# 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.

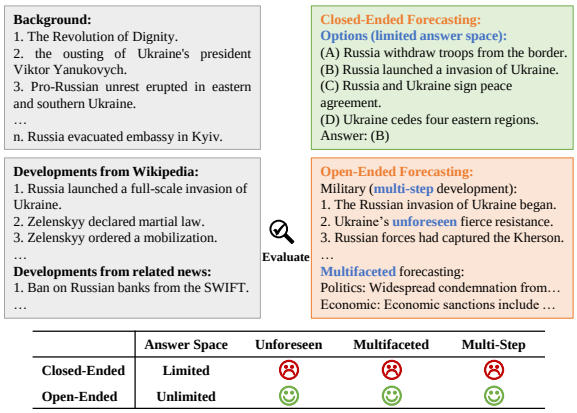


Figure 1: Comparison between open-ended and closed-ended event forecasting for complex events. Closed-ended forecasting is constrained to a limited answer space, while open-ended forecasting facilitates unforeseen, multifaceted, and multi-step predictions.

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.