omit any atomic events. If two or more atomic events are omitted, the event timeline is annotated as incomplete.

- *Temporal Correctness* evaluates whether the extracted event timelines follow the correct chronological order. Annotators need to check the temporal relationships throughout the extracted event timelines according to the original articles. If any atomic events are in the wrong order, the event timeline in this dimension is annotated as incorrect.

- *Factual Consistency* examines whether the extracted atomic events contain factual errors. Annotators need to verify the atomic events using the original articles, ensuring factual consistency in event type, subject, object, time, location, etc. If two or more atomic events are incorrect, the entire event timeline is annotated as incorrect.

In these dimensions, the event relevance dimension corresponds to the dataset collection, while the other dimensions correspond to the event timeline annotation.

The annotation consistency ratios between the two annotators in *Event Relevance*, *Completeness*, *Temporal Correctness*, and *Factual Consistency* are 98.5%, 93.0%, 96.0%, and 98.5% respectively, indicating substantial agreement. OpenForecast achieves an *Event Relevance* of 98.0%, *Completeness* of 90.5%, *Temporal Correctness* of 94.5%, and *Factual Consistency* of 97.5%, demonstrating the effectiveness of our event timeline construction pipeline. In the *Completeness* dimension, our pipeline provides comprehensive event development, with minimal omissions of atomic events in the background and aftermath.

A.4 The Quality of the Question Annotation

In this work, we propose three open-ended tasks and three closed-ended tasks. The gold answers in open-ended tasks come from the extracted event timeline and are already proved in the event timeline construction evaluation (see Appendix A.3). For the closed-ended tasks, except the gold answers, noisy candidates and argument-level questions in MCAC are generated using LLMs and need further evaluation. To evaluate their quality in closed-ended tasks, we randomly select 200 questions from MCNC and MCAC and ask two human annotators[7] to evaluate them from multiple dimensions, as outlined below:

- *Clearness* examines whether the samples in the MCNC and MCAC tasks explicitly clarify the event background and questions. For the MCAC task, annotators also need to check the quality of the argument-level questions.

- *Answerable* examines whether the questions and candidates are answerable. Specifically, given a question, annotators should be able to identify the correct answer from four candidates by analyzing the event background.

- *Uniqueness* evaluate the correctness of three noisy candidates. All noisy candidates should be different from the gold answer in wording and semantics, ensuring that only one correct answer is provided.

The annotation consistency ratios between the two annotators in *Clearness*, *Answerable*, and *Uniqueness* are 97.0%, 93.5%, and 98.5% respectively, indicating substantial agreement. The closed-ended questions in OpenForecast achieve a *Clearness* of 97.5%, *Answerable* of 91.0%, and *Uniqueness* of 98.5%, demonstrating the effectiveness of the generated questions.

B Experiments

B.1 LLMs in Experiments

The large language models in our experiments are listed below:

- *Llama2-series* contain multiple instruction-finetuned model in different scale: Llama2-7b-chat[8], Llama2-13b-chat[9], Llama2-70b-chat[10]. The pretraining data has a cutoff of September 2022, ensuring no knowledge leakage.

- *WizardLM-13B-V1.2*[11] empowers LLMs to follow complex instructions by creating large amounts of instruction data with varying levels of

[7] Our annotation team consists of two graduate students engaged in information processing. Another annotator will review their annotation results and eliminate their discrepancies.

[8] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[9] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[10] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

[11] https://huggingface.co/WizardLMTeam/WizardLM-13B-V1.2

complexity and finetuning on Llama2-13b. The pretraining data of Llama2-13b has a cutoff of September 2022, ensuring no knowledge leakage.

- *Vicuna-13b-v1.5*[12] finetunes Llama2-13b on user-shared conversations collected from ShareGPT, ensuring no knowledge leakage.

- *Falcon-40b-instruct*[13] is a chat model based on Falcon-40b with 40B parameters. Since it was released in May 2023, there is no knowledge leakage issue.

- *Mixtral-8x7B-Instruct-v0.1*[14] is a pretrained generative Sparse Mixture of Experts and is released on December 11, 2023. Considering that Llama3-70b is released on April 18, 2024 and its pretraining data has a cutoff of December 2023 (5 months ago), there is no knowledge leakage issue for the evaluation on the testset with a cutoff of September 2023 (3 months ago).

- *Llama3-8b-Instruct*[15] is pretrained on over 15 trillion tokens of data from publicly available sources. Although Llama3-8b was released on April 18, 2024, the pretraining data has a cutoff of March 2023, ensuring no knowledge leakage.

B.2 Human Evaluation for Open-ended Event Forecasting

The human evaluation reflects the true performance of open-ended event forecasting. To evaluate the LLMs on three open-ended tasks, we randomly select 100 samples for each open-ended task and collect the corresponding forecasting results from nine LLMs. After the atomic event partitioning, the number of atomic events reaches approximately 9,000 for both STF and LTF, rendering the human evaluation labor-intensive.

Then, We ask two groups of annotators[16] to judge whether the prediction is correct at the atomic level. Annotators are provided with detailed information, including questions, predictions, gold labels, backgrounds, and original articles. For each sample, annotators must investigate the background, development, aftermath, and reactions of the event via Wikipedia and web searching. Then annotators need to find supporting evidence and determine whether the atomic predictions really occurred. The annotation should follow the following criteria:

- If a prediction can be directly confirmed by specific gold answers, it is annotated as correct, and the corresponding gold answer should be documented.

