

# OpenEP: Open-Ended Future Event Prediction

Yong Guan  
Tsinghua University  
Beijing, China  
gy2022@mail.tsinghua.edu.cn

Hao Peng  
Tsinghua University  
Beijing, China  
peng\_21@mail.tsinghua.edu.cn

Xiaozhi Wang  
Tsinghua University  
Beijing, China  
wangxz20@mail.tsinghua.edu.cn

Lei Hou  
Tsinghua University  
Beijing, China  
houlei@tsinghua.edu.cn

Juanzi Li  
Tsinghua University  
Beijing, China  
lijuanzi@tsinghua.edu.cn

## Abstract

Future event prediction (FEP) is a long-standing and crucial task in the world, as understanding the evolution of events enables early risk identification, informed decision-making, and strategic planning. Existing work typically treats event prediction as classification tasks and confines the outcomes of future events to a fixed scope, such as yes/no questions, candidate set, and taxonomy, which is difficult to include all possible outcomes of future events. In this paper, we introduce **OpenEP** (an Open-Ended Future Event Prediction task), which generates flexible and diverse predictions aligned with real-world scenarios. This is mainly reflected in two aspects: firstly, the predictive questions are diverse, covering different stages of event development and perspectives; secondly, the outcomes are flexible, without constraints on scope or format. To facilitate the study of this task, we construct **OpenEPBench**, an open-ended future event prediction dataset. For question construction, we pose questions from seven perspectives, including time, location, event development, event outcome, event impact, event response, and other, to facilitate an in-depth analysis and understanding of the comprehensive evolution of events. For outcome construction, we collect free-form text containing the outcomes as ground truth to provide semantically complete and detail-enriched outcomes. Furthermore, we propose **StkFEP**, a stakeholder-enhanced future event prediction framework, that incorporates event characteristics for open-ended settings. Our method extracts stakeholders involved in events to extend questions to gather diverse information. We also collect historically events that are relevant and similar to the question to reveal potential evolutionary patterns. Experiment results indicate that accurately predicting future events in open-ended settings is challenging for existing LLMs. In addition, we thoroughly summarize the problems encountered in prediction, hoping to provide insights for future research.

## CCS Concepts

- Computing methodologies → Information extraction.

## ACM Reference Format:

Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2024. OpenEP: Open-Ended Future Event Prediction. In . ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/nnnn.nnn>

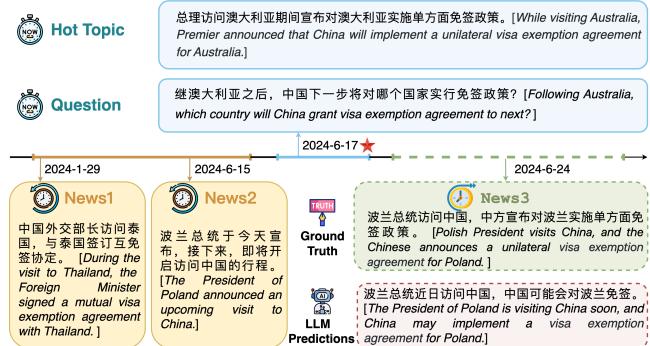


Figure 1: Example of Future Event Prediction.

## 1 Introduction

Future event prediction (FEP) aims to predict the potential future outcomes based on the historical events and the precursors [33]. In Figure 1, given the predictive question “*Which country will China grant visa exemption agreement to next?*”, a FEP model needs to collect historical events, such as “*The President of Poland announced an upcoming visit to China*”, to predict the outcomes of future events like “*Poland*”. Accurate anticipation of future events is crucial in the modern world, as it provides scientific support for making more rational and efficient decisions and possesses significant practical and application value. Such as, in fields law [21], finance [29], and healthcare [4], event prediction technology helps identify potential risks and uncertainties, enhancing safety and risk management capabilities through logical analysis and predicting.

Early research mainly utilizes statistical machine learning methods, fitting predefined statistical models with historical data for prediction [8, 12]. These methods often required the integration of domain knowledge and involved complex feature engineering. With the rapid advancement of big data and deep learning technologies, predicting through data-driven neural networks has emerged as an appealing alternative [13, 15, 35]. Especially in recent years, the emergence of large language models (LLMs) [5, 18] have exhibited astonishing performance in tasks previously thought to require human cognitive abilities. Despite some work focusing on LLM-based event prediction [7, 19, 23, 30], **open-ended future event prediction task is still being neglected**. Existing work typically treats event prediction as classification tasks and confines the outcomes of future events to a fixed scope, such as yes/no questions [7, 23, 30],

candidate set [2, 16, 34], and taxonomy [13]. The fixed scope of outcomes typically comprises a uniform format, such as single words or short phrases, resulting in predicted outcomes that often lack rich semantics and details. In contrast, freely generated outcomes in real-world usually contain longer and semantically complete responses, enriched with more details.

In this paper, we introduce **OpenEP** (an Open-Ended Future Event Prediction task), which generates more flexible and diverse predictions aligned with real-world scenarios. This is mainly reflected in two aspects. (1) The predictive questions are diverse, covering different stages of event development and perspectives, facilitating a comprehensive analysis. (2) The outcomes are flexible, without constraints on scope, format, or length, which can provide semantically complete responses enriched with more details. To facilitate the study of this task, we first construct **OpenEPBench**, an open-ended future event prediction dataset, and concurrently propose **StkFEP**, a stakeholder-enhanced future event prediction framework for open-ended settings.

For OpenEPBench construction, we need to address the following three key questions: (*Q1*) How to determine the data source? (*Q2*) How to generate predictive questions? (*Q3*) How to annotate the outcomes of future events? *For data source*, we select hot topics from two platforms widely used for discussing daily events, utilizing Zhihu for Chinese data and Google News for English data. *For question generation*, we pose questions from seven perspectives, including location, time, event development, event outcome, event impact, event response, and other, to facilitate in-depth analysis and understanding of the comprehensive evolution of events. *For outcome annotation*, due to the advanced comprehension capabilities of LLMs, they can evaluate event predictions from a semantic perspective just like human evaluators. Therefore, we extract segments from the original texts containing outcomes as ground truth, without employing a fixed scope or format. In addition, we design corresponding LLM-based evaluation metrics, which measure the predictions from five dimensions, including accuracy, completeness, relevance, specificity, and reasonableness.

For framework StkFEP, it contains three modules: *Retrieval*, *Integration*, and *Prediction*. The *Retrieval* aims to collect the diverse information from news sources to mitigate the semantic gap between the question and predictions. The evolution of events depends on salient entities involved, regard as *stakeholders*. Knowing these entities aids in question expansion and facilitates the retrieval of diverse information. For instance, in Figure 1, given the stakeholders *China* and *premier* can effectively retrieve the news “*The President of Poland announced an upcoming visit to China*”. Thus, we extract stakeholders to extend original questions to gather the diverse information. In addition to retrieving news directly related to the question, referred to as *relevant events*, we also collect historically occurred events that are similar to the question, known as *similar events*. These similar events can serve as references for predicting future events. For example, by considering news 1, it can be inferred from news 2 whether China will sign a visa exemption agreement with Poland. The *Integration* module employs clustering method to clarify the dependencies between events and reduce redundant information. At last, the *Prediction* aims to predict the outcomes based on the information of relevant and similar events.

During the testing phase, tests are conducted immediately after daily question annotations are completed to minimize the risk of information leakage. To summarize our main contributions:

- We introduce **OpenEP**, an open-ended future event prediction task that generates flexible and diverse predictions aligned with real-world scenarios.
- We construct **OpenEPBench**, an open-ended future event prediction dataset with diverse predictive questions and flexible outcomes, facilitating comprehensive analysis. In addition, we design LLM-based metrics to evaluate the model predictions.
- We propose **StkFEP**, a stakeholder-enhanced future event prediction framework that incorporates the characteristics of event evolution for open-ended settings.
- Extensive experiments demonstrate that accurately predicting future events in open-ended settings is challenging for existing LLMs. Furthermore, we have thoroughly summarized the problems encountered in prediction.

