

- [32] Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. 2024. Large Language Models as Evaluators for Recommendation Explanations. arXiv:2406.03248 [cs.IR]
- [33] Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–37.
- [34] Fangqi Zhu, Jun Gao, Changlong Yu, Wei Wang, Chen Xu, Xin Mu, Min Yang, and Ruifeng Xu. 2023. A generative approach for script event prediction via contrastive fine-tuning. In *Proceedings of the AAAI*. Article 1576, 9 pages.
- [35] Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. 2022. Forecasting Future World Events With Neural Networks. In *Proceedings of the NeurIPS*, Vol. 35. 27293–27305.

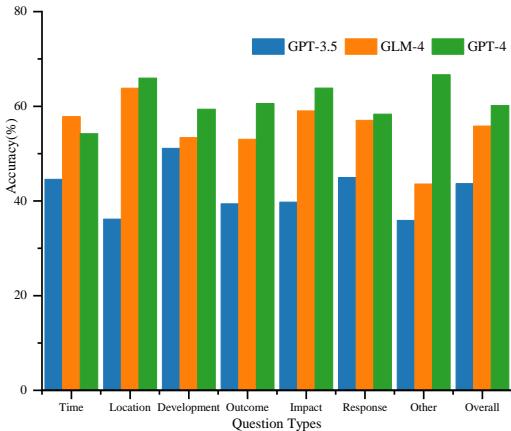


Figure 9: Performance of data validity verification on different LLMs.

A Data Validity Verification

We adopt a data construction strategy that involves annotating questions on the same day they are tested, with answers collected after the prediction window for evaluation. This means that at the time of question annotation, it is unknown whether there will be outcomes, inevitably leading to instances where no answers are available. Such instances are termed as invalid questions, and are excluded from the dataset. Notably, we collect 983 valid questions and annotate 286 invalid ones. Although these questions lack answers, they are valuable for evaluating the model's capacity to identify question validity. Therefore, based on these 286 invalid questions, we extract an equivalent number of valid questions to test whether three powerful LLMs, such as GPT-4, GLM-4, and GPT-3.5, can identify which questions would have answers and which would not during the prediction window.

The results on different types of questions are shown in Figure 9. The results indicate that event identify the data validity is challenging for existing LLMs. We also have the following three observations: (1) For model perspectives, GPT-4 achieves the best performance, except in time-related questions where GLM-4 slightly outperforms. (2) Regarding question types, time-related questions exhibit the lowest accuracy, and GLM-4 achieves the highest accuracy in this category with 57.83%. (3) In terms of overall scores, GPT-4 leads in performance. However, GPT-3.5 achieves an accuracy score of 43.67%.

B Performance on Individual Dimension

This section aims to provide a detailed analysis of the specific scores across different question types for each evaluation dimension. Figures 10 and 11 display the experimental results for GPT-3.5, while Figures 12 and 13 present the results for GLM-4. Overall, the experimental outcomes exhibit similar trends across both LLMs. Completeness scores are the lowest, indicating that making comprehensive future predictions is highly challenging. Reasonableness scores are relatively higher, suggesting that the predictions generated by the large models are logically consistent.

C Details of Dataset Construction

We design the following six principles to better assist LLMs and human annotators in constructing the data.

(1) **Real-time Principle.** Events must be currently occurring. Data related to events not happening in real-time should be discarded, with potential scenarios including: (a) An event that occurred years ago has become a hot topic, such as “*A female employee was fired for using an umbrella at work to avoid exposure, and the court ruled the company's termination legal*”. (b) An event that happened some time ago and has been a hot topic for a while, such as “*How should one evaluate the role of a full-time postdoctoral fellow at Sichuan University*”.

(2) **Answerability Principle.** For a predictive question, it must be ascertainable and answerable to warrant annotation. Unanswerable questions should be discarded. For example, “*How might the performance of the Chinese team in the next 15 days affect its status in international football?*” On one hand, a few matches alone are insufficient to determine impacts on international status. On the other, the outcomes of such questions may not become apparent within 15 days and should therefore be discarded.

(3) **Specificity Principle.** Vague questions and those with broad, indeterminate answers should be discarded. For example, “*What impact might the STSS epidemic have on Japan in the next 15 days?*” This question is unclear as the impacts could span multiple aspects, including economic and political, and should therefore be discarded.

(4) **Continuity Principle.** An event must still be unfolding and not concluded to justify its annotation. Events that have already ended should be discarded.

(5) **Short-Term Principle.** The current task of future event prediction primarily predicts events that may occur within the next 15 days. Therefore, it is necessary to analyze whether a predictive question can yield results within this prediction window. For example, “*What new laws might be proposed in response to construction safety incidents?*” Legislative proposals typically do not yield results within 15 days and should be discarded.

(6) **Truthfulness Principle.** Events that are annotated must be real and currently occurring. People may pose discussions about events that have not actually happened. For example, “*Is there a future in opening all-female nursing homes for older single women?*” or “*Do you remember what you did on the night after the college entrance exam ended?*”.

Furthermore, we have developed an annotation system. The question annotation interface is depicted in Figure 14, and the ground truth annotation interface is shown in Figure 15.

