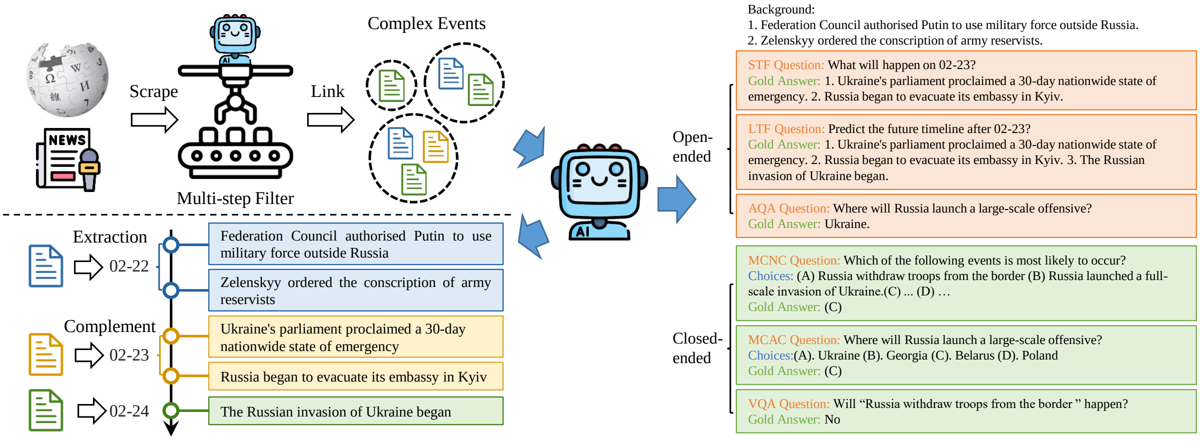

Federation Council authorised Putin to use military force outside Russia

Zelenskyy ordered the conscription of army reservists

Ukraine's parliament proclaimed a 30-day nationwide state of emergency

Russia began to evacuate its embassy in Kyiv

The Russian invasion of Ukraine began

Background:
1. Federation Council authorised Putin to use military force outside Russia.
2. Zelenskyy ordered the conscription of army reservists.

STF Question: What will happen on 02-23?
Gold Answer: 1. Ukraine's parliament proclaimed a 30-day nationwide state of emergency. 2. Russia began to evacuate its embassy in Kyiv.

LTF Question: Predict the future timeline after 02-23?
Gold Answer: 1. Ukraine's parliament proclaimed a 30-day nationwide state of emergency. 2. Russia began to evacuate its embassy in Kyiv. 3. The Russian invasion of Ukraine began.

AQA Question: Where will Russia launch a large-scale offensive?
Gold Answer: Ukraine.

MCNC Question: Which of the following events is most likely to occur?
Choices: (A) Russia withdraw troops from the border (B) Russia launched a full-scale invasion of Ukraine.(C) ... (D) …
Gold Answer: (C)

MCAC Question: Where will Russia launch a large-scale offensive?
Choices:(A). Ukraine (B). Georgia (C). Belarus (D). Poland
Gold Answer: (C)

VQA Question: Will “Russia withdraw troops from the border ” happen?
Gold Answer: No

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

4.2 Event Timeline Annotation

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

4.3 Question Annotation

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

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

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

4.4 Dataset Statistics

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

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

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

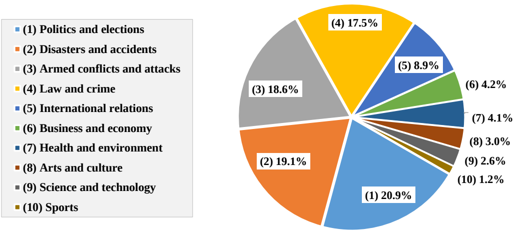

Figure 3: The category distributions of OpenForecast.

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

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

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

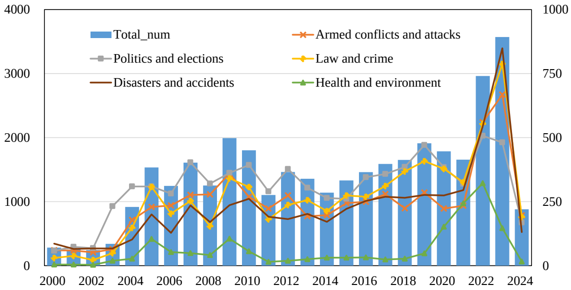

Figure 4: The temporal evolution of total event counts and five distinct event categories from 2000 to 2024.

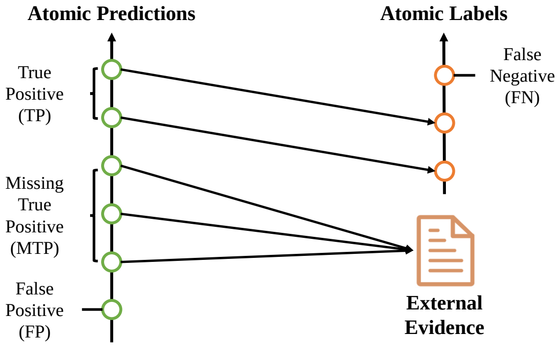

Figure 5: Illustration for the LLM-based retrieval-augmented evaluation.

5 LLM-based Retrieval-Augmented Evaluation

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

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

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

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

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

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

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

6 Experiments

6.1 Experimental Setup

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

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

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

6.2 Main Results

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

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