## 2 OpenEPBench

In this section, we will describe the OpenEPBench dataset. First, we introduce the overall dataset construction (Sec 2.1). Next, we introduce the construction process, which includes the data source (Sec. 2.2), constructing the predictive questions (Sec. 2.3), and their corresponding outcomes (Sec. 2.4). Once the dataset is built, we analyze its distribution and perform quality checks (Sec. 2.5). Finally, we introduce the evaluation metrics (Sec. 2.6). Further construction and annotation details, including annotation interfaces, examples, and annotation guidelines, are provided in Appendix C.

### 2.1 Overview

This section aims to introduce the overall dataset construction procedure. Our goal is to build an open-ended FEP dataset featuring diverse predictive questions and flexible outcomes that are unconstrained by scope or format. Predictive questions are annotated on a daily basis, and outcomes are collected at future time points. The model is tested daily after the predictive questions are constructed, and it is evaluated when the outcomes are collected.

Prediction window is the time span from the current moment into which future events or values are projected. Considering that people’s attention to hot topics generally lasts around 7 days [11] and that attention to major health emergencies can extend to over 13 days [14]. Therefore, we set the prediction window to 15 days, predicting events that might occur within the next 15 days.

Data annotation, while labor-intensive and costly due to the substantial resources and domain expertise required, ensures high accuracy. The advent of advanced LLMs, exemplified by GPT-4 [18], offers a transformative opportunity for the data annotation process. Consequently, we employ a combination of LLMs and human verification to construct the dataset. The LLMs automate initial annotations, significantly reducing manual labor, while human checks ensure the accuracy and relevance of the annotations. The data construction process consists of two stages: Question Construction and Outcome Construction. For *Question Construction*, collect daily hot topics, use the LLMs to generate multiple potential predictive questions for each hot topic, and manually validate and filter these

Field	Content
Hot Topic	总理访问澳大利亚期间宣布对澳大利亚实施单方面免签政策。[While visiting Australia, Premier announced that China will implement a unilateral visa exemption agreement for Australia.]
Background	中方把澳大利亚纳入单方面免签国家。双方同意互为旅游、探亲人员审发多次入境签证。[China has included Australia in the list of countries eligible for unilateral visa exemptions, and both sides have agreed to issue multiple-entry visas for tourists and family visitors.]
Questions	<p>澳大利亚游客入境数量的高峰期会出现在哪个时间段? [During which period does the peak of Australian tourist arrivals occur?]</p> <p>中国下一步将对哪个国家实行免签政策? [Which country will China grant visa exemption agreement to next?]</p> <p>访华旅游的人数会如何变化? [How will the number of people traveling to China change?]</p> <p>澳大利亚游客对来华旅游的满意度如何? [How satisfied are Australian tourists with their travel to China?]</p> <p>免签政策对中国旅游也会造成什么影响? [What impact will the visa exemption agreement have on Chinese tourism?]</p> <p>国内旅游业对免签政策将作何反应? [How will the domestic tourism industry react to the visa exemption agreement?]</p> <p>免签政策如何促进中澳之间的文化交流? [How does the visa exemption promote cultural exchange between China and Australia?]</p>
Key Dates	Question Date: 2024-06-17   Prediction Window: 15

Figure 2: Example from the OpenEPBench dataset.

questions. For *Outcome Construction*, after the prediction window for individual question, use LLMs to collect news within the period, score the news articles, and then manually validate the event outcomes.

## 2.2 Data Source

Data Source aims to identify and select reliable and relevant sources of data. To construct a prediction dataset that aligns with real-world scenarios, we utilize news data from the internet, which is continuously updated and comprehensive. LLMs are trained with vast amounts of data across various languages, giving them multilingual understanding capabilities. However, their performance still varies across different languages. Therefore, to test the robustness of these models, we have constructed separate datasets in Chinese and English. We focus on constructing predictive questions based on daily hot topics. We have chosen two widely used platforms for this purpose: Zhihu<sup>1</sup> for Chinese data, where daily discussions on hot topics are directly utilized as hot topics, and Google News<sup>2</sup> for English data, where the headlines of news are regard as hot topics.

## 2.3 Question Construction

Question Construction aims to build predictive questions based on hot topics. The overall process involves collecting hot topics, using LLMs to generate candidate questions, and then manually verifying the candidate questions.

Events evolve dynamically and undergo various phases. By formulating predictive questions from multiple perspectives, we can conduct a more thorough analysis and understanding of the event evolution. Therefore, incorporating the elements of an event 5W1H (who, where, when, what, why, how) [24], along with external feedback, we propose posing questions from seven perspectives about hot topics, as shown in Figure 2.

- Time. The date on which a future event is likely to occur.

<sup>1</sup><https://www.zhihu.com/knowledge-plan/hot-question/hot>

<sup>2</sup><https://news.google.com/topics>

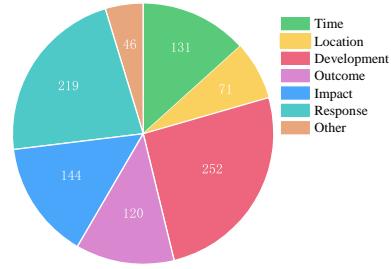


Figure 3: Data distribution across different types of questions.

- Location. The specific place or location where a future event will occur.
- Event Development. The progression of a future event, including how the event unfolds, potential movements, or turning points.
- Event Outcome. The direct results or outcomes after a future event has concluded.
- Event Impact. The impact of a future event on the surrounding environment, economy, or other relevant sectors.
- Event Response. The reactions of different stakeholders to an event, including the public, governments, markets, or specific groups' behavioral and emotional responses.
- Other. Any additional aspects or perspectives that require further clarification.

**Construction process.** Question construction employs a combination of LLMs and manual verification, comprising the following five steps: (1) *Hot Topics Collection*. Collect daily hot topics from Zhihu and Google News. (2) *Background Collection*. Hot topics collected online include a background typically generated from a single news article. To enrich the background, we retrieve related news articles using the hot topic and use LLMs to regenerate a background for supplementation. (3) *Hot Topic Validity Verification*. Not all hot topics are suitable for generating questions. For instance, many hot topics may involve discussions of past events that have resurfaced. Hence, based on LLMs, we verify each hot topic for aspects such as continuity, initially filtering out those that do not meet the criteria. (4) *Candidate Question Generation*. For each hot topic, the LLM generates multiple predictive questions from the previously mentioned seven perspectives. Each perspective may contain multiple questions. (5) *Human Verification*. Manually verify questions generated by the LLM based on answerability, specificity, and real-time, selecting suitable predictive questions.

## 2.4 Outcome Construction

Outcome Construction aims to collect the ground truth for the corresponding predictive questions. The overall process involves using LLMs to collect news, score the articles, and manually verify the outcomes. The outcomes of future events are not constrained by fixed scopes, formats, or lengths. Relying solely on manual generation of event outcomes undoubtedly increases the workload. Interestingly, LLMs have inherent strengths in comprehension and generation, effectively grasping contextual semantic information. Currently, LLMs are widely used to evaluate model performance [17, 32],

providing a semantic perspective that surpasses traditional assessments like ROUGE or BLEU, which measure word co-occurrence. Therefore, we extract segments from news articles that contain the outcomes as ground truth.

**Construction process.** Outcome construction utilizes a combination of LLMs and manual verification, consisting of the following three steps: (1) *News Collection*. For each question, retrieving news from news sources within the prediction window. (2) *News Rerank*. Each news article is scored by LLMs to assess if it contains the event outcomes, with higher scores indicating a greater probability. The news articles are then ranked by the scores, and the valid news are selected as candidates. (3) *Human Verification*. The selected news articles are manually verified to extract segments containing the event outcomes, forming the final ground truth.

## 2.5 Data Analysis

**Data quality.** To ensure data quality, for automatic annotation, we select GPT-4, currently the best-performing LLM. For human annotation, we invited two experts with PhDs in natural language processing to help check the data. The specific process involves each individual independently verifying the results from the LLM during both the question construction and outcome construction phases. After validation, any inconsistencies in the data are discussed between the two experts. Data agreed upon through discussion is accepted, while data with unresolved discrepancies is discarded.

