Table 2: Model performance of different types of questions on English data (%).

Models	Methods	Time	Location	Development	Outcome	Impact	Response	Other	Overall
GPT-3.5	DR + Summ	35.18	29.98	32.93	46.24	50.51	35.96	42.93	37.41
	DR + Summ-o-Summ	38.12	37.44	29.21	49.84	53.55	38.74	48.34	39.85
	GQR + Summ-o-Summ	42.87	34.65	33.47	48.29	57.59	45.77	51.89	43.58
	StkFEP	**44.85**	**38.63**	**35.81**	**50.42**	**60.68**	**49.74**	**52.03**	**46.03**
GLM-4	DR + Summ	33.53	35.79	34.32	39.51	46.87	32.37	32.72	35.86
	DR + Summ-o-Summ	38.26	30.91	35.88	40.81	47.51	36.15	**50.40**	39.54
	GQR + Summ-o-Summ	42.88	45.47	33.68	40.59	51.24	40.24	36.52	42.62
	StkFEP	**43.31**	**48.31**	**36.40**	**41.39**	**54.70**	**40.63**	37.57	**43.11**
Llama3-8B	DR + Summ	26.34	22.92	28.24	48.73	42.80	34.11	25.29	31.34
	DR + Summ-o-Summ	29.15	23.17	25.77	46.82	44.90	40.50	31.96	32.51
	GQR + Summ-o-Summ	34.50	11.13	28.18	38.98	**55.04**	38.65	32.92	35.16
	StkFEP	**38.66**	**29.50**	**28.35**	**48.93**	50.93	**41.44**	**46.29**	**38.77**
Mistral-7B	DR + Summ	32.26	32.52	30.48	35.32	46.48	32.83	41.20	34.12
	DR + Summ-o-Summ	35.87	31.06	30.98	37.25	54.36	29.88	38.06	36.70
	GQR + Summ-o-Summ	39.12	**43.01**	31.56	36.72	**55.84**	32.03	36.81	38.94
	StkFEP	**41.38**	32.53	**31.79**	**43.07**	51.58	**37.93**	**57.84**	**41.24**

questions based on the question or background for retrieval, similar to existing work [3, 7].

For Integration, we select two comparison methods: (1) *Summ*, which generates a summary for each retrieved document, similar to existing work [7, 27]; (2) *Summ-over-Summ*, which first generates summaries for each document and then produces a brief description of all summaries.

Finally, for each backbone LLM, we employ three combination strategies as baselines, including *DR + Summ*, *DR + Summ-over-Summ*, and *GQR + Summ-over-Summ*. For the prediction module, all baselines utilize the same prediction framework.

4.3 Overall Results

The comparative performances of various methods on Chinese and English data are detailed in Tables 1 and 2 respectively. Our approach StkFEP, which integrates stakeholders insights and information from similar events, consistently outperformed other methods. We also have four key observations:

(1) For time-related questions, the current best result is 44.85%. In our experiments, we set these questions as multiple-choice format and divided the prediction window into three intervals. Additionally, we tested the performance of GPT-3.5 over five intervals and found a decrease to 25.23%, indicating that these questions remain highly challenging.

(2) From the perspective of retrieval methods, using prediction questions directly for retrieval yields the poorest results, while employing LLMs to generate diverse questions shows improvement.

(3) In terms of information integration, the *Summ-o-Summ* approach, which uses summarization twice, performs better than a single summarization *Summ*, indicating that this method can further refine content.

(4) From the perspective of different languages, the model exhibits similar trends across all languages. The performance on questions related to *Development* is relatively lower.

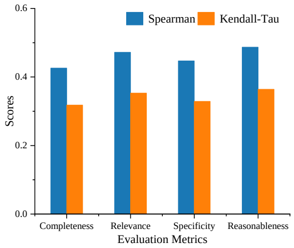


Figure 5: The correlations between human and LLMs.

