and attacks* from 2022 to 2023 and in *health and environment* from 2020 to 2022, corresponding to the Russia-Ukraine war, the Israel–Hamas conflict, and the COVID-19 pandemic.

Through human evaluation on event timeline construction and question annotation from seven dimensions, OpenForecast demonstrates superior quality, with average scores in 98.0%, 94.2%, and 95.7% on dataset collection, timeline annotation, and question annotation, respectively. To prevent knowledge leakage in the evaluation, we take the data before 2023/06/30 as the trainset, data between 2023/07/01 and 2023/08/31 as the validation set, and data between 2023/09/01 and 2024/03/31 as the testset. For the evaluation of new LLMs, we can use the subset of the testset or update OpenForecast with novel corpora, thus ensuring its alignment with the continual advancements in LLMs. Details of the dataset collection, statistics, and evaluation can be found in Appendices A.1, A.2, A.3, and A.4, respectively.

5 LLM-based Retrieval-Augmented Evaluation

Open-ended evaluations involve open-ended responses and thus necessitate semantic matching, unlike closed-ended event forecasting featuring a limited answer space. Additionally, as depicted in Figure 1 and 5, the gold answers and predictions of open-ended tasks STF and LTF contain multiple atomic events, resulting in many-to-many matching. In the human evaluation, we observe that some true atomic predictions are not included in gold answers, resulting in an underestimated evaluation. Inspired by the fact confirmation pipeline of human and RAG, we propose the LLM-based Retrieval-Augmented Evaluation (LRAE) to address these issues. Specifically, for the many-to-many matching issue, we partition prediction into atomic predictions and iteratively verify them. For the underestimated evaluation issue, we retrieve relevant contents from the web as supplementary evidences. The detailed process is shown as follows:

1) The original prediction is partitioned into atomic predictions using the regular expression.

2) LRAE uses a search API (Serper, which takes the Wikipedia title or news title as input) to retrieve relevant websites, crawls their contents using Newspaper3k and WebBaseLoader, and segments them into paragraph-level chunks as supplementary evidences.

3) Using LLMs (Llama3-8b), LRAE iteratively verifies each atomic prediction against labels to determine whether the prediction corresponds to a specific label. Additionally, LLMs should identify the specific label to which the prediction corresponds. The number of correct predictions corresponds to the *TP*. The number of unrecalled labels corresponds to the *FN*. The false predictions in this step will be double-verified using Steps 4 and 5.

4) For the remaining atomic predictions from Step 2, using text embedding model (bge-large-en-1.5), LRAE retrieves Top-K (5 in this work) similar chunks as external evidences.

5) LRAE iteratively verifies each atomic prediction against retrieved chunks to determine whether the prediction is supported by these chunks (corresponding to the *MTP*). The atomic predictions that are verified as false correspond to the *FP*.

Finally, we compute precision, recall, and F1-score using the formula (2) in Appendix B.3, which alleviates the underestimated evaluation issue.

6 Experiments

6.1 Experimental Setup

Models To comprehensively evaluate the performances of LLMs, we select a variety of open-source LLMs with different scales, including Llama2 series (7b, 13b, 70b), Vicuna-13b, Wizardlm-13b, Falcon-40b, Mixtral-8x7b, and Llama3-8b. All the LLMs have been instruction-tuned and exhibit no knowledge leakage[2]. For comparison, we finetune BertMultipleChoice models as baselines for closed-ended tasks. Furthermore, using multi-task training, we conduct fine-tuning on Llama3-8b-instruct (Llama3-8b-SFT).

[2]The pretraining data for these LLMs has a cutoff before 2023/09/01. Details can be found in Appendix B.1

Settings For the evaluation of LLMs, the decoding temperature is set to 0.0 and the same prompt templates in Table 10 are adopted for all LLMs. For the training of BertMultipleChoice, we use the Adam optimizer with a learning rate of 5e-5 over 3 epochs. Similarly, for the training of Llama3-8b-SFT, we used the Adam optimizer with a learning rate of 2e-5 over 3 epochs. Experiments are conducted on four NVIDIA Tesla A100 GPUs with 80GB of RAM each.

Evaluation Metrics For tasks including AQA, MCNC, MCAC, and VQA, we evaluate them with accuracy. For many-to-many matching tasks including STF and LTF, we report them using modified precision, recall, and F1-score in Appendix B.3. For **closed-ended tasks**, we adopt automatic answer matching. For **open-ended tasks**, we conduct human evaluation (using original articles, event timelines, and web searching, see Appendix B.2) to reflect the true performance by randomly sampling 100 questions for each task and their corresponding responses from nine models, resulting in a total of 900 unique answers. After atomic event partitioning, the number of atomic predictions reaches approximately 9,000 for STF and LTF, rendering the human evaluation labor-intensive.

6.2 Main Results

Table 2 shows the experimental results on the open-ended and close-ended event forecasting tasks, from which we have the following observations: (1) **Open-ended tasks is much more challenging than closed-ended tasks.** On closed-ended tasks, despite the moderate performances of LLMs, the baseline BertMultipleChoice achieved 91.5%, 80.3%, and 83.7% on MCNC, MCAC, and VQA, respectively. Additionally, Llama3-8b-SFT yields significant improvement over it. However, the performances of LLMs on open-ended AQA, STF, and LTF are lower than 65%. (2) **Long-term forecasting is more challenging**. The best performances of LLMs on AQA, STF, and LTF are 63.5%, 57.2%, and 50.7%, respectively, with LTF achieving the lowest accuracy. For long-term complex events, their developments might be changed by various unforeseen events, making LTF the most challenging. (3) **LLMs exhibit strong potential in open-ended forecasting.** Despite suboptimal performances on open-ended tasks, human evaluations indicate that LLMs excel in making comprehensive predictions, including multifaceted impacts (politics, economy, and diplomacy, etc), long-term impacts, interna-

Model	AQA	STF			LTF			MCNC	MCAC	VQA
		P	R	F1	P	R	F1			
BertMultipleChoice				/				91.5	80.3	83.7
Llama2-7b	62.0	35.1	58.8	41.6	47.8	53.8	48.2	30.5	52.5	57.5
Llama2-13b	57.5	41.1	58.5	45.9	46.2	53.9	46.5	42.8	51.6	55.3
Llama2-70b	**63.5**	42.4	66.2	48.8	**48.5**	60.8	**50.7**	47.0	67.8	57.9
Vicuna-13b	44.5	25.0	45.3	29.0	45.4	48.4	44.7	47.3	65.3	51.5
Wizardlm-13b	50.5	33.3	43.5	35.8	37.4	49.5	38.5	45.3	61.7	54.8
Falcon-40b	56.5	29.5	30.3	28.1	21.6	31.0	23.2	41.0	56.5	59.0
Mixtral-8x7b	61.0	52.3	**70.6**	**57.2**	45.2	**61.1**	47.7	55.2	67.7	64.6
Llama3-8b	63.0	**52.4**	70.1	56.8	47.2	59.5	49.5	55.7	67.3	57.0
Llama3-8b-SFT	61.5	27.8	21.2	22.9	27.8	21.2	22.9	**96.1**	**89.2**	**89.3**

Table 2: Experimental results (%) of diverse models on open-ended (AQA, STF, and LTF) and closed-ended (MCNC, MCAC, and VQA) tasks.