**Data Distribution.** For Chinese data, from June 1, 2024, to July 10, 2024, we collect 192 valid hot topics over a 40-day period and generate 869 valid predictive questions, with an average of 4.52 predictive questions per hot topic. For English data, from July 1, 2024, to July 10, 2024, we collect 27 valid hot topics over a 10-day period and generate 114 valid predictive questions, with an average of 4.22 predictive questions per hot topic.

The questions in our dataset cover a very wide variety of perspectives. We design seven types of predictive questions, including time, location, event development, event outcome, event impact, event response, and other. Figure 3 shows the data distribution across each predictive question category.

## 2.6 Evaluation Metrics

The predictive questions include seven types: time, location, event development, event outcome, event impact, event response, and other. When evaluating time-type questions, we convert the question into a multiple-choice format, using accuracy as the evaluation criterion. Specifically, the prediction window is divided into three periods, each covering five days, plus an additional option indicating no outcome, creating a total of four options. The prediction model must output one of these options as the result.

Apart from time-type prediction questions, the outcomes for other types of questions are presented in free-form text, without constraints on scope, format, or length. Traditional automatic evaluation metrics, which measure word co-occurrence, are no longer suitable, necessitating a more human-like approach to assessment from a semantic perspective. More recently, LLMs exhibit astonishing performance in tasks previously thought to require human cognitive abilities and are increasingly used to evaluate model performance. Inspired by existing work [17, 32], we utilize LLMs, such

as GPT-4, to evaluate event prediction performance from the following five dimensions:

- Accuracy. Measures the extent to which the predicted content matches the actual outcomes or states that occurred, with a primary focus on the precision of the predictions.
- Completeness. Assesses whether the prediction covers the different relevant aspects of the actual outcomes, evaluating the thoroughness of the information provided.
- Relevance. Evaluates how pertinent the prediction is to the actual outcomes, ensuring that the prediction does not veer into unrelated details.
- Specificity. Analyzes the sharpness and focus of the prediction, ensuring that it is neither overly broad nor vague.
- Reasonableness. Measures the logical coherence and believability of the prediction, checking whether the prediction aligns with general world knowledge and appears plausible.

When measuring *Accuracy*, for each prediction question, the actual outcomes may contain multiple aspects of information. If the prediction hits at least one aspect, it scores 1; otherwise, it scores 0. For *Completeness*, the score is calculated as the proportion of accurately predictions relative to the actual outcomes. For the other three dimensions—*Relevance*, *Specificity*, and *Reasonableness*—scores for each dimension are on a scale from 1 to 5, where higher scores indicate better performance. Existing research indicates that LLMs can be overconfident [26, 28]. Therefore, when LLMs provide scores, we require them to also offer probabilities. The final score for each dimension is calculated as follows:

$$\text{score} = \sum_{i=1}^n \sigma(s_i) * p(s_i) \quad (1)$$

where  $\sigma(\cdot)$  aims to map the score to the range 0-1,  $p(\cdot)$  represents the probability of the score, and  $n$  denotes the size of the data.

Ultimately, we aggregate the scores from different dimensions and question types to gauge overall performance. In addition to automatic evaluation, we also conduct human evaluations on a subset of the entire dataset. Similar to automatic evaluation, human evaluators provide scores from the aforementioned dimensions.

## 3 StkFEP

In this section, we introduce StkFEP, a stakeholder-enhanced future event prediction framework for open-ended settings. We first present the task description (Sec. 3.1). Next, we detail the StkFEP framework, comprising three modules: Retrieval (Sec. 3.2), Integration (Sec. 3.3), and Prediction (Sec. 3.4), as depicted in Figure 4.

### 3.1 Task Description

Each item  $x_i$  in the dataset  $\mathcal{D}$  can be represented as a quintuple  $x_i = (q, t, w, b, o)$ , where  $q$  is the predictive question,  $t$  is the time the question is built, noted as a timestamp<sup>3</sup>,  $w$  is the prediction window,  $b$  is the background of the question, and  $o$  is the actual outcomes of the question. Relevant events *SE* refer to news information that is directly related to the question. Similar events *RE* are historically occurred events that are similar to the question. Assume  $q'$  is the expanded set of question  $q$ ,  $f$  is the prediction system, and  $o'$  represents the predicted outcomes of  $f$ .

<sup>3</sup>Each timestamp represents a day, formatted in “YYYY-MM-DD”.

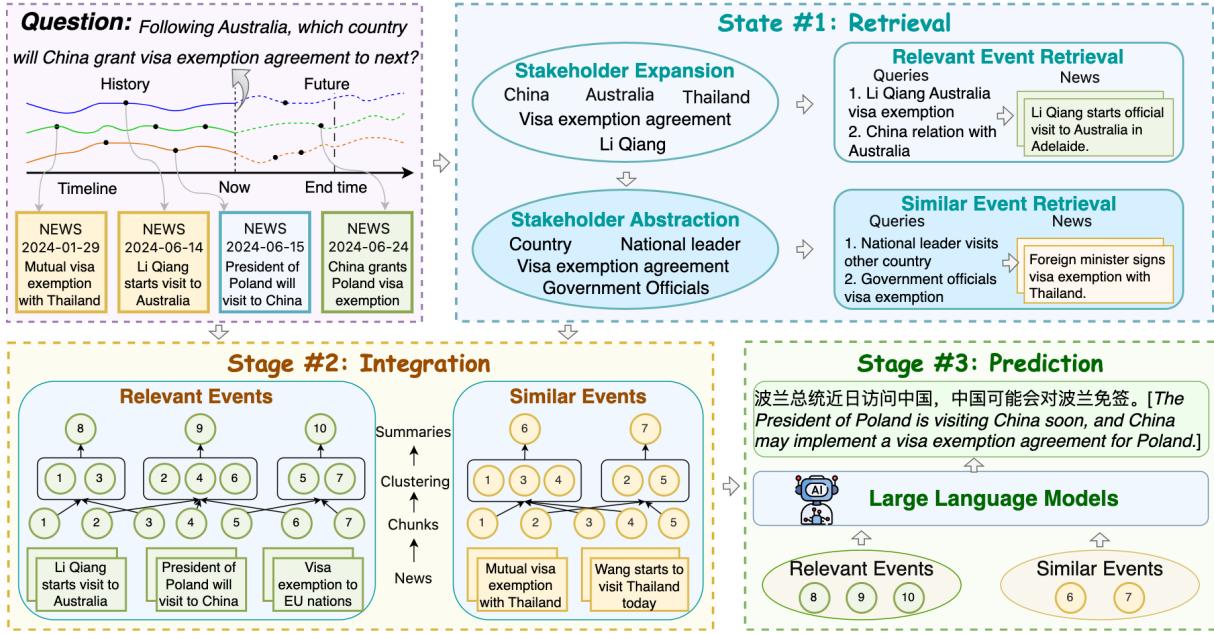


Figure 4: The framework of StkFEP.

Given  $q$  and  $b$ , the prediction system  $f$  is required to expand  $q$  into  $q'$  at time  $t$ , using both  $q$  and  $q'$  to retrieve  $SE$  and  $RE$  from news sources. Based on  $SE$  and  $RE$ , the system predicts potential outcomes  $o'$ . The performance of the model is then evaluated after constructing  $o$  following the prediction window  $w$ .

### 3.2 Retrieval

The Retrieval module aims to collect diverse information from news sources to support the prediction. It consists of 3 steps: question expansion, relevant event retrieval, and similar event retrieval.

**Question Expansion.** This module aims to expand the original question to facilitate the retrieval of diverse information. The information retrieved using the original question is insufficient for event prediction, necessitating the expansion of the question. However, existing methods mainly focus on the capabilities of LLMs, allowing these models to autonomously generate multiple questions while overlooking the characteristics of event. The evolution of events depends on salient entities involved, regard as *stakeholders* [10]. Knowing these entities aids in question expansion and enhances information retrieval. Therefore, we extract stakeholders to extend the original questions and gather comprehensive information. Specifically, we first use the original question to retrieve news from news sources and prompt the LLM to assess the relevancy and filter out irrelevant news. Then, extracting stakeholders from each news article. Based on the original question, background, and stakeholders, we use the LLM to generate various questions.