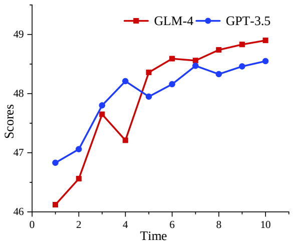


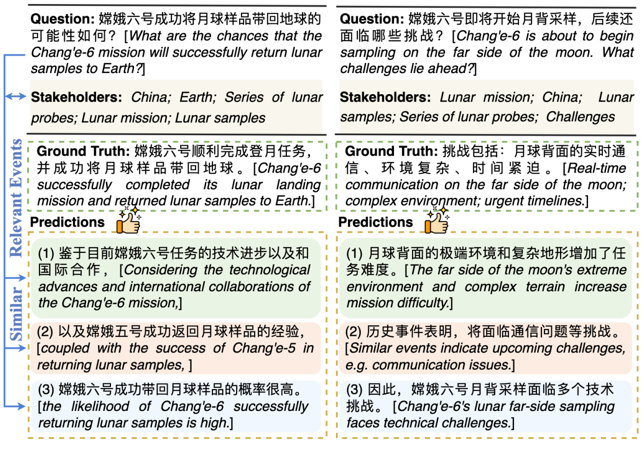
Figure 6: Performance of daily prediction.

4.4 Human Evaluation

In this section, we expand our evaluation methodology beyond model-based metric. We conduct an additional human evaluation to compare 50 predictions generated by GPT-3.5. We invite annotators to assess the model outputs from four dimensions: *Completeness*, *Relevance*, *Specificity*, and *Reasonableness*, using the same criteria as the automatic evaluation method. We report the Spearman and Kendall-Tau correlations between human expert-annotated scores and GPT-4 assigned scores in Figure 5. We find that GPT-4 achieves a Spearman correlation of around 0.45, which indicates that recent LLMs perform predictions evaluations that are reasonably valid to a meaningful extent.

4.5 Analysis of Daily Prediction

We conduct daily predictions to capture the trends of predictions changing over time. To achieve this, we select 22 questions that will yield results after 10 days, organizing a test each day. The experimental results, as shown in Figure 6, indicate that the model performance generally improves over time with updates in information. Upon deeper analysis, we observe that during the initial days, the scale of information is substantial, encompassing both



Question: 嫦娥六号成功将月球样品带回地球的可能性如何？[*What are the chances that the Chang'e-6 mission will successfully return lunar samples to Earth?*]

Stakeholders: *China; Earth; Series of lunar probes; Lunar mission; Lunar samples*

Ground Truth: 嫦娥六号顺利完成登月任务，并成功将月球样品带回地球。[*Chang'e-6 successfully completed its lunar landing mission and returned lunar samples to Earth.*]

Predictions

(1) 鉴于目前嫦娥六号任务的技术进步以及和国际合作，[*Considering the technological advances and international collaborations of the Chang'e-6 mission,*]

(2) 以及嫦娥五号成功返回月球样品的经验，[*coupled with the success of Chang'e-5 in returning lunar samples,*]

(3) 嫦娥六号成功带回月球样品的概率很高。[*the likelihood of Chang'e-6 successfully returning lunar samples is high.*]

(a) Enhancing Prediction Credibility

Question: 嫦娥六号即将开始月背采样，后续还面临哪些挑战？[*Chang'e-6 is about to begin sampling on the far side of the moon. What challenges lie ahead?*]

Stakeholders: *Lunar mission; China; Lunar samples; Series of lunar probes; Challenges*

Ground Truth: 挑战包括：月球背面的实时通信、环境复杂、时间紧迫。[*Real-time communication on the far side of the moon; complex environment; urgent timelines.*]

Predictions

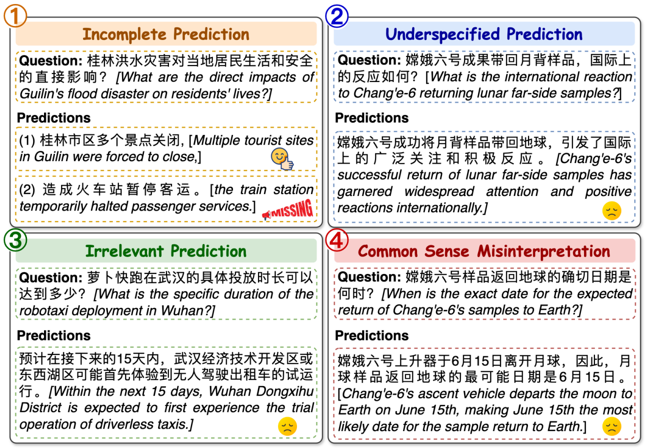
(1) 月球背面的极端环境和复杂地形增加了任务难度。[*The far side of the moon's extreme environment and complex terrain increase mission difficulty.*]