Model	Positive			Negative			T/F
	P	R	F1	P	R	F1	
Llama2-7b	54.5	83.2	65.8	66.8	32.8	43.9	3.01
Llama2-13b	52.7	89.9	66.4	69.1	21.9	33.2	5.21
Llama2-70b	54.2	92.2	68.4	76.9	25.0	37.7	5.06
Vicuna-13b	65.5	3.0	5.8	51.2	**98.5**	67.4	0.02
Wizardlm-13b	52.4	89.8	66.2	68.0	21.0	32.1	5.39
Falcon-40b	56.6	71.1	63.0	62.8	47.2	53.9	1.62
Mixtral-8x7b	60.2	82.3	69.6	73.5	47.4	57.7	2.05
Llama3-8b	53.5	**96.5**	68.8	**84.7**	18.8	30.8	7.84
SFT	**78.9**	71.4	**75.0**	74.3	81.3	**77.7**	0.96

Table 3: The significant disparity between positive and negative samples in the VQA task (%). The *T/F* denotes the ratio of the number of samples predicted as positive to those predicted as negative. The ratio of positive and negative samples (label) is 0.97.

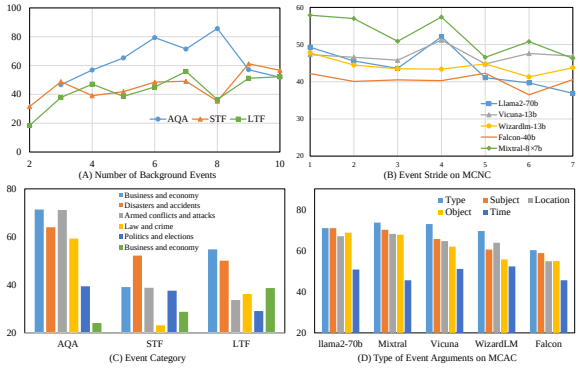

Figure 6: Performances (%) analysis of influencing factors.

tional reactions, etc. (4) **LLMs tend to make pessimistic forecasting.** For instance, when presented with an event background involving protest, impeachment, or border clash, LLMs often make pessimistic predictions such as government collapse, long-term instability, or large-scale wars (as shown in Table 7). (5) **LLMs tend to predict the occurrence of candidate events.** As demonstrated in Table 3, most LLMs achieve much higher recall but lower precision for positive candidates than negative ones and show a notable tendency to predict true rather than false.

Influencing Factors for Event Forecasting. Compared to other LLMs, Llama2-70b, Mixtral-8x7b, and Llama3-8b perform better, which aligns with their ranking on the leaderboard, thus confirming the importance of general capabilities for forecasting. We further investigate four influencing factors, as depicted in Figure 6. Due to the limited number of human evaluations on open-ended tasks, the average scores of six LLMs, including Mixtral-8x7b, Llama2-70b, Vicuna-13b, Wizardlm-13b, Falcon-40b, and Llama3-8b, are adopted in Figure 6 (A) and (C). The results in Figure 6 (A) indicate that more event backgrounds might improve prediction accuracy. Figure 6 (B) indicates that larger event strides (the distance between background and target event) are much harder due to increased uncertainty. In Figure 6 (C), the performances of armed conflicts and attacks, international relations, law and crime, and politics and elections are inferior to those of disasters and accidents, because the disaster responses (such as post-disaster relief and material assistance) are quite similar. Figure 6 (D) indicates that the performance distribution of LLMs across argument types remains consistent, with the best results in *Type* while the worst results in *Time*. Given the inherent challenges in time prediction (Zhao, 2022), it may be beneficial

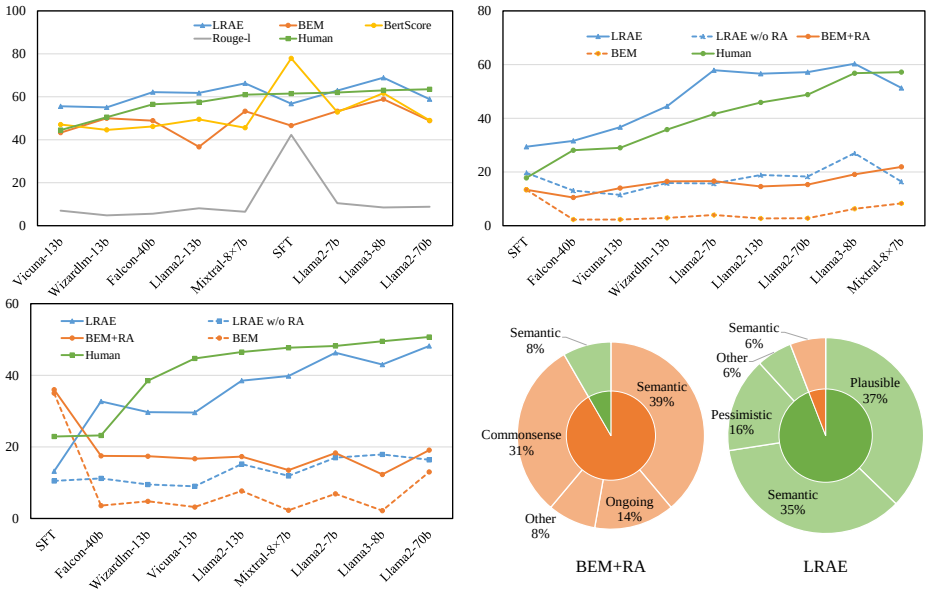

Figure 7: Analysis of open-ended evaluation methods. The upper left, upper right, and lower left figures depict the comparative performance (%) of human and automatic evaluation methods across AQA, STF, and LTF respectively. The lower right graph illustrates the statistic of error types for the BEM+RA and LRAE, with orange segments representing missing errors (positive predictions are misclassified as negative) and green segments representing overestimation errors (negative predictions are misclassified as positive). Detailed definitions for the error types are shown in Table 6.

to exclude time during fine-tuning.

The analysis on SFT and event representation format can be found in Appendix B.4 and B.5.

6.3 Experiments on LRAE

In this section, we evaluate four automatic evaluation methods on open-ended tasks: (1) **Rouge-L** (F1), a lexical matching method; (2) **BertScore** (F1), a token-level semantic matching method; (3) **BEM**[3] (accuracy for AQA and F1 for STF and LTF), a sentence-level semantic matching method; (4) **LRAE**[4] (accuracy for AQA and F1 for STF and LTF), a LLM-based semantic matching method.

