# 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.

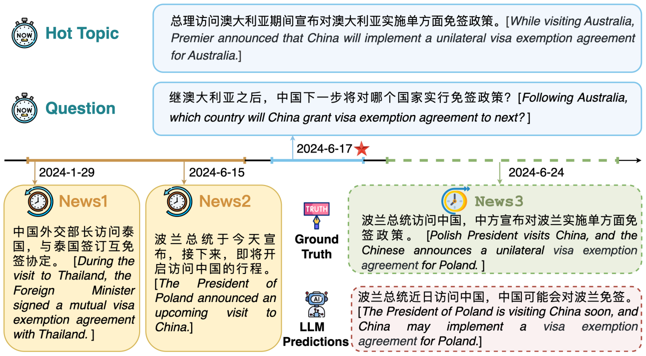
## CCS Concepts

- **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

*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnn.nnn



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.*]

继澳大利亚之后，中国下一步将对哪个国家实行免签政策？[*Following Australia, which country will China grant visa exemption agreement to next?*]

中国外交部长访问泰国，与泰国签订互免签协定。 [*During the visit to Thailand, the Foreign Minister signed a mutual visa exemption agreement with Thailand.* ]

波兰总统于今天宣布，接下来，即将开启访问中国的行程。[*The President of Poland announced an upcoming visit to China.*]

波兰总统访问中国，中方宣布对波兰实施单方面免签政策。 [*Polish President visits China, and the Chinese announces a unilateral visa exemption agreement for Poland.* ]

波兰总统近日访问中国，中国可能会对波兰免签。[*The President of Poland is visiting China soon, and China may implement a visa exemption agreement for Poland.*]

**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

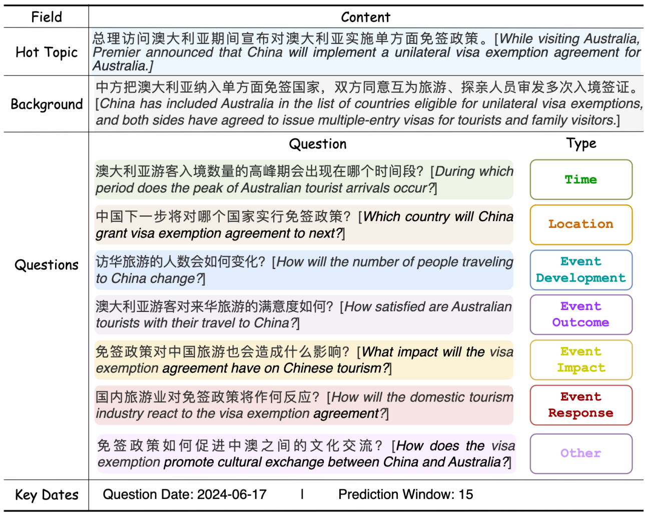
In this section, we will describe the OpenEPBench dataset. First, we introduce the overall dataset construction (Sec 2.1). Next, we introduce the construction process, which includes the data source (Sec. 2.2), constructing the predictive questions (Sec. 2.3), and their corresponding outcomes (Sec. 2.4). 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 employs 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 | Question | Type |
| | 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?*] | Time |
| | 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?*] | Location |
| | 访华旅游的人数会如何变化？[*How will the number of people traveling to China change?*] | Event Development |
| | 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?*] | Event Outcome |
| | 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?*] | Event Impact |
| | 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?*] | Event Response |
| | 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?*] | Other |
| 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 [1] for Chinese data, where daily discussions on hot topics are directly utilized as hot topics, and Google News [2] 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.
- 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.

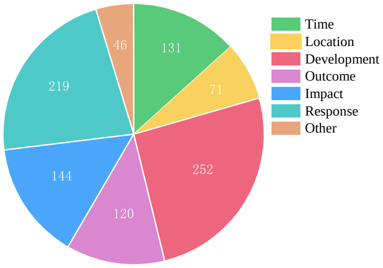


**Figure 3: Data distribution across different types of questions.**

**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],

[1] https://www.zhihu.com/knowledge-plan/hot-question/hot
[2] https://news.google.com/topics