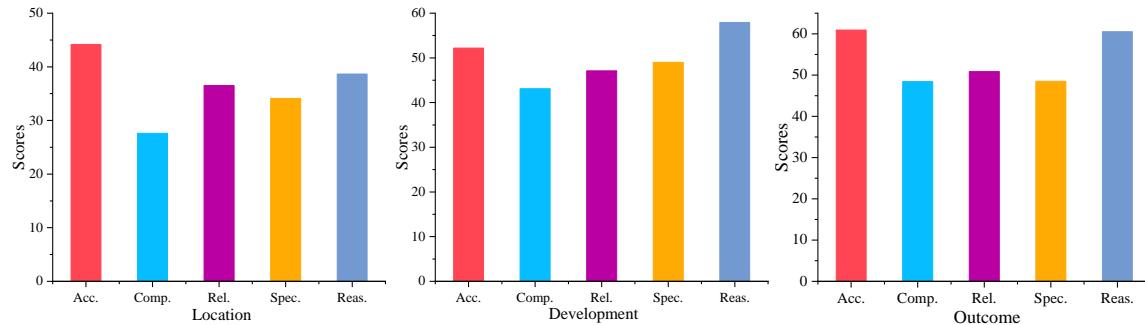


Figure 10: GPT-3.5 Performance of individual dimension on Location, Event Development, and Event Outcome. Acc., Comp., Rel., Spec., and Reas. are abbreviations for Accuracy, Completeness, Relevance, Specificity, and Reasonableness, respectively.

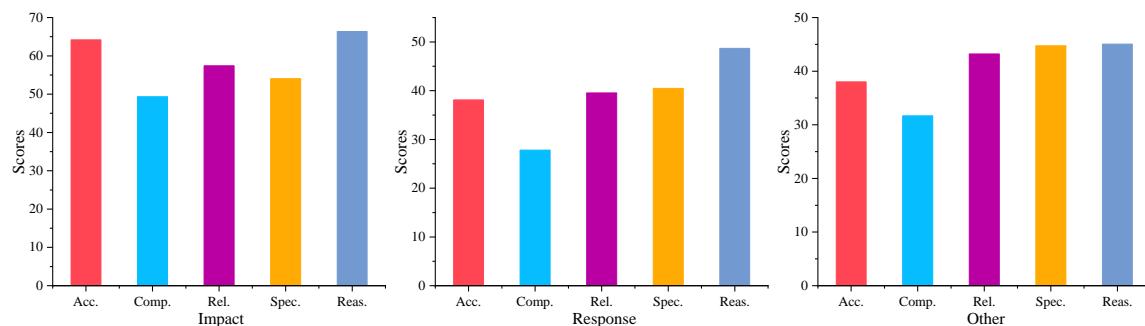


Figure 11: GPT-3.5 Performance of individual dimension on Event Impact, Event Response, and Other.

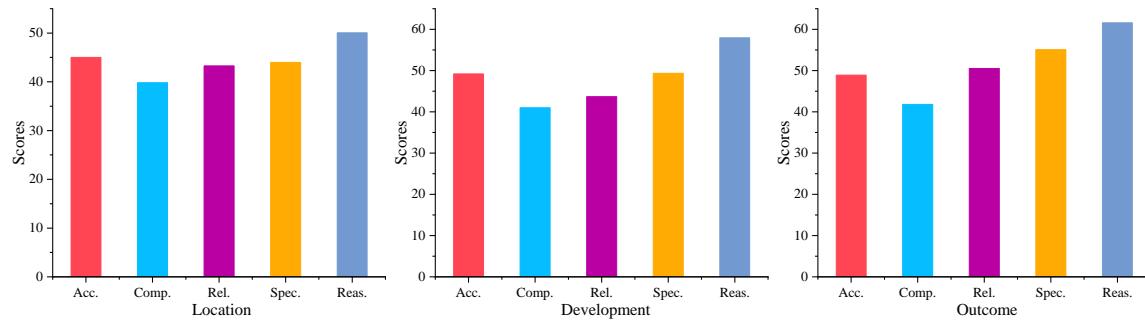


Figure 12: GLM-4 Performance of individual dimension on Location, Event Development, and Event Outcome.

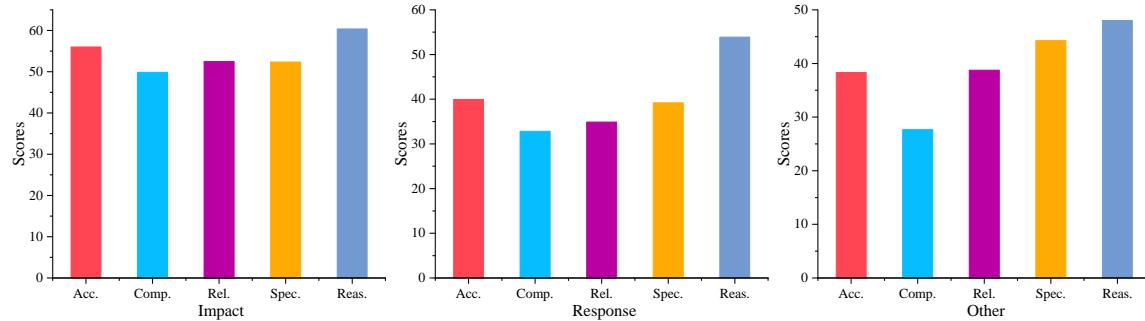


Figure 13: GLM-4 Performance of individual dimension on Event Impact, Event Response, and Other.