As depicted in Figure 7, we sort the LLMs (on the horizontal axis) according to their performance using human evaluations, yielding a progressively ascending line. In the AQA task (upper left), a notable discrepancy is observed between Rouge-L and human evaluations. This is because the answers of LLMs are lengthy while the labels are short, resulting in a low overlap. Additionally, other evaluation methods show good consistency with human evaluation, with our LRAE performing the best, indicating that language models, especially LLMs, can match semantically equivalent answers despite significant textual differences.

For the many-to-many matching tasks (STF: upper right, LTF: lower left), LRAE and BEM are selected[5]. However, the gaps between them and human evaluations become larger. Unlike AQA (argument-level one-to-one matching), STF and LTF are required to match multiple atomic events, each with multiple event arguments (event-level many-to-many matching), thereby significantly increasing the difficulties. Notably, LRAE exhibits the best consistency with human evaluation, particularly on STF, due to the strong language understanding ability of LLMs and the retrieval augmentation module (RA). After removing RA, there is a significant performance decline for LRAE w/o RA, demonstrating the effectiveness of RA. The detailed results of LRAE and BEM are presented in Table 8, and 9. However, LRAE exhibits minor ranking discrepancies compared to human evaluation, indicating the need for further refinement.

Furthermore, on STF and LTF, we collect samples with the absolute F1-score differences between

[3]For AQA, BEM conducts once matching between prediction and label. For STF and LTF, BEM conducts multiple one-to-one matching for each atomic prediction-label pair.

[4]For AQA, we remove the retrieval augmentation from LRAE and employ once matching using LLM.

[5]For an atomic prediction-label pair, BertScore and Rouge-L provide continuous outputs, rather than binary 'Yes' or 'No' classifications, thus are not applicable to many-to-many matching.

human and automatic evaluation exceeding 0.3. We then investigate the reasons for these failures of BEM+RA and LRAE, as depicted in Figure 7 (lower right). For BEM+RA, 92% of errors are matching missing (positive predictions are judged as negative). This issue stems primarily from suboptimal semantic modeling ability. Additionally, BEM lacks commonsense reasoning ability and fails in handling ongoing events (as shown in Table 6). In contrast, LRAE alleviates the matching missing issues but introduces minor overestimation errors (negative predictions are judged as positive). Among these errors, LRAE often regards plausible and pessimistic predictions as correct, even in the absence of relevant content supporting the prediction. These failures indicate that LLM-based evaluation may suffer from hallucinations and thus needs further optimization of LLMs.

7 Conclusion

To promote event forecasting from the closed-ended paradigm to the open-ended paradigm, we introduce OpenForecast, an open-ended event forecasting dataset characterized by defining three open-ended tasks and automatically annotating a large-scale dataset from Wikipedia and news. Additionally, we introduce LRAE for the automatic evaluation of open-ended tasks. Using human evaluations and experiments, we demonstrate the quality and challenges of OpenForecast, as well as LRAE's superior consistency with human evaluations. Future work will focus on exploring advanced fine-tuning methods for open-ended tasks and increasing consistency with human evaluation.

Acknowledgement

This research is sponsored by the Xinjiang Uygur Autonomous Region "Tianshan Talents" Scientific and Technological Innovation Leading Talent Project (2022TSYCLJ0035), the Youth Talents Support Project of Xinjiang Uyghur Autonomous Region (2023TSYCQNTJ0037), the "Tianshan Elite" Science and Technology Topnotch Youth Talents Program (2022TSYCCX0059), Tianshan Talent Training Program (2023TSYCCX0041), the Outstanding Member Program of the Youth Innovation Promotion Association of Chinese Academy of Sciences (Y2021112, Y2023118), and the Natural Science Foundation of Xinjiang Uyghur Autonomous Region (2022D01D04, 2022D01B207).

Limitations

In this section, we discuss several limitations in our work. First, the construction of OpenForecast relies on the performance of LLMs and necessitates substantial computational resources. Second, despite demonstrating superior consistency, our LRAE exhibits minor ranking discrepancies when compared to human evaluation. Further research is needed to enhance its robustness.

Ethics Statement

In our study, OpenForecast was developed using open-source projects, including Wikipedia and WCEP. These resources have been widely employed in other studies, ensuring that no ethical standards were compromised.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes LLM better reasoner. *CoRR*, abs/2310.20689.

Matin N. Ashtiani and Bijan Raahemi. 2023. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Syst. Appl.*, 217:119509.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3676–3687. Association for Computational Linguistics.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 291–305. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 77–86. Association for Computational Linguistics.

David Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? 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 15607–15631. Association for Computational Linguistics.

Li Du, Xiao Ding, Yue Zhang, Ting Liu, and Bing Qin. 2022. A graph enhanced BERT model for event prediction. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2628–2638. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166.

Alberto García-Durán, Sebastijan Dumancic, and Mathias Niepert. 2018. Learning sequence encoders for temporal knowledge graph completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4816–4821. Association for Computational Linguistics.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2727–2733. AAAI Press.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. *CoRR*, abs/2402.18563.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. Forecastqa: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4636–4650. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of 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 5591–5606. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 193–203. European Association for Machine Translation.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare R. Voss. 2021a. The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5203–5215. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4201–4207. ijcai.org.

Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. 2021b. Temporal knowledge graph reasoning based on evolutional representation learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 408–417. ACM.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *CoRR*, abs/2305.20050.

Sijia Liu, Patrick Lange, Behnam Hedayatnia, Alexandros Papangelis, Di Jin, Andrew Wirth, Yang Liu, and Dilek Hakkani-Tur. 2023a. Towards credible human evaluation of open-domain dialog systems using interactive setup. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13264–13272. AAAI Press.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. 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 4140–4170. Association for Computational Linguistics.

Ruilin Luo, Tianle Gu, Haoling Li, Junzhe Li, Zicheng Lin, Jiayi Li, and Yujiu Yang. 2024. Chain of history: Learning and forecasting with llms for temporal knowledge graph completion. *CoRR*, abs/2401.06072.

Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, Liang Pang, and Tat-Seng Chua. 2023. Structured, complex and time-complete temporal event forecasting. *CoRR*, abs/2312.01052.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2024. Is summary useful or not? an extrinsic human evaluation of text summaries on downstream tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 9389–9404. ELRA and ICCL.

Fengcai Qiao, Pei Li, Jingsheng Deng, Zhaoyun Ding, and Hui Wang. 2015. Graph-based method for detecting occupy protest events using GDELT dataset. In *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2015, Xi'an, China, September 17-19, 2015*, pages 164–168. IEEE Computer Society.