**Relevant Event Retrieval.** This module aims to retrieve relevant events based on the expanded questions. Relevant events are those directly related to the predictive question and can help provide a comprehensive background for the question. News are retrieved

from news sources based on both the expanded questions and original question. However, not all retrieved news articles are relevant. To filter out irrelevant news, we utilize the LLM to score each article and remove those with low scores.

**Similar Event Retrieval.** This module aims to retrieve similar events to reveal potential evolutionary patterns. Similar events refer to historically occurred events that are similar to the current question. Similar events can serve as references for predicting future event. For example, in Figure 1, by considering news 1, it can be inferred from news 2 whether China will sign a visa exemption agreement with Poland. However, similar events have often been overlooked in previous research.

Since the extracted stakeholders are mostly specific instances, such as *Australia* and *Li Qiang*, using these stakeholders primarily retrieves relevant events. Retrieving similar events, however, requires more abstract question formulations. To address this, we abstract the stakeholders and then use the abstracted role information to retrieve similar events. Specifically, we first use LLM to abstract the stakeholders, obtaining role information such as *country* and *government officials*. Then, based on the original question, background, and stakeholder roles, we generate diverse questions and use these questions to retrieve news about similar events from news sources. Each news article represents a similar event. Since a single news article may not provide comprehensive information, we use the LLM to generate multiple questions to further expand the information about the similar events.

### 3.3 Integration

The Integration module employs clustering method to clarify the dependencies between events and reduce redundant information. After obtaining the relevant and similar events, due to the large

**Table 1: Model performance of different types of questions on Chinese data (%). Detailed experimental results for each dimension can be found in the appendix B.**

Models	Methods	Time	Location	Development	Outcome	Impact	Response	Other	Overall
GPT-3.5	DR + Summ	32.65	47.83	41.22	45.39	46.18	37.65	38.86	37.52
	DR + Summ-o-Summ	38.77	46.21	43.06	46.81	47.68	37.37	<b>44.43</b>	41.12
	GQR + Summ-o-Summ	39.79	48.73	43.49	47.29	46.28	38.51	41.03	43.78
	StkFEP	<b>45.92</b>	<b>49.21</b>	<b>45.88</b>	<b>50.84</b>	<b>54.26</b>	<b>38.93</b>	40.53	<b>46.95</b>
GLM-4	DR + Summ	38.65	28.33	34.92	40.70	39.07	32.90	35.24	37.08
	DR + Summ-o-Summ	40.18	35.68	35.69	42.57	40.92	34.44	34.02	38.72
	GQR + Summ-o-Summ	42.50	38.17	34.39	39.98	41.72	38.37	31.53	40.05
	StkFEP	<b>45.25</b>	<b>44.43</b>	<b>48.23</b>	<b>51.57</b>	<b>54.20</b>	<b>40.14</b>	<b>39.57</b>	<b>46.27</b>
Llama3-8B	DR + Summ	28.24	38.01	33.38	38.66	39.95	36.52	<b>59.35</b>	32.84
	DR + Summ-o-Summ	31.01	35.67	35.47	37.61	<b>41.89</b>	39.60	46.08	34.64
	GQR + Summ-o-Summ	35.47	39.07	34.41	35.59	42.30	42.64	50.28	37.54
	StkFEP	<b>38.75</b>	<b>39.21</b>	<b>37.61</b>	<b>40.73</b>	41.87	<b>43.24</b>	53.68	<b>39.26</b>

scale of retrieved information, models often fail to fully utilize long-range contexts, and performance tends to decrease as context length increases. In addition, models do not rely on all retrieved information, which contains a considerable amount of redundancy. Therefore, before making predictions, it is necessary to clarify the dependencies between events and eliminate redundant information. Prior work often employs summarization methods [7, 22], which generate summaries for each document. However, this method struggles to effectively remove redundancy due to overlapping information among different news articles.

To address this, we propose a clustering method that organizes text segments into cohesive groups. Specifically, we first extract supportable content segments for prediction from news articles using LLMs and remove any duplicate segments. Next, we cluster all extracted segments. Following existing work [6], we use the K-means clustering algorithm. To determine the optimal number of clusters, we employ the Bayesian information criterion, which not only penalizes model complexity but also rewards goodness of fit [1]. After dividing the segments into different clusters, we prompt the LLM to generate a description for each cluster. Finally, these cluster descriptions are utilized to support predictions.

Both relevant and similar events undergo this processing procedure. However, since the outcomes are unknown for predictions, relevant events focus more on gathering comprehensive information, such as in Figure 1, the news 2 “*The President of Poland announced an upcoming visit to China*”. For similar events, where the outcomes are known, the focus is on the outcome and causes of the events, such as retrieved news 1 “*visits Thailand and signs a mutual exemption agreement*”.

### 3.4 Prediction

The Prediction module aims to predict outcomes based on the information gathered about relevant and similar events. LLMs employ step-by-step reasoning or self-reflection to enhance their ability to answer questions. However, LLMs often display overconfidence or high randomness [26, 28], frequently providing stubborn or inconsistent feedback [31], which can lead to potential bias and

miscalibration. To address this, we implement an aggregate strategy to obtain final prediction outcomes. This strategy involves collecting potential answers from various perspectives of relevant events and similar events, comparing differences between these answers to reach the final result. For time-related questions, the prediction model outputs the time interval with the highest probability. For the remaining questions, the model outputs free-form text.

## 4 Experiments

### 4.1 Implementation Details

**Dataset Details.** We use GPT-4 to assist in building the dataset. We annotate predictive questions daily, complete annotations the same day, and conduct testing immediately. After the prediction window, we construct ground truth and complete the evaluation. We utilize Bing API to retrieve the news.

**Framework Details.** We employ multiple advanced LLMs as the backbone, including GPT-3.5<sup>4</sup>, GLM-4 [5], Llama3-8B [25], and Mistral-7B [9]. Due to the limited Chinese data in training Mistral model, it cannot well support Chinese understanding. So Mistral is tested only on English data. We use embedding models, such as Sentence-BERT [20], to encode the text for clustering.

**Evaluation Details.** For automatic evaluation by LLM, we use GPT-4 to assess the performance of model predictions. To fully leverage the capabilities of the LLMs, we conduct tests for each dimension separately, such as *Accuracy*, *Completeness*, *Relevance*, *Specificity*, and *Reasonableness*.

### 4.2 Baselines

Due to the lack of existing LLM-based methods for open-ended future event prediction, we integrate widely used techniques from different stages to construct the baselines.

For Retrieval, we select two comparison methods: (1) *DR*, which uses the original predictive question to retrieve information directly; (2) *GQR*, where the LLMs automatically generates multiple

<sup>4</sup><https://chat.openai.com/chat>

**Table 2: Model performance of different types of questions on English data (%).**

Models	Methods	Time	Location	Development	Outcome	Impact	Response	Other	Overall
GPT-3.5	DR + Summ	35.18	29.98	32.93	46.24	50.51	35.96	42.93	37.41
	DR + Summ-o-Summ	38.12	37.44	29.21	49.84	53.55	38.74	48.34	39.85
	GQR + Summ-o-Summ	42.87	34.65	33.47	48.29	57.59	45.77	51.89	43.58
	StkFEP	<b>44.85</b>	<b>38.63</b>	<b>35.81</b>	<b>50.42</b>	<b>60.68</b>	<b>49.74</b>	<b>52.03</b>	<b>46.03</b>
GLM-4	DR + Summ	33.53	35.79	34.32	39.51	46.87	32.37	32.72	35.86
	DR + Summ-o-Summ	38.26	30.91	35.88	40.81	47.51	36.15	<b>50.40</b>	39.54
	GQR + Summ-o-Summ	42.88	45.47	33.68	40.59	51.24	40.24	36.52	42.62
	StkFEP	<b>43.31</b>	<b>48.31</b>	<b>36.40</b>	<b>41.39</b>	<b>54.70</b>	<b>40.63</b>	37.57	<b>43.11</b>
Llama3-8B	DR + Summ	26.34	22.92	28.24	48.73	42.80	34.11	25.29	31.34
	DR + Summ-o-Summ	29.15	23.17	25.77	46.82	44.90	40.50	31.96	32.51
	GQR + Summ-o-Summ	34.50	11.13	28.18	38.98	<b>55.04</b>	38.65	32.92	35.16
	StkFEP	<b>38.66</b>	<b>29.50</b>	<b>28.35</b>	<b>48.93</b>	50.93	<b>41.44</b>	<b>46.29</b>	<b>38.77</b>
Mistral-7B	DR + Summ	32.26	32.52	30.48	35.32	46.48	32.83	41.20	34.12
	DR + Summ-o-Summ	35.87	31.06	30.98	37.25	54.36	29.88	38.06	36.70
	GQR + Summ-o-Summ	39.12	<b>43.01</b>	31.56	36.72	<b>55.84</b>	32.03	36.81	38.94
	StkFEP	<b>41.38</b>	32.53	<b>31.79</b>	<b>43.07</b>	51.58	<b>37.93</b>	<b>57.84</b>	<b>41.24</b>

questions based on the question or background for retrieval, similar to existing work [3, 7].