- If a prediction cannot be confirmed by the gold answers but can be confirmed through external sources such as Wikipedia or reliable web sources, it is also annotated as correct.

- If a prediction cannot be directly confirmed by gold answers or external sources, annotators should judge whether the prediction is reasonable. If the prediction does not contradict existing facts and can be inferred as correct, it is annotated as correct.

- If no evidence supports the prediction and it contradicts existing facts, it will be annotated as incorrect.

After annotation, we calculate the F1-score using the formula (2) in Appendix B.3 for each sample. The average F1-score difference between the two annotation teams are 0.02, indicating substantial agreement. To alleviate the disagreement between the annotation groups, we filter out those samples with the absolute F1-score difference exceeding 0.1. Then, we employ another annotator to review the annotation results and eliminate their discrepancies.

B.3 Open-Ended Evaluation Metric for List-Style tasks

This section introduces the evaluation metric for list-style tasks: LTF and STF, which means each gold answer and prediction contains multiple atomic events. Traditional *precision*, *recall*, and *F1-score* are defined as formula 1. Here, *TP* denotes true predictions that can be verified by given labels, *FP* denotes false predictions that aren't included in given labels, and *FN* denotes missing

[12] https://huggingface.co/lmsys/vicuna-13b-v1.5

[13] https://huggingface.co/tiiuae/falcon-40b-instruct

[14] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[15] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[16] Our annotation team consists of four graduate students engaged in information processing.

labels that are not recalled.

$$
\begin{aligned}
Precision &= \frac{TP}{TP+FP} \\
Recall &= \frac{TP}{TP+FN} \\
F1 &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}
\end{aligned}
\tag{1}
$$

However, as discussed above, the misalignment between predictions of LLMs and labels, selective documentation, reporting granularity, etc cause the incomplete labeling issue. Therefore, some true positive predictions are ignored if the labels are incomplete, resulting in lower *precision*, *recall*, and *F1-score* than actual performance. Nevertheless, it is impractical to encompass all atomic events within the gold labels, as even Wikipedia only documents the key events and lacks fine-grained developments. Therefore, we propose a modified method to calculate the precision, recall, and F1-score for STF and LTF as follows. In the **modified formulas**, we additionally introduce *MTP* which denotes missing true predictions that can be verified through external evidence or reasoning rather than given gold labels.

$$
\begin{aligned}
Precision_{open} &= \frac{TP+MTP}{TP+MTP+FP} \\
Recall_{open} &= \frac{TP+MTP}{TP+MTP+FN} \\
F1_{open} &= 2 \cdot \frac{Precision_{open} \cdot Recall_{open}}{Precision_{open} + Recall_{open}}
\end{aligned}
\tag{2}
$$

B.4 Analysis on Finetuning

We finetuned the Llama3-8b using balanced multi-task datasets. The instruction part includes the prompt template, event background, and gold answer.

As shown in Table 2, although SFT yields substantial improvements in closed-ended tasks, they exhibit performance degradation in open-ended tasks. Based on the analysis in Section 6.2, three primary reasons may contribute to this. (1) Discrepancy between predictions and labels. The original predictions of LLMs differ greatly from the labels in terms of pessimism bias, number of atomic events, text style, etc. (2) Missing true predictions. LLMs show potential in making comprehensive predictions, including multifaceted impacts (politics, economy, and diplomacy, etc), long-term impacts, international reactions. However, some correct predictions require additional web searching or commonsense reasoning for verification but are not included in the original articles and labels. Consequently, these predictions are erroneously regarded as negative during finetuning. The discrepancy and incomplete labeling together introduce significant instability into the finetuning. (3) The gap between closed-ended and open-ended tasks. Unlike closed-ended tasks, open-ended tasks feature unconstrained answer spaces, which might make the multi-task finetuning unbalanced and unstable. To alleviate these two problems, advanced training methods such as process-supervised RL Lightman et al., 2023) and correction-based learning (An et al., 2023) might be promising.

Model	MCNC	MCNC*
Llama2-7b	30.5	25.3
Llama2-13b	42.8	26.0
Llama2-70b	47.0	30.2
Vicuna-13b	47.3	27.0
Wizardlm-13b	45.3	26.1
Falcon-40b	41.0	26.2
Mixtral-8x7b	55.2	36.2
Llama3-8b	55.7	31.8
Llama3-8b-SFT	96.1	74.5
Avg	51.2	33.7

Table 5: The comparison of the structured and unstructured representation format. The MCNC* task represents the structured variant of MCNC.

B.5 Structured or Unstructured?

Based on the MCNC dataset, using LLMs, we perform open event extraction on the background texts and candidate answers to create a structured MCNC dataset. Then we convert the structured events to text as other works (Luo et al., 2024) and employ the same prompt templates as MCNC for evaluation (as shown in MCNC* column). We evaluate the *extraction accuracy* of this open event extraction using LLMs. For each event timeline from A.3, we randomly select one atomic event and its corresponding structured result. Following the open information extraction, the extraction is deemed as correct if the arguments within the atomic event are correctly extracted. The annotation consistency ratio between the two annotators' is 89.5%, indicating substantial agreement. The open event extraction achieves an extraction accuracy of 87.0%. Most errors arise from the extraction omission and the reversal of subject and object, which pose challenges for forecasting within the structured format.

As illustrated in Table 5, there is a dramatic decline across all LLMs, including Mixtral-8x7b