Jie Ruan, Xiao Pu, Mingqi Gao, Xiaojun Wan, and Yuesheng Zhu. 2024. Better than random: Reliable NLG human evaluation with constrained active sampling. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18915–18923. AAAI Press.

Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Zhang, Jun Zhou, Chenhao Tan, and Hongyuan Mei. 2023. Language models can improve event prediction by few-shot abductive reasoning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lihong Wang, Juwei Yue, Shu Guo, Jiawei Sheng, Qianren Mao, Zhenyu Chen, Shenghai Zhong, and Chen Li. 2021. Multi-level connection enhanced representation learning for script event prediction. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3524–3533. ACM / IW3C2.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023. Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization. *CoRR*, abs/2306.05087.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1963–1974. ACM.

Mengqi Zhang, Yuwei Xia, Qiang Liu, Shu Wu, and Liang Wang. 2023. Learning long- and short-term representations for temporal knowledge graph reasoning. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 2412–2422. ACM.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Liang Zhao. 2022. Event prediction in the big data era: A systematic survey. *ACM Comput. Surv.*, 54(5):94:1–94:37.

Pengpeng Zhou, Bin Wu, Caiyong Wang, Hao Peng, Juwei Yue, and Song Xiao. 2022. What happens next? combining enhanced multilevel script learning and dual fusion strategies for script event prediction. *Int. J. Intell. Syst.*, 37(11):10001–10040.

Fangqi Zhu, Jun Gao, Changlong Yu, Wei Wang, Chen Xu, Xin Mu, Min Yang, and Ruifeng Xu. 2023. A generative approach for script event prediction via contrastive fine-tuning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14056–14064. AAAI Press.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. 2022. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

	Train	Dev	Test
CEs	41,424	474	1,519
AQA	82,846	948	3,038
STF	54,619	469	1,588
LTF	72,355	948	3,038
MCNC	86,326	886	3,697
MCAC	82,846	948	3,038
VQA	86,326	886	3,697

Table 4: Statistics of the data splitting of OpenForecast.

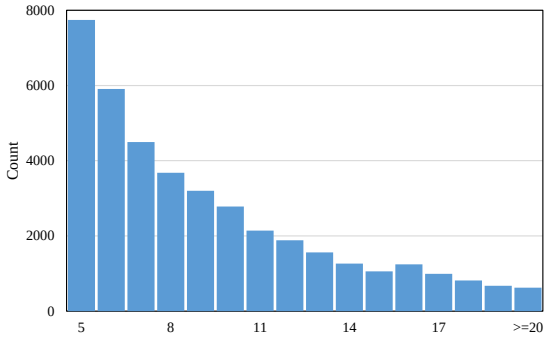

Figure 8: Number of complex events with varying atomic events number.

A Dataset

A.1 Details of Dataset Collection

In this work, the Wikipedia dump (20240320) and WCEP data before 2024 April are selected as the original sources. For the Wikipedia dump, we initially get 4,056,152 articles after excluding those with fewer than 200 words. We then implement a multi-step filtration method to efficiently filter event-related articles. The steps are as follows: (1) **Section Title Filtering.** A statistical analysis of section titles across Wikipedia articles reveals that event-related articles often contain sections such as "Background", "Development", "Aftermath", and "Reaction". Consequently, we collect all the section titles that may be relevant to the event and filter articles containing these section titles. (2) **Category Filtering.** While the section title filtering step excludes most non-event articles, a significant amount of noisy articles remained. To address this, we employ the Mixtral-8x7b to categorize the articles, discarding those that belong to individuals, locations, organizations, nations, etc. (3) **Time Filtering.** Events from different periods exhibit distinct evolution patterns. So we leverage Mixtral-8x7b to identify the occurrence dates of events and filter out those before 1950. For the WCEP dataset, which only documents current events after 2000, filtration is not employed. We first crawl and extract all metadata including event summaries, external links, time of occurrence, event categories, and subheaders from the WCEP website. Then we leverage the Newspaper3k project to scrape the news articles of external links. Finally, events sharing the same subheaders are aggregated into complex events.

By category filtering in data collection, we recognize the regular events (such as annual festivals, exhibitions, and conferences) by LLMs and the number is only 925/43419 $\approx$ 2.1%. Coupled with the diversity of problems, the impact of regular events is very small.

The prompt templates for the dataset construction are shown in Table 11.

A.2 Details of Dataset Splitting

To prevent knowledge leakage for the evaluation, we take the data before 2023/06/30 as the trainset, data between 2023/07/01 and 2023/08/31 as the validation set, and data between 2023/09/01 and 2024/03/31 as the testset. The detailed statistics of data splits of OpenForecast are presented in Table 4. Figure 8 illustrates the distribution of atomic event counts of complex events. Long-term complex events typically encompass more atomic events.

A.3 The Quality of the Event Timeline Construction

To evaluate the event timeline construction pipeline, which includes dataset collection and event timeline annotation, we randomly select 200 complex events (the selected samples encompass all event types with various article lengths) corresponding to 200 event timelines and ask two human annotators[6] to evaluate the extracted event timelines from multiple dimensions, as outlined below:

- *Event Relevance* examines whether non-event articles, including those related to individuals, locations, organizations, nations, festivals, entertainment, concepts, etc from Wikipedia, are successfully excluded.

- *Completeness* evaluates the completeness of the extracted event timelines. Annotators first review the articles to grasp the overall event timeline and then determine whether the extracted timelines

[6]Our annotation team consists of two graduate students engaged in information processing. Another annotator will review their annotation results and eliminate their discrepancies.

omit any atomic events. If two or more atomic events are omitted, the event timeline is annotated as incomplete.

- *Temporal Correctness* evaluates whether the extracted event timelines follow the correct chronological order. Annotators need to check the temporal relationships throughout the extracted event timelines according to the original articles. If any atomic events are in the wrong order, the event timeline in this dimension is annotated as incorrect.

- *Factual Consistency* examines whether the extracted atomic events contain factual errors. Annotators need to verify the atomic events using the original articles, ensuring factual consistency in event type, subject, object, time, location, etc. If two or more atomic events are incorrect, the entire event timeline is annotated as incorrect.

In these dimensions, the event relevance dimension corresponds to the dataset collection, while the other dimensions correspond to the event timeline annotation.

The annotation consistency ratios between the two annotators in *Event Relevance*, *Completeness*, *Temporal Correctness*, and *Factual Consistency* are 98.5%, 93.0%, 96.0%, and 98.5% respectively, indicating substantial agreement. OpenForecast achieves an *Event Relevance* of 98.0%, *Completeness* of 90.5%, *Temporal Correctness* of 94.5%, and *Factual Consistency* of 97.5%, demonstrating the effectiveness of our event timeline construction pipeline. In the *Completeness* dimension, our pipeline provides comprehensive event development, with minimal omissions of atomic events in the background and aftermath.