For Integration, we select two comparison methods: (1) *Summ*, which generates a summary for each retrieved document, similar to existing work [7, 27]; (2) *Summ-over-Summ*, which first generates summaries for each document and then produces a brief description of all summaries.

Finally, for each backbone LLM, we employ three combination strategies as baselines, including *DR + Summ*, *DR + Summ-over-Summ*, and *GQR + Summ-over-Summ*. For the prediction module, all baselines utilize the same prediction framework.

### 4.3 Overall Results

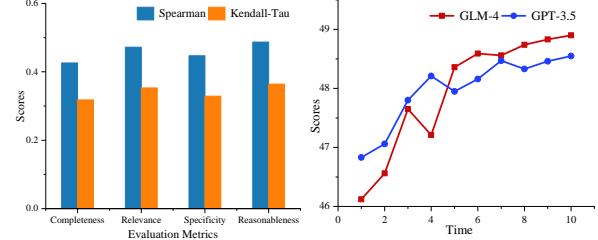
The comparative performances of various methods on Chinese and English data are detailed in Tables 1 and 2 respectively. Our approach StkFEP, which integrates stakeholders insights and information from similar events, consistently outperformed other methods. We also have four key observations:

(1) For time-related questions, the current best result is 44.85%. In our experiments, we set these questions as multiple-choice format and divided the prediction window into three intervals. Additionally, we tested the performance of GPT-3.5 over five intervals and found a decrease to 25.23%, indicating that these questions remain highly challenging.

(2) From the perspective of retrieval methods, using prediction questions directly for retrieval yields the poorest results, while employing LLMs to generate diverse questions shows improvement.

(3) In terms of information integration, the *Summ-o-Summ* approach, which uses summarization twice, performs better than a single summarization *Summ*, indicating that this method can further refine content.

(4) From the perspective of different languages, the model exhibits similar trends across all languages. The performance on questions related to *Development* is relatively lower.



**Figure 5: The correlations between human and LLMs.**

### 4.4 Human Evaluation

In this section, we expand our evaluation methodology beyond model-based metric. We conduct an additional human evaluation to compare 50 predictions generated by GPT-3.5. We invite annotators to assess the model outputs from four dimensions: *Completeness*, *Relevance*, *Specificity*, and *Reasonableness*, using the same criteria as the automatic evaluation method. We report the Spearman and Kendall-Tau correlations between human expert-annotated scores and GPT-4 assigned scores in Figure 5. We find that GPT-4 achieves a Spearman correlation of around 0.45, which indicates that recent LLMs perform predictions evaluations that are reasonably valid to a meaningful extent.

### 4.5 Analysis of Daily Prediction

We conduct daily predictions to capture the trends of predictions changing over time. To achieve this, we select 22 questions that will yield results after 10 days, organizing a test each day. The experimental results, as shown in Figure 6, indicate that the model performance generally improves over time with updates in information. Upon deeper analysis, we observe that during the initial days, the scale of information is substantial, encompassing both

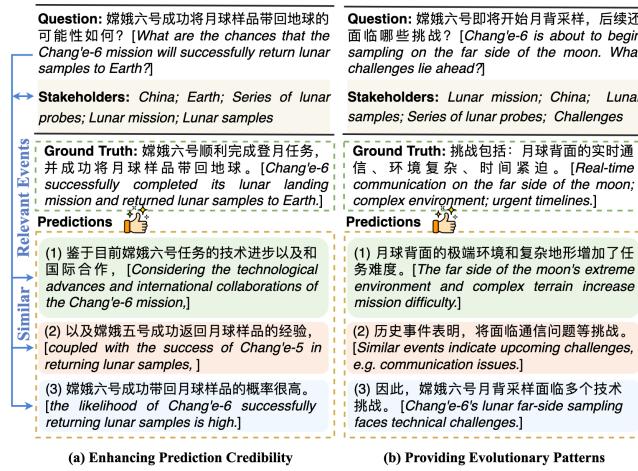


Figure 7: Cases for model predictions.

redundant and critical details, leading to significant fluctuations. As time progresses, public discussion about the issues diminishes, resulting in smaller fluctuations during this phase.

## 4.6 Case Study

To better understand the results shown in Table 3, we conduct a case study to explicitly illustrate the effectiveness of the event prediction framework. The cases are shown in Figure 7. For the first case, by identifying stakeholders such as *Lunar mission*, *Lunar samples*, and *China*, model can effectively retrieve similar events like “*Chang'e-5 successfully returning lunar samples*”. By then incorporating relevant events, it can significantly enhance the credibility of the predictions. For the second case, similar events can provide potential evolutionary patterns to support prediction. Retrieving similar events allows us to learn about challenges faced by previous lunar sampling missions, such as *communication issues*, and combining this with the progress and breakthroughs in current research, can enhance the effectiveness of event prediction.

## 4.7 Error Analysis

To enrich the understanding and better advance future research, we conduct a detailed analysis of the problems encountered in existing research. The common problems can primarily be divided into four categories: (1) **Incomplete Prediction** refers to scenarios where the predictions made are not comprehensive enough to cover all aspects or variables related to the event. As shown in case 1 of Figure 8, the model overlooks the outcome “*the train station temporarily halted passenger services*”. (2) **Underspecified Prediction** occurs when predictions are too vague or general, lacking specific details necessary for them to be actionable or useful. As shown in case 2, the model outputs “*Chang'e-6's successful return of lunar far-side samples has garnered widespread attention and positive reactions internationally*”. The predictions of model lacks value because it does not provide any salient entity information, resulting in an output too generic to effectively address the specific question. (3) **Irrelevant Prediction** describes predictions unrelated to the question

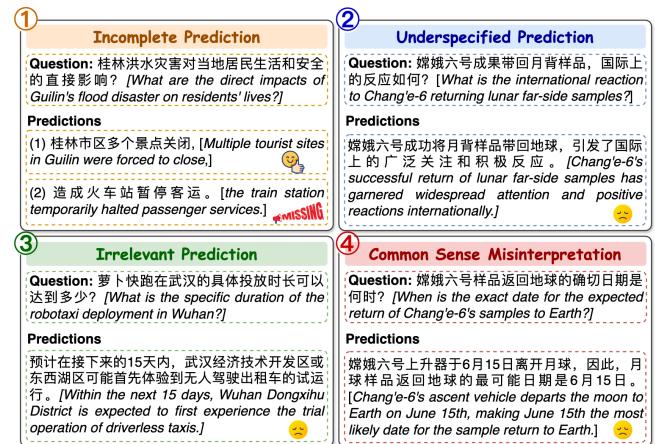


Figure 8: Error analysis of the model predictions.

posed, essentially providing answers that do not address the question. As shown in case 3, the question asks about time information, but the model responds with a location information “*Wuhan Dongxihu District*”. (4) **Common Sense Misinterpretation** arises when predictions contradict basic common sense, resulting in outcomes that are implausible or logically inconsistent with known facts. This undermines the credibility of the predictions and may lead to mistrust or disregard of model outputs. In case 4, the statement “*Chang'e-6's ascent vehicle departs the moon for Earth on June 15th*” is predicted, however, the model overlooks the common sense that it is impossible to return from the moon to Earth within a day.