(2) 历史事件表明，将面临通信问题等挑战。[*Similar events indicate upcoming challenges, e.g. communication issues.*]

(3) 因此，嫦娥六号月背采样面临多个技术挑战。[*Chang'e-6's lunar far-side sampling faces technical challenges.*]

(b) Providing Evolutionary Patterns

Figure 7: Cases for model predictions.



Question: 桂林洪水灾害对当地居民生活和安全的直接影响？ *[What are the direct impacts of Guilin's flood disaster on residents' lives?]*

Predictions

(1) 桂林市区多个景点关闭, [*Multiple tourist sites in Guilin were forced to close,*]

(2) 造成火车站暂停客运。[*the train station temporarily halted passenger services.*]

Question: 嫦娥六号成果带回月背样品，国际上的反应如何？[*What is the international reaction to Chang'e-6 returning lunar far-side samples?*]

Predictions

嫦娥六号成功将月背样品带回地球，引发了国际上的广泛关注和积极反应。 *[Chang'e-6's successful return of lunar far-side samples has garnered widespread attention and positive reactions internationally.]*

Question: 萝卜快跑在武汉的具体投放时长可以达到多少？ *[What is the specific duration of the robotaxi deployment in Wuhan?]*

Predictions

预计在接下来的15天内，武汉经济技术开发区或东西湖区可能首先体验到无人驾驶出租车的试运行。*[Within the next 15 days, Wuhan Dongxihu District is expected to first experience the trial operation of driverless taxis.]*

Question: 嫦娥六号样品返回地球的确切日期是何时？ *[When is the exact date for the expected return of Chang'e-6's samples to Earth?]*

Predictions

嫦娥六号上升器于6月15日离开月球，因此，月球样品返回地球的最可能日期是6月15日。[*Chang'e-6's ascent vehicle departs the moon to Earth on June 15th, making June 15th the most likely date for the sample return to Earth.*]

Figure 8: Error analysis of the model predictions.

redundant and critical details, leading to significant fluctuations. As time progresses, public discussion about the issues diminishes, resulting in smaller fluctuations during this phase.

4.6 Case Study

To better understand the results shown in Table 3, we conduct a case study to explicitly illustrate the effectiveness of the event prediction framework. The cases are shown in Figure 7. For the first case, by identifying stakeholders such as *Lunar mission*, *Lunar samples*, and *China*, model can effectively retrieve similar events like "*Chang'e-5 successfully returning lunar samples*". By then incorporating relevant events, it can significantly enhance the credibility of the predictions. For the second case, similar events can provide potential evolutionary patterns to support prediction. Retrieving similar events allows us to learn about challenges faced by previous lunar sampling missions, such as *communication issues*, and combining this with the progress and breakthroughs in current research, can enhance the effectiveness of event prediction.

4.7 Error Analysis

To enrich the understanding and better advance future research, we conduct a detailed analysis of the problems encountered in existing research. The common problems can primarily be divided into four categories: (1) **Incomplete Prediction** refers to scenarios where the predictions made are not comprehensive enough to cover all aspects or variables related to the event. As shown in case 1 of Figure 8, the model overlooks the outcome "*the train station temporarily halted passenger services*". (2) **Underspecified Prediction** occurs when predictions are too vague or general, lacking specific details necessary for them to be actionable or useful. As shown in case 2, the model outputs "*Chang'e-6's successful return of lunar far-side samples has garnered widespread attention and positive reactions internationally*". The predictions of model lacks value because it does not provide any salient entity information, resulting in an output too generic to effectively address the specific question. (3) **Irrelevant Prediction** describes predictions is unrelated to the question posed, essentially providing answers that do not address the question. As shown in case 3, the question asks about time information, but the model responds with a location information "*Wuhan Dongxihu District*". (4) **Common Sense Misinterpretation** arises when predictions contradict basic common sense, resulting in outcomes that are implausible or logically inconsistent with known facts. This undermines the credibility of the predictions and may lead to mistrust or disregard of model outputs. In case 4, the statement "*Chang'e-6's ascent vehicle departs the moon for Earth on June 15th*" is predicted, however, the model overlooks the common sense that it is impossible to return from the moon to Earth within a day.