A.4 The Quality of the Question Annotation

In this work, we propose three open-ended tasks and three closed-ended tasks. The gold answers in open-ended tasks come from the extracted event timeline and are already proved in the event timeline construction evaluation (see Appendix A.3). For the closed-ended tasks, except the gold answers, noisy candidates and argument-level questions in MCAC are generated using LLMs and need further evaluation. To evaluate their quality in closed-ended tasks, we randomly select 200 questions from MCNC and MCAC and ask two human annotators[7] to evaluate them from multiple dimensions, as outlined below:

- *Clearness* examines whether the samples in the MCNC and MCAC tasks explicitly clarify the event background and questions. For the MCAC task, annotators also need to check the quality of the argument-level questions.

- *Answerable* examines whether the questions and candidates are answerable. Specifically, given a question, annotators should be able to identify the correct answer from four candidates by analyzing the event background.

- *Uniqueness* evaluate the correctness of three noisy candidates. All noisy candidates should be different from the gold answer in wording and semantics, ensuring that only one correct answer is provided.

The annotation consistency ratios between the two annotators in *Clearness*, *Answerable*, and *Uniqueness* are 97.0%, 93.5%, and 98.5% respectively, indicating substantial agreement. The closed-ended questions in OpenForecast achieve a *Clearness* of 97.5%, *Answerable* of 91.0%, and *Uniqueness* of 98.5%, demonstrating the effectiveness of the generated questions.

B Experiments

B.1 LLMs in Experiments

The large language models in our experiments are listed below:

- *Llama2-series* contain multiple instruction-finetuned model in different scale: Llama2-7b-chat[8], Llama2-13b-chat[9], Llama2-70b-chat[10]. The pretraining data has a cutoff of September 2022, ensuring no knowledge leakage.

- *WizardLM-13B-V1.2*[11] empowers LLMs to follow complex instructions by creating large amounts of instruction data with varying levels of

[7] Our annotation team consists of two graduate students engaged in information processing. Another annotator will review their annotation results and eliminate their discrepancies.

[8] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[9] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[10] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

[11] https://huggingface.co/WizardLMTeam/WizardLM-13B-V1.2

complexity and finetuning on Llama2-13b. The pretraining data of Llama2-13b has a cutoff of September 2022, ensuring no knowledge leakage.

- *Vicuna-13b-v1.5*[12] finetunes Llama2-13b on user-shared conversations collected from ShareGPT, ensuring no knowledge leakage.

- *Falcon-40b-instruct*[13] is a chat model based on Falcon-40b with 40B parameters. Since it was released in May 2023, there is no knowledge leakage issue.

- *Mixtral-8x7B-Instruct-v0.1*[14] is a pretrained generative Sparse Mixture of Experts and is released on December 11, 2023. Considering that Llama3-70b is released on April 18, 2024 and its pretraining data has a cutoff of December 2023 (5 months ago), there is no knowledge leakage issue for the evaluation on the testset with a cutoff of September 2023 (3 months ago).

- *Llama3-8b-Instruct*[15] is pretrained on over 15 trillion tokens of data from publicly available sources. Although Llama3-8b was released on April 18, 2024, the pretraining data has a cutoff of March 2023, ensuring no knowledge leakage.

B.2 Human Evaluation for Open-ended Event Forecasting

The human evaluation reflects the true performance of open-ended event forecasting. To evaluate the LLMs on three open-ended tasks, we randomly select 100 samples for each open-ended task and collect the corresponding forecasting results from nine LLMs. After the atomic event partitioning, the number of atomic events reaches approximately 9,000 for both STF and LTF, rendering the human evaluation labor-intensive.

Then, We ask two groups of annotators[16] to judge whether the prediction is correct at the atomic level. Annotators are provided with detailed information, including questions, predictions, gold labels, backgrounds, and original articles. For each sample, annotators must investigate the background, development, aftermath, and reactions of the event via Wikipedia and web searching. Then annotators need to find supporting evidence and determine whether the atomic predictions really occurred. The annotation should follow the following criteria:

- If a prediction can be directly confirmed by specific gold answers, it is annotated as correct, and the corresponding gold answer should be documented.

- If a prediction cannot be confirmed by the gold answers but can be confirmed through external sources such as Wikipedia or reliable web sources, it is also annotated as correct.

- If a prediction cannot be directly confirmed by gold answers or external sources, annotators should judge whether the prediction is reasonable. If the prediction does not contradict existing facts and can be inferred as correct, it is annotated as correct.

- If no evidence supports the prediction and it contradicts existing facts, it will be annotated as incorrect.

After annotation, we calculate the F1-score using the formula (2) in Appendix B.3 for each sample. The average F1-score difference between the two annotation teams are 0.02, indicating substantial agreement. To alleviate the disagreement between the annotation groups, we filter out those samples with the absolute F1-score difference exceeding 0.1. Then, we employ another annotator to review the annotation results and eliminate their discrepancies.

B.3 Open-Ended Evaluation Metric for List-Style tasks

This section introduces the evaluation metric for list-style tasks: LTF and STF, which means each gold answer and prediction contains multiple atomic events. Traditional *precision*, *recall*, and *F1-score* are defined as formula 1. Here, *TP* denotes true predictions that can be verified by given labels, *FP* denotes false predictions that aren't included in given labels, and *FN* denotes missing

[12] https://huggingface.co/lmsys/vicuna-13b-v1.5

[13] https://huggingface.co/tiiuae/falcon-40b-instruct

[14] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[15] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[16] Our annotation team consists of four graduate students engaged in information processing.

labels that are not recalled.

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ F1 &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \end{aligned} \quad (1)$$

However, as discussed above, the misalignment between predictions of LLMs and labels, selective documentation, reporting granularity, etc cause the incomplete labeling issue. Therefore, some true positive predictions are ignored if the labels are incomplete, resulting in lower *precision*, *recall*, and *F1-score* than actual performance. Nevertheless, it is impractical to encompass all atomic events within the gold labels, as even Wikipedia only documents the key events and lacks fine-grained developments. Therefore, we propose a modified method to calculate the precision, recall, and F1-score for STF and LTF as follows. In the **modified formulas**, we additionally introduce *MTP* which denotes missing true predictions that can be verified through external evidence or reasoning rather than given gold labels.

$$\begin{aligned} Precision_{open} &= \frac{TP+MTP}{TP+MTP+FP} \\ Recall_{open} &= \frac{TP+MTP}{TP+MTP+FN} \\ F1_{open} &= 2 \cdot \frac{Precision_{open} \cdot Recall_{open}}{Precision_{open} + Recall_{open}} \end{aligned} \quad (2)$$

B.4 Analysis on Finetuning

We finetuned the Llama3-8b using balanced multi-task datasets. The instruction part includes the prompt template, event background, and gold answer.