Table 3: Ablation study.

Methods	GPT-3.5	GLM-4
StkFEP	46.95	46.27
w/o Cluster-Summ	46.11	45.38
w/o Similar Events	45.65	44.79
w/o Stakeholders	44.28	42.80

## 4.8 Ablation Study

To more specifically validate the different modules within the event prediction framework, we conduct experiments to ablate the clustering-over-summarization method (*w/o Cluster-Summ*) for information integration, similar events (*w/o Similar Events*), and stakeholders (*w/o Stakeholders*). From the results in Table 3, we can see that: (1) For the scenario without cluster summarization (*w/o Cluster-Summ*), where we used Summ-over-Summ for information integration, the model performance decreased, indicating that our method can more effectively refine information and organize dependencies between events. (2) For the scenario without similar events (*w/o Similar Events*), relying only on relevant events for predictions, the model results also declined, mainly because similar events provide potential evolutionary patterns that support the final predictions. (3) For the scenario without stakeholders (*w/o*

*Stakeholders*), ignoring stakeholders resulted in the most substantial drop in model performance. This demonstrates that utilizing stakeholders not only enhances the diversity of retrieval but also enables more accurate retrieval of similar events.

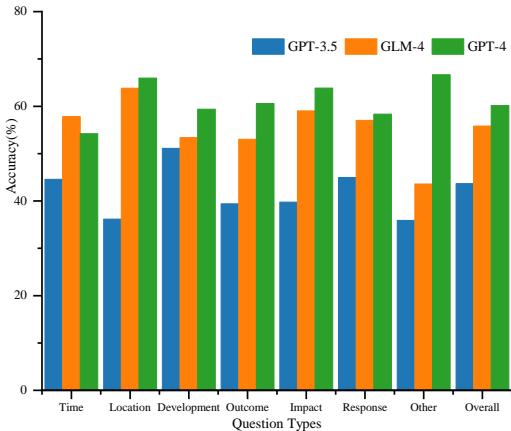
## 5 Conclusions

In this paper, we introduce OpenEP (an open-ended future event prediction task), which generates flexible and diverse predictions aligned with real-world scenarios. To facilitate the study of this task, we first construct OpenEPBench, an open-ended future event prediction dataset. For question construction, we pose questions from seven perspectives, including location, time, event development, event outcome, event impact, event response, and other, to facilitate an in-depth analysis and understanding of the comprehensive evolution of events. For outcome construction, we collect free-form text containing the outcomes as ground truth to provide semantically complete and detail-enriched outcomes. Furthermore, we propose StkFEP, a stakeholder-enhanced future event prediction framework that incorporates the characteristics of event evolution for open-ended settings. Our method extracts stakeholders involved in events to extend questions and collects historical events that are relevant and similar to the question to gather diverse and comprehensive information to support model prediction. Extensive experiments on Chinese and English data demonstrate that accurately predicting future events in open-ended settings is challenging for existing large language models.

## References

- [1] Samuel Adeyemo and Debangsu Bhattacharyya. 2024. Optimal nonlinear dynamic sparse model selection and Bayesian parameter estimation for nonlinear systems. *Computers & Chemical Engineering* 180 (2024), 108502.
- [2] Long Bai, Saiping Guan, Zixuan Li, Jiafeng Guo, Xiaolong Jin, and Xueqi Cheng. 2023. Rich event modeling for script event prediction. In *Proceedings of the AAAI*. Article 1409, 9 pages.
- [3] Junyan Cheng and Peter Chin. 2024. SocioDojo: Building Lifelong Analytical Agents with Real-world Text and Time Series. In *Proceedings of the ICLR*.
- [4] Walter H. Dempsey, Alexander Moreno, Christy K. Scott, Michael L. Dennis, David H. Gustafson, Susan A. Murphy, and James M. Rehg. 2017. iSurvive: An Interpretable, Event-time Prediction Model for mHealth. In *Proceedings of the ICML (Proceedings of Machine Learning Research, Vol. 70)*. 970–979.
- [5] Team GLM, , Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Ha Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Luceen Zhong, Mingdiao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793 [cs.CL]
- [6] Yong Guan, Xiaozhi Wang, Lei Hou, Juanzi Li, Jeff Z. Pan, Jiaoyan Chen, and Freddy Lecue. 2024. TacoERE: Cluster-aware Compensation for Event Relation Extraction. In *Proceedings of the LREC-COLING*. 15511–15521.
- [7] Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching Human-Level Forecasting with Language Models. arXiv:2402.18563 [cs.LG]
- [8] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features. In *Proceedings of the ACL*. 987–997.
- [9] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL]
- [10] Alapan Kuila and Sudeshna Sarkar. 2024. From Text to Context: An Entailment Approach for News Stakeholder Classification. In *Proceedings of the SIGIR*. 2426–2430.
- [11] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. 591–600.
- [12] Srivatsan Laxman, Vikram Tankasali, and Ryan W. White. 2008. Stream prediction using a generative model based on frequent episodes in event sequences. In *Proceedings of the SIGKDD*. 453–461.
- [13] Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021. The Future is not One-dimensional: Complex Event Schema Induction by Graph Modeling for Event Prediction. In *Proceedings of the EMNLP*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). 5203–5215.
- [14] Xiaoyan Liu, Jiarui Zhao, Ran Liu, and Kai Liu. 2022. Event history analysis of the duration of online public opinions regarding major health emergencies. *Frontiers in Psychology* 13 (2022). 954559.
- [15] Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, and Tat-Seng Chua. 2023. Context-aware Event Forecasting via Graph Disentanglement. In *Proceedings of the SIGKDD*. 1643–1652.
- [16] Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, and Tat-Seng Chua. 2023. Context-aware Event Forecasting via Graph Disentanglement. In *Proceedings of the SIGKDD*. 1643–1652.
- [17] Yaswanth Narusupalli, Abhranil Chandra, Sreevatsa Muppirlala, Manish Gupta, and Pawan Goyal. 2024. Review-Feedback-Reason (ReFeR): A Novel Framework for NLG Evaluation and Reasoning. arXiv:2407.12877 [cs.CL]
- [18] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 <https://arxiv.org/pdf/2303.08774.pdf>
- [19] Sarah Pratt, Seth Blumberg, Pietro Kreitlon Carolino, and Meredith Ringel Morris. 2024. Can Language Models Use Forecasting Strategies? arXiv:2406.04446 [cs.LG]
- [20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the EMNLP-JCNLP*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). 3982–3992.
- [21] Shakila Khan Rumi, Ke Deng, and Flora D. Salim. 2018. Theft prediction with individual risk factor of visitors. In *Proceedings of the SIGSPATIAL*. 552–555.
- [22] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. arXiv:2401.18059 [cs.CL]
- [23] Philipp Schoenegger, Indre Tuminauskaitė, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock. 2024. Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy. arXiv:2402.19379 [cs.CY]
- [24] Smriti Sharma, Rajesh Kumar, Pawan Bhadana, and Sumita Gupta. 2013. News event extraction using 5W1H approach & its analysis. *International Journal of Scientific & Engineering Research* 4, 5 (2013), 2064–2068.
- [25] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhowmik, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Eisobou, Jude Fernandes, Jeremy Fu, Wenjin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madiam Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Miaylov, Pushkar Mishra, Igor Molbyog, Yixin Nie, Andrew Poultney, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]
- [26] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *Proceedings of the ICLR*.
- [27] Qi Yan, Raihan Seraj, Jiawei He, Lili Meng, and Tristan Sylvain. 2024. AutoCast++: Enhancing World Event Prediction with Zero-shot Ranking-based Context Retrieval. arXiv:2310.01880 [cs.LG]
- [28] Haoyan Yang, Xiyuan Wang, Xingyu Xu, Hanyuan Zhang, and Yirong Bian. 2024. Can We Trust LLMs? Mitigate Overconfidence Bias in LLMs through Knowledge Transfer. arXiv:2405.16856 [cs.CL]
- [29] Yiyang Yang, Zhongyu Wei, Qin Chen, and Libo Wu. 2019. Using External Knowledge for Financial Event Prediction Based on Graph Neural Networks. In *Proceedings of the CIKM*. 2161–2164.
- [30] Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. 2024. MIRAI: Evaluating LLM Agents for Event Forecasting. arXiv:2407.01231 [cs.CL]
- [31] Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueteng Zhuang, and Weiming Lu. 2024. Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives. arXiv:2401.02009 [cs.CL]

- [32] Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. 2024. Large Language Models as Evaluators for Recommendation Explanations. arXiv:2406.03248 [cs.IR]
- [33] Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–37.
- [34] 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 *Proceedings of the AAAI*. Article 1576, 9 pages.
- [35] 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 *Proceedings of the NeurIPS*, Vol. 35. 27293–27305.



