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:

$$score = \sum_{i=1}^{n} \sigma(s_i) * \rho(s_i) \tag{1}$$

where $\sigma(\cdot)$ aims to map the score to the range 0-1, $\rho(\cdot)$ 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[3], 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.

[3]Each timestamp represents a day, formatted in "YYYY-MM-DD".

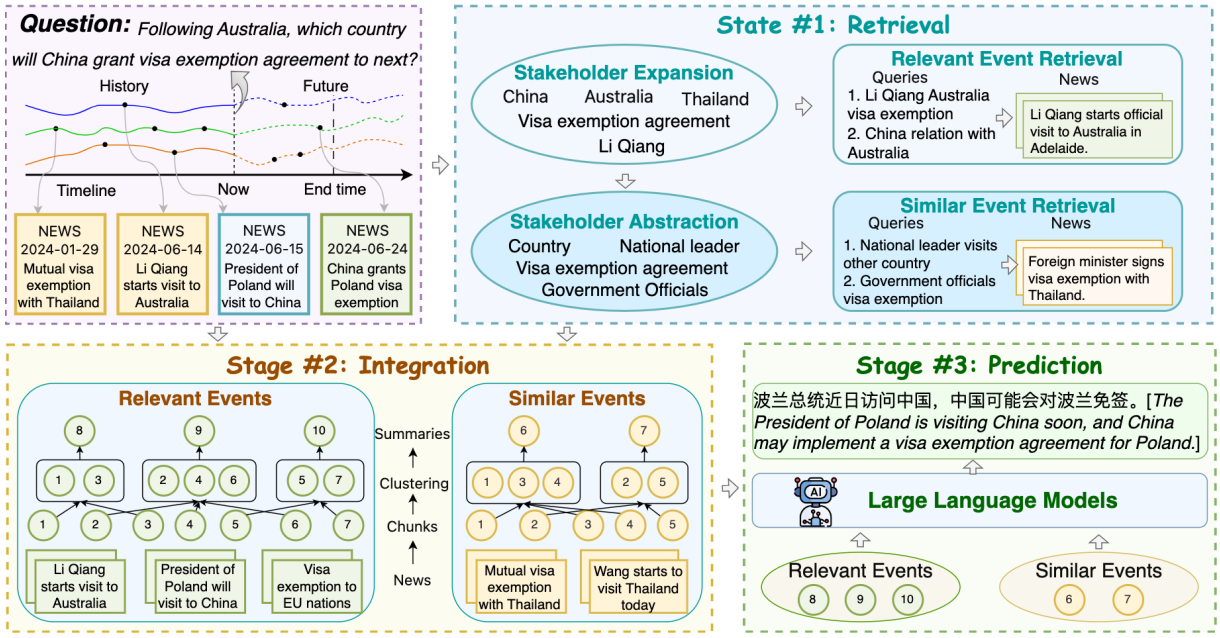

Figure 4: The framework of StkFEP.

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	**44.43**	41.12
	GQR + Summ-o-Summ	39.79	48.73	43.49	47.29	46.28	38.51	41.03	43.78
	StkFEP	**45.92**	**49.21**	**45.88**	**50.84**	**54.26**	**38.93**	40.53	**46.95**
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	**45.25**	**44.43**	**48.23**	**51.57**	**54.20**	**40.14**	**39.57**	**46.27**
Llama3-8B	DR + Summ	28.24	38.01	33.38	38.66	39.95	36.52	**59.35**	32.84
	DR + Summ-o-Summ	31.01	35.67	35.47	37.61	**41.89**	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	**38.75**	**39.21**	**37.61**	**40.73**	41.87	**43.24**	53.68	**39.26**

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[4], 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

[4]https://chat.openai.com/chat