As shown in Table 2, although SFT yields substantial improvements in closed-ended tasks, they exhibit performance degradation in open-ended tasks. Based on the analysis in Section 6.2, three primary reasons may contribute to this. (1) Discrepancy between predictions and labels. The original predictions of LLMs differ greatly from the labels in terms of pessimism bias, number of atomic events, text style, etc. (2) Missing true predictions. LLMs show potential in making comprehensive predictions, including multifaceted impacts (politics, economy, and diplomacy, etc), long-term impacts, international reactions. However, some correct predictions require additional web searching or commonsense reasoning for verification but are not included in the original articles and labels. Consequently, these predictions are erroneously regarded as negative during finetuning. The discrepancy and incomplete labeling together introduce significant instability into the finetuning. (3) The gap between closed-ended and open-ended tasks. Unlike closed-ended tasks, open-ended tasks feature unconstrained answer spaces, which might make the multi-task finetuning unbalanced and unstable. To alleviate these two problems, advanced training methods such as process-supervised RL Lightman et al., 2023) and correction-based learning (An et al., 2023) might be promising.

Model	MCNC	MCNC*
Llama2-7b	30.5	25.3
Llama2-13b	42.8	26.0
Llama2-70b	47.0	30.2
Vicuna-13b	47.3	27.0
Wizardlm-13b	45.3	26.1
Falcon-40b	41.0	26.2
Mixtral-8x7b	55.2	36.2
Llama3-8b	55.7	31.8
Llama3-8b-SFT	96.1	74.5
Avg	51.2	33.7

Table 5: The comparison of the structured and unstructured representation format. The MCNC* task represents the structured variant of MCNC.

B.5 Structured or Unstructured?

Based on the MCNC dataset, using LLMs, we perform open event extraction on the background texts and candidate answers to create a structured MCNC dataset. Then we convert the structured events to text as other works (Luo et al., 2024) and employ the same prompt templates as MCNC for evaluation (as shown in MCNC* column). We evaluate the *extraction accuracy* of this open event extraction using LLMs. For each event timeline from A.3, we randomly select one atomic event and its corresponding structured result. Following the open information extraction, the extraction is deemed as correct if the arguments within the atomic event are correctly extracted. The annotation consistency ratio between the two annotators' is 89.5%, indicating substantial agreement. The open event extraction achieves an extraction accuracy of 87.0%. Most errors arise from the extraction omission and the reversal of subject and object, which pose challenges for forecasting within the structured format.

As illustrated in Table 5, there is a dramatic decline across all LLMs, including Mixtral-8x7b

(19.0%), Llama3-8b (23.9%), and Llama3-8b-SFT (21.6%). These results indicate that the **unstructured format outperforms the structured format for LLMs in script event prediction** (unsure for temporal knowledge graph completion).

B.6 Prompt Templates

The prompt templates for six tasks are shown in Table 10. The prompt templates for dataset construction are shown in Table 11 and Table 12.

Type		Definition
Misssing	Semantic	The predictions and labels refer to the same event but differ in textual representation.
	Commensense	The predictions cannot be verified through existing evidences but can be easily inferred from them through commonsense reasoning. For example, "the escalating tension between the two countries" can be easily inferred from the "increased military deployment and confrontation in the border region", even if there is no relevant reports.
	Ongoing	The predictions are ongoing events of given background but are not mentioned in labels and subsequent reports. For example, given event background including assistance, there will be ongoing material and medical assistance until the rescue operation is finished, even if there is no relevant content in subsequent report.
overestimation	Semantic	The prediction and label describe different events but are erroneously matched.
	Pessimistic	The model makes a pessimistic forecasting but actually not. For example, in testing sample about "2023 Chitral cross-border attacks", the prediction is "The Pakistani military will launch a major offensive against TTP insurgents in the border regions, resulting in a significant increase in violence and displacement". In contrast, the real development is "Pakistani ambassador Asif Durrani led a visit of delegation to Afghanistan and Afghan authorities took actions to mitigate the conflict".
	plausible	The prediction is plausible, but not happen according to the event background and relevant reports.

Table 6: The definitions of the error types in open-ended evaluation methods.

Task	**2023 Chitral cross-border attacks**
Extracted Timeline	(1) On 7 September 2023, the TTP attacked five military checkposts from Afghan territory, killing 4 Pakistani soldiers. (2) On 7 September 2023, the Pakistani military thwarted the attack and killed 12 TTP insurgents. (3) Following the attack, Pakistan's primary Torkham border crossing with Afghanistan remained closed for a second consecutive day on 7 September 2023, resulting in a backlog of trucks carrying goods. (4) The Pakistani army initiated a combing operation in response to the attacks, targeting the neighboring Afghan provinces of Kunar and Nuristan. (5) On 8 September 2023, Pakistan summoned the head of the Afghan diplomatic mission in Islamabad to lodge a robust protest concerning the attacks. (6) On 8 September 2023, the Foreign Office (FO) issued a stern demarche to the Afghan interim government, emphasizing that such attacks and the utilization of Afghan territory by the TTP would pose a significant threat to peace and security. (7) The Caretaker Foreign Minister, Jalil Abbas Jilani, described the recent terrorist attack on two military posts in the Lower Chitral district of Khyber Pakhtunkhwa as an ïsolated incident.̈
Prediction on 8 September	(1) The TTP will continue to carry out cross-border attacks from Afghanistan into Pakistan. (2) The Pakistani military will respond with retaliatory strikes against TTP hideouts in Afghanistan. (3) The Pakistani government will request the Afghan government to take action against TTP insurgents operating from its territory. (4) The Afghan government will deny the presence of TTP insurgents on its soil and refuse to take action. (5) The Pakistani government will seek international support to pressure the Afghan government to take action against TTP insurgents. (6) The TTP will intensify its attacks on Pakistani military and civilian targets, leading to increased tensions between Pakistan and Afghanistan. (7) The Pakistani military will launch a major offensive against TTP insurgents in the border regions, resulting in a significant increase in violence and displacement. (8) The international community will call for a ceasefire and a negotiated settlement to the conflict. (9) The Pakistani government and the TTP will enter into peace talks, but the talks will break down due to irreconcilable differences. (10) The conflict will continue to simmer, with occasional outbreaks of violence, for the foreseeable future.

Table 7: An example of Mixtral-8x7b on STF task. The model makes pessimistic forecasting while actually not.