**Figure 9: Performance of data validity verification on different LLMs.**

## A Data Validity Verification

We adopt a data construction strategy that involves annotating questions on the same day they are tested, with answers collected after the prediction window for evaluation. This means that at the time of question annotation, it is unknown whether there will be outcomes, inevitably leading to instances where no answers are available. Such instances are termed as invalid questions, and are excluded from the dataset. Notably, we collect 983 valid questions and annotate 286 invalid ones. Although these questions lack answers, they are valuable for evaluating the model's capacity to identify question validity. Therefore, based on these 286 invalid questions, we extract an equivalent number of valid questions to test whether three powerful LLMs, such as GPT-4, GLM-4, and GPT-3.5, can identify which questions would have answers and which would not during the prediction window.

The results on different types of questions are shown in Figure 9. The results indicate that event identify the data validity is challenging for existing LLMs. We also have the following three observations: (1) For model perspectives, GPT-4 achieves the best performance, except in time-related questions where GLM-4 slightly outperforms. (2) Regarding question types, time-related questions exhibit the lowest accuracy, and GLM-4 achieves the highest accuracy in this category with 57.83%. (3) In terms of overall scores, GPT-4 leads in performance. However, GPT-3.5 achieves an accuracy score of 43.67%.

## B Performance on Individual Dimension

This section aims to provide a detailed analysis of the specific scores across different question types for each evaluation dimension. Figures 10 and 11 display the experimental results for GPT-3.5, while Figures 12 and 13 present the results for GLM-4. Overall, the experimental outcomes exhibit similar trends across both LLMs. Completeness scores are the lowest, indicating that making comprehensive future predictions is highly challenging. Reasonableness scores are relatively higher, suggesting that the predictions generated by the large models are logically consistent.

## C Details of Dataset Construction

We design the following six principles to better assist LLMs and human annotators in constructing the data.

(1) **Real-time Principle.** Events must be currently occurring. Data related to events not happening in real-time should be discarded, with potential scenarios including: (a) An event that occurred years ago has become a hot topic, such as “*A female employee was fired for using an umbrella at work to avoid exposure, and the court ruled the company's termination legal*”. (b) An event that happened some time ago and has been a hot topic for a while, such as “*How should one evaluate the role of a full-time postdoctoral fellow at Sichuan University*”.

(2) **Answerability Principle.** For a predictive question, it must be ascertainable and answerable to warrant annotation. Unanswerable questions should be discarded. For example, “*How might the performance of the Chinese team in the next 15 days affect its status in international football?*” On one hand, a few matches alone are insufficient to determine impacts on international status. On the other, the outcomes of such questions may not become apparent within 15 days and should therefore be discarded.

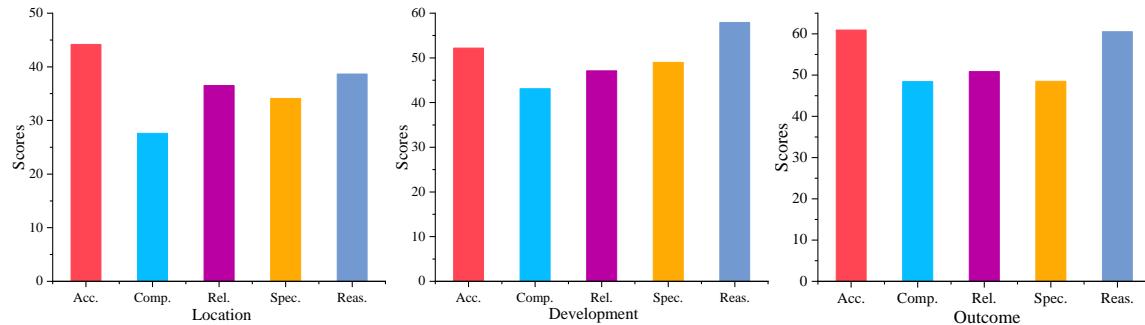
(3) **Specificity Principle.** Vague questions and those with broad, indeterminate answers should be discarded. For example, “*What impact might the STSS epidemic have on Japan in the next 15 days?*” This question is unclear as the impacts could span multiple aspects, including economic and political, and should therefore be discarded.

(4) **Continuity Principle.** An event must still be unfolding and not concluded to justify its annotation. Events that have already ended should be discarded.

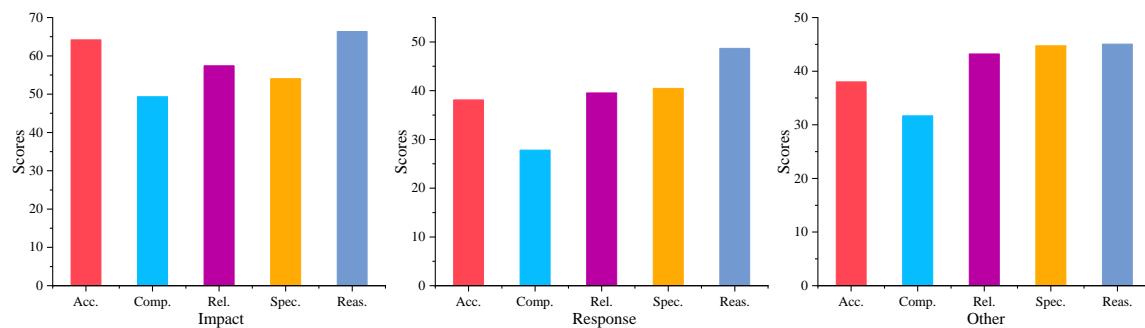
(5) **Short-Term Principle.** The current task of future event prediction primarily predicts events that may occur within the next 15 days. Therefore, it is necessary to analyze whether a predictive question can yield results within this prediction window. For example, “*What new laws might be proposed in response to construction safety incidents?*” Legislative proposals typically do not yield results within 15 days and should be discarded.

(6) **Truthfulness Principle.** Events that are annotated must be real and currently occurring. People may pose discussions about events that have not actually happened. For example, “*Is there a future in opening all-female nursing homes for older single women?*” or “*Do you remember what you did on the night after the college entrance exam ended?*”.

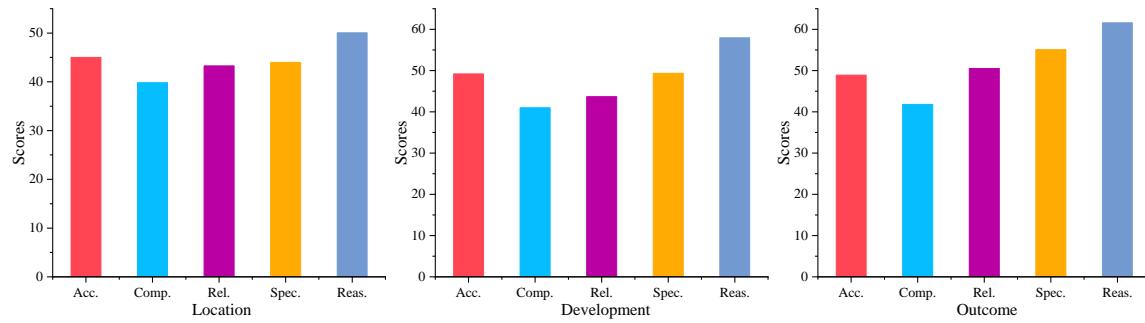
Furthermore, we have developed an annotation system. The question annotation interface is depicted in Figure 14, and the ground truth annotation interface is shown in Figure 15.