Model	STF						LTF					
	LRAE w/o RA			LRAE			LRAE w/o RA			LRAE		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama2-7b	9.6	51.7	15.7	47.5	83.7	57.9	11.7	42.3	17.0	39.7	64.3	46.3
Llama2-13b	11.9	57.4	18.9	47.0	81.4	56.6	11.0	34.0	15.2	33.3	54.6	38.5
Llama2-70b	11.1	61.0	18.3	42.7	81.7	57.2	10.9	45.7	16.4	41.4	69.5	48.2
Vicuna-13b	7.3	35.9	11.5	30.2	59.0	36.7	6.5	18.3	9.0	25.5	42.7	29.6
Wizardlm-13b	11.3	35.0	15.9	40.3	57.6	44.5	6.7	23.6	9.5	26.6	42.6	29.7
Falcon-40b	11.0	19.7	13.1	30.3	35.1	31.6	9.6	20.6	11.2	31.9	41.4	32.7
Mixtral-8x7b	10.6	49.2	16.4	42.8	74.2	51.3	8.0	33.2	11.9	35.4	54.1	39.8
Llama3-8b	19.9	50.0	27.0	58.7	69.5	60.3	13.6	34.2	17.9	38.0	55.4	43.0
SFT	24.0	17.9	19.7	39.6	25.0	29.4	10.1	14.1	10.5	12.9	16.2	13.2

Table 8: Evaluation results on STF and LTF tasks using LRAE evaluation. *RA* denotes the retrieval augmentation module.

Model	STF						LTF					
	BEM			BEM+RA			BEM			BEM+RA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama2-7b	2.5	12.0	4.0	11.8	32.4	16.6	4.6	19.2	6.9	14.5	33.6	18.3
Llama2-13b	1.6	8.4	2.7	11.5	25.3	14.6	5.2	20.4	7.7	13.9	30.0	17.3
Llama2-70b	1.7	8.7	2.8	11.4	28.9	15.3	8.7	36.8	13.0	13.6	43.8	19.1
Vicuna-13b	1.6	7.7	2.3	11.6	21.4	14.0	2.2	6.8	3.2	15.7	22.2	16.7
Wizardlm-13b	2.0	5.8	2.9	14.9	21.9	16.5	4.0	7.9	4.8	15.6	22.4	17.4
Falcon-40b	2.0	3.7	2.3	10.2	12.8	10.5	3.0	6.5	3.6	17.4	21.1	17.5
Mixtral-8x7b	2.3	8.3	3.5	18.6	31.8	21.9	1.5	6.1	2.3	10.8	21.5	13.5
Llama3-8b	6.0	6.9	6.3	21.9	18.7	19.1	1.9	2.8	2.2	11.9	13.9	12.3
SFT	13.5	13.9	13.4	13.6	13.9	13.4	36.5	43.9	35.0	38.0	44.9	36.0

Table 9: Evaluation results on STF and LTF tasks using BEM evaluation. *RA* denotes the retrieval augmentation module.

Task	Prompt
AQA	Based on the given event background, please answer the given question with the most likely answer. The output format should be Brief Analysis: xxx \n Answer: xxx \n Event background: \n {background} \n Question: {question} \n Brief Analysis and Answer:
STF	Based on the given event background, please predict the most likely events to occur at the given time. Note that the event description should be brief, accurate, and complete, especially the name (replace the pronouns with the corresponding name). \n **Example:** \n *Event background: \n1. Starting from March 2021, Ukrainian forces intensified their troop deployments on the frontline and began frequent clashes with Donbas militants.\n 2. From March to April 2021, Russia began a large-scale military buildup near the border.\n 3. From October 2021 to February 2022, Russia and Belarus carried out a second buildup. \n 4. Throughout, the Russian government repeatedly denied it had plans to attack Ukraine.\n 5. On 21 February at 22:35, Putin announced that the Russian government would diplomatically recognize the Donetsk and Luhansk People's Republics.\n 6. The same evening, Putin directed that Russian troops deploy into Donbas.\n 7. On 22 February, the Federation Council unanimously authorized Putin to use military force outside Russia. \n 8. The following day, Ukraine's parliament proclaimed a 30-day nationwide state of emergency and ordered the mobilisation of all reservists.\n 9. Russia began to evacuate its embassy in Kyiv.\n The most likely events on February 24 is: 1. Russia will launch a full-scale invasion of Ukraine.\n 2. Zelenskyy will declare martial law and a general mobilisation of all male Ukrainian citizens between 18 and 60, who will be banned from leaving the country.*\n\n Now, based on the following event background, please predict the events that are most likely to occur on {time}.\n Event background: \n{background} \n The most likely events on {time} are:
LTF	Based on the given event background, please predict the event timeline that is most likely to occur in the future. \n **Rules:** \n 1. The predicted events should be output in in chronological order.\n2. The event descriptions for each event should be brief, accurate, and complete, especially names (replace pronouns with their corresponding names). \n 3. The events output format should be: 1. brief predicted event; 2. brief predicted event; etc. \n **Example:** \n *Event background: \n1. Starting from March 2021, Ukrainian forces intensified their troop deployments on the frontline and began frequent clashes with Donbas militants.\n2. From March to April 2021, Russia began a large-scale military buildup near the border.\n3. From October 2021 to February 2022, Russia and Belarus carried out a second buildup.\n4. Throughout, the Russian government repeatedly denied it had plans to attack Ukraine.\n Predicted Event Timeline: \n1. Putin will announce that the Russian government would diplomatically recognize the Donetsk and Luhansk People's Republics.\n2. Putin will direct Russian troops to deploy into Donbas.\n3. The Federation Council will unanimously authorise Putin to use military force outside Russia.\n4. Zelenskyy will order the conscription of army reservists. \n5. Ukraine's parliament will proclaim a 30-day nationwide state of emergency and orders the mobilisation of all reservists.\n6. Russia will launch a full-scale invasion of Ukraine.*\n\n Now, based on the following event background, please predict the event timeline that is most likely to occur in the future.\n Event background: \n{background} \n Predicted Event Timeline:
MCNC	Based on the given event background, please select one option from the four candidate event options that is most likely to occur in the future. Your final answer should be a single option letter, in the form (option letter) such as (A), at the end of your response. \n Event background: \n{background} \n Event options: \n{options}\n Answer:
MCAC	Based on the given event background and question, please select the most likely option from the four candidate options. Your final answer should be a single option letter, in the form (option letter) such as (A), at the end of your response. \n Event background: \n{background} \n Question: {question}\n Event options: \n{options}\n Answer:
VQA	Based on the given event background, please predict whether the given candidate event will occur with high probability. Your final answer should be Yes or No, at the end of your response. \n Event background: \n{background} \n Candidate Event: \n{options}\n Answer:

Table 10: Prompt templates for the evaluation on six tasks.

Task	Prompt
Category Filtering	Please categorize the given Wikipedia article. Candidate categories include: [People, Country, Region, Organization, Sports Competition, Entertainment, item Definition, Holiday, Event].\n Wikipedia article title: {title}\n The article: {article}\n Please determine the category of the given Wikipedia article. The prediction category from [People, Country, Region, Organization, Sports Competition, Entertainment, item Definition, Holiday, Event] is:
Date Filtering	Please extract the occurrence time of the following event. **Rules:** \n1.The time format must follow "yyyy", "yyyy-MM", or "yyyy-MM-dd", where yyyy, MM, and dd refer to the year, month, and day respectively. \n2.Notice, the dates should be as precise as possible. If only the year is known, use the "yyyy" format. If the year and month are known, use the "yyyy-MM" format.\n3.If the event spans a certain period of time, please output the end time of the event.The Wikipedia title: {title}\n The article:{article}\n The year of the occurrence of the event is:
Event Timeline Extraction	You are a event timeline assistant. Based on the following article, please extract the complete event timeline in the article and sequentially output the key objective events in chronological order. The event timeline should contain the background, development, aftermath, investigation, and reactions if they are introduced in the article. If there is no content related to the event timeline, please directly output "None". \n The article title: {title}\n The article: {article}\n The events timeline is:
Event Timeline Completion	You are an event timeline completion assistant. Please complete the given event timeline based on the following news. **Rules:** \n 1. New events should only be supplemented into the given event timeline, without altering the descriptions of events already specified within the timeline.\n 2. Only add objectively occurring key events that are not included in the given event timeline.\n 3. The newly added events should be limited to a small number of key events, excluding irrelevant events and subjective events.\n 4. The newly added events encompass events preceding, during, and subsequent to the given chain of events, which should be added to the front, middle, and end of the timeline respectively, keeping the completed event timeline in chronological order.\n 5. Event descriptions should be brief, accurate, and complete, especially for numbers, names (replace pronouns with their corresponding names), and dates.\n 6. If there are no new events, simply output the given event timeline.\n 7. The events output format should be: 1. brief event; 2. brief event; etc. \n The given event timeline before {time} is: {event chain}\n Now, given the following news: \n The News release time: {date}\n The News title:{title} \n The news article: {article}\n The completed event timeline in chronological order is:
Event Timeline Adjustment	Please adjust the sequence of events according to the given article to correct any chronological errors in given event timeline, ensuring that the events are arranged in chronological order \n The article title: {title}\n The article: {article}\n The original event timeline: {timeline}\n The adjusted event timeline is:
Open Event Extraction	Please extract structured event information from the given event text. The **rules** are as follows: \n1. Structured event information includes event trigger words and parameters such as time, location, subject, and object. If the above parameters are not enough to express the event clearly, you can also use other elements such as announcement content, condemnation content, boycott content, and other elements (excluding causes and consequences); \n 2. Event elements such as subject and object should use full names instead of Pronouns (such as he, she, they, etc.);\n3. Event text can contain multiple structured events. Please output events in chronological order. If there are no structured events, output None. The format is as shown in the example. \n An example is as follows: {exemple} \n Now, please perform event extraction on the following event text. Event text:{text}\n\nThe extracted structured event list is:\n

Table 11: Prompt templates for the dataset construction of OpenForecast.

Task	Prompt
Noisy Candidates Generation for MCNC	According to the given known event timeline, subsequent evolution events, and one true candidate event, please generate three candidate noisy events for the given true candidate event to the event forecasting problem. \nThe **rules** are as follows: \n1. The true candidate event comes from the given subsequent evolution events. The given subsequent evolution events are used to prevent the generation of candidate noisy events that actually occur. \n 2. The three candidate noisy events should be challenging and similar to the given true candidate events. And the three candidate noisy events should occur at the same time as the true candidate event, but they should explicitly not occur (according to the known events timeline and subsequent evolution events). \n3.You can generate candidate noisy events by replacing the elements of the true candidate event with other arguments in the given known event timeline and subsequent evolution events. The replaceable event elements include event type, subject, object, and location, but not time. \n4. The output format should be: Noisy Event A: xxx\n Noisy Event B: xxx\n Noisy Event C: xxx\n\n An example is provided below: \n {exemple} \n\nNow, based on the above rules and example, please generate three candidate noisy answers. The known event timeline: {timeline} \nThe subsequent evolution events: {evolution} \nThe true candidate event:{true event}\nThe three candidate noisy answers are:
Question Generation for MCAC	According to the given known event timeline, the target evolution event, and other subsequent evolution events, please design an event prediction question on the {args} argument of the given target evolution event, and generate the true answer and three candidate noisy answers for this question. The **rules** are as follows:\n1. The generated question should focus on the {args} argument of the given target evolution event and include all other arguments of the given target evolution event except for the {args} argument.\n2.Based on the generated question, extract the true answer from the target evolution event.\n3.The three candidate noisy answers should be challenging and you can involve event argument from the given event timeline or other subsequent evolution events, but the noisy answers should explicitly not occur (not occur in both given event timeline and other subsequent evolution events).\n4.The output format should be as follows: \nEvent argument prediction question: xxx\n True answer: xxx\n Noisy answer A: xxx\n Noisy answer B: xxx\n Noisy answer C: xxx\n Example 1:\n {exemple} \n Now, based on the above rules and examples, please generate a {args} argument prediction questions and candidate answers for the following event: \n Known event timeline: \n{timeline \n Target evolution event: \n {true event} \n Other subsequent evolution events: \n{evolution} \n Generated event {args} argument prediction question and candidate answers:

Table 12: Prompt templates for the question generation of MCNC and MCAC.

Task	Prompt
Stage 1 Prompt for AQA	Question: {question} \nLable: {label} \n Prediction: {prediction} \nPlease judge whether the prediction is correct. If the prediction matches the label clearly, then the prediction is correct, otherwise wrong. The answer (Yes or No) is: "
Stage 1 Prompt for STF and LTF	Following natural language inference, please judge whether the prediction is in accord with the given labels. **Rules:** \n1. Based on the given labels only. \n2.Note that the tense of the prediction should be disregarded. \n3.If certain label can clearly demonstrate the prediction, then the prediction is correct, otherwise wrong. \n4. If the prediction is wrong, output "No". If the prediction is correct, please output "Yes" along with the corresponding index in the list of labels, for example: "According to the given information, the answer is Yes, and the index is: (1)". \nThe labels are {label}\n The prediction is {prediction} The answer is:
Stage 1 Prompt for STF and LTF	Following natural language inference, please judge whether the prediction is in accord with the given label. Rules: \n1. Based on the given label only. \n2.Note that the tense of the prediction should be disregarded. \n3.If certain content from label can clearly demonstrate the prediction, then the prediction is correct, otherwise wrong. \nThe label is {label}\n The prediction is "{prediction}" The answer (Yes or No) is:

Table 13: Prompt templates for LLM-based evaluation. The STF and LTF adopt the RAE with two-stage evaluation. The stage 1 (step 3 in 5) of RAE evaluates the prediction with gold answers. The stage 2 (step 5 in 5) evaluates the prediction with retrieved web contents.