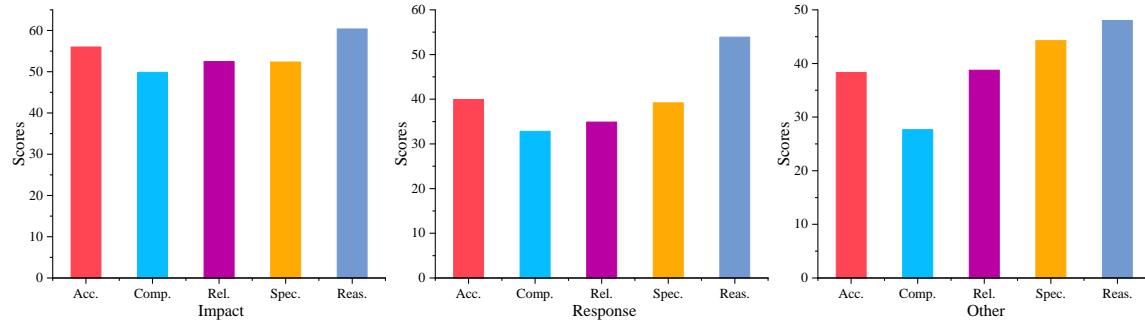
**Figure 10: GPT-3.5 Performance of individual dimension on Location, Event Development, and Event Outcome. Acc., Comp., Rel., Spec., and Reas. are abbreviations for Accuracy, Completeness, Relevance, Specificity, and Reasonableness, respectively.**



**Figure 11: GPT-3.5 Performance of individual dimension on Event Impact, Event Response, and Other.**



**Figure 12: GLM-4 Performance of individual dimension on Location, Event Development, and Event Outcome.**



**Figure 13: GLM-4 Performance of individual dimension on Event Impact, Event Response, and Other.**

## Candidate Question Annotation

[Home Page](#) [Annotation Instructions](#) [Annotation Progress](#)

Hot topic: Beryl strengthens into the earliest Category 5 Atlantic hurricane on record after devastating Windward Islands

Candidate Questions	Background of Hot Topic	News Content
<b>Time</b>  <input type="checkbox"/> During which period is Hurricane Beryl expected to reach its maximum intensity as it approaches the central Caribbean? <input type="checkbox"/> In what timeframe will Hurricane Beryl likely weaken from a Category 5 to a lower category as it progresses?	Hot Topic: Beryl strengthens into the earliest Category 5 Atlantic hurricane on record after devastating Windward Islands  Date Query: 2024-07-01  URL: <a href="https://edition.cnn.com/2024/07/01/weather/hurricane-beryl-caribbean-landfall-monday/index.html">https://edition.cnn.com/2024/07/01/weather/hurricane-beryl-caribbean-landfall-monday/index.html</a>	DatePublished: 2024-07-01  URL: <a href="https://weather.com/storms/hurricane/news/2024-07-01-hurricane-beryl-category-5">https://weather.com/storms/hurricane/news/2024-07-01-hurricane-beryl-category-5</a>  Sign up for the Morning Brief email newsletter to get weekday updates from The Weather Channel and our meteorologists.
<b>Location</b>  <input type="checkbox"/> Which regions in the central Caribbean are forecasted to be in the direct path of Hurricane Beryl as it maintains its intensity?	Original Background: Hurricane Beryl has strengthened into a Category 5 Atlantic hurricane — the earliest on record — as it powers across the Caribbean after bringing devastation to the Windward Islands, where at least one person is dead. Its intensity also marks just the second time an Atlantic hurricane has reached Category 5 status in July after Emily did so on July 17, 2005, according to the National Hurricane Center. Beryl's maximum sustained winds have increased to near 160 mph, with higher gusts, the NHC said.	Hurricane Beryl has become the first Category 5 hurricane of the season and the earliest hurricane of such strength on record. Beryl is the earliest to reach Category 5 strength on record in the Atlantic, and reached the record earlier than the previous storm by more than two weeks. The previous earliest Category 5 was Hurricane Emily on July 16 during the hyperactive 2005 Atlantic Hurricane Season.
<b>Event Development</b>  <input type="checkbox"/> How will Hurricane Beryl's path and intensity change as it approaches Jamaica? <input type="checkbox"/> What are the expected changes in Hurricane Beryl's intensity as it interacts with the landmass of the central Caribbean?	Generated Background: Hurricane Beryl has escalated into a Category 5 storm, marking it as the earliest	Beryl's rapid intensification and extreme intensity are likely due to record warm water temperatures.  Earlier today, Beryl's center moved over Carriacou Island in the Grenadines with maximum sustained winds of 150 mph, making Beryl a strong Category 4.
<b>Event Outcome</b> Please manually enter the appropriate predictive question:	<small>See our full annotation tool link for more on where Beryl is headed next</small>	
Question Type: (time: 0, location: 1, event development: 2, event outcome: 3, event impact: 4, event response: 5, other: 6)		
<input type="button" value="Submit"/>		

Figure 14: Question annotation interface.

## Ground Truth Annotation

[Home Page](#) [Annotation Instructions](#) [Annotation Progress](#)

Prediction Question: During which period is Hurricane Beryl expected to reach its maximum intensity as it approaches the central Caribbean?

Annotation List	Question Details	News Content
Question: During which period is Hurricane Beryl expected to reach its maximum intensity as it approaches the central Caribbean?  Please manually enter the revised question:	Hot Topic: Beryl strengthens into the earliest Category 5 Atlantic hurricane on record after devastating Windward Islands  Date Published: 2024-07-01	News Title: Beryl Becomes the First Major 2024 Atlantic Hurricane  URL: <a href="https://gpm.nasa.gov/applications/weather/news/beryl-becomes-first-major-2024-atlantic-hurricane">https://gpm.nasa.gov/applications/weather/news/beryl-becomes-first-major-2024-atlantic-hurricane</a>  Published Date: 2024-07-09  Score: 5
Please manually enter the answer URL: <small>Separate each url with a semicolon (;)</small>	<b>News List</b>  <input type="checkbox"/> Beryl Becomes the First Major 2024 Atlantic Hurricane <input type="checkbox"/> Beryl's 'alarming' characteristics: A deep dive into its rapid... <input type="checkbox"/> July 4, 2024: Hurricane Beryl brings powerful winds and... - CNN <input type="checkbox"/> Hurricane Beryl drops to Category 4 strength after breaking records... <input type="checkbox"/> Hurricane Beryl weakens as Caribbean islands assess damage : NPR <input type="checkbox"/> Category 5 Hurricane Beryl makes explosive start to 2024 Atlantic ... <input type="checkbox"/> Beryl's records: Impacts, track, what's known and unknown <input type="checkbox"/> Watch: Satellite video tracks Beryl's path tearing	On the morning of Sunday, June 30, Hurricane Beryl became a rare early season major hurricane when it reached Category 3 status with sustained winds of 115 mph as it was moving across the Atlantic 420 miles east-southeast of Barbados in the direction of the Leeward Islands. Then, just a few hours later at 11:35 a.m. EDT, Beryl became the earliest Category 4 hurricane on record in the Atlantic with sustained winds reported at 130 mph by the National Hurricane Center (NHC), surpassing Hurricane Dennis from the epic 2005 Atlantic hurricane season. Beryl also became historic as the strongest and easternmost hurricane and major hurricane in June in the Atlantic. Beryl originated from a tropical wave of low pressure that moved off the coast of Africa on June 22. Known as African Easterly Waves (AEWs), these waves typically move westward from Africa across the tropical Atlantic and into the Caribbean at semi-regular intervals and can serve as seedlings for tropical storms and hurricanes. However, what makes Beryl unusual is that it formed from an AEW so early in the season.
Please manually enter the answer:		
<input type="checkbox"/> There is no answer to this question <input type="checkbox"/> The answer to this question has not yet emerged		
Please manually enter the new questions:  Question Type: (time: 0, location: 1, event development: 2, event outcome: 3, event impact: 4, event response: 5, other: 6)		
<input type="button" value="Submit"/>		

Figure 15: Ground truth annotation interface.