Table 3: Ablation study.

Methods	GPT-3.5	GLM-4
StkFEP	46.95	46.27
w/o Cluster-Summ	46.11	45.38
w/o Similar Events	45.65	44.79
w/o Stakeholders	44.28	42.80

4.8 Ablation Study

To more specifically validate the different modules within the event prediction framework, we conduct experiments to ablate the clustering-over-summarization method (*w/o Cluster-Summ*) for information integration, similar events (*w/o Similar Events*), and stakeholders (*w/o Stakeholders*). From the results in Table 3, we can see that: (1) For the scenario without cluster summarization (*w/o Cluster-Summ*), where we used Summ-over-Summ for information integration, the model performance decreased, indicating that our method can more effectively refine information and organize dependencies between events. (2) For the scenario without similar events (*w/o Similar Events*), relying only on relevant events for predictions, the model results also declined, mainly because similar events provide potential evolutionary patterns that support the final predictions. (3) For the scenario without stakeholders (*w/o*

Stakeholders), ignoring stakeholders resulted in the most substantial drop in model performance. This demonstrates that utilizing stakeholders not only enhances the diversity of retrieval but also enables more accurate retrieval of similar events.

5 Conclusions

In this paper, we introduce OpenEP (an open-ended future event prediction task), which generates flexible and diverse predictions aligned with real-world scenarios. To facilitate the study of this task, we first construct OpenEPBench, an open-ended future event prediction dataset. For question construction, we pose questions from seven perspectives, including location, time, event development, event outcome, event impact, event response, and other, to facilitate an in-depth analysis and understanding of the comprehensive evolution of events. For outcome construction, we collect free-form text containing the outcomes as ground truth to provide semantically complete and detail-enriched outcomes. Furthermore, we propose StkFEP, a stakeholder-enhanced future event prediction framework that incorporates the characteristics of event evolution for open-ended settings. Our method extracts stakeholders involved in events to extend questions and collects historical events that are relevant and similar to the question to gather diverse and comprehensive information to support model prediction. Extensive experiments on Chinese and English data demonstrate that accurately predicting future events in open-ended settings is challenging for existing large language models.

References

[1] Samuel Adeyemo and Debangsu Bhattacharyya. 2024. Optimal nonlinear dynamic sparse model selection and Bayesian parameter estimation for nonlinear systems. *Computers & Chemical Engineering* 180 (2024), 108502.

[2] Long Bai, Saiping Guan, Zixuan Li, Jiafeng Guo, Xiaolong Jin, and Xueqi Cheng. 2023. Rich event modeling for script event prediction. In *Proceedings of the AAAI*. Article 1409, 9 pages.

[3] Junyan Cheng and Peter Chin. 2024. SocioDojo: Building Lifelong Analytical Agents with Real-world Text and Time Series. In *Proceedings of the ICLR*.

[4] Walter H. Dempsey, Alexander Moreno, Christy K. Scott, Michael L. Dennis, David H. Gustafson, Susan A. Murphy, and James M. Rehg. 2017. iSurvive: An Interpretable, Event-time Prediction Model for mHealth. In *Proceedings of the ICML (Proceedings of Machine Learning Research, Vol. 70)*. 970–979.

[5] Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793 [cs.CL]

[6] Yong Guan, Xiaozhi Wang, Lei Hou, Juanzi Li, Jeff Z. Pan, Jiaoyan Chen, and Freddy Lecue. 2024. TacoERE: Cluster-aware Compression for Event Relation Extraction. In *Proceedings of the LREC-COLING*. 15511–15521.

[7] Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching Human-Level Forecasting with Language Models. arXiv:2402.18563 [cs.LG]

[8] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features. In *Proceedings of the ACL*. 987–997.

[9] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL]

[10] Alapan Kuila and Sudeshna Sarkar. 2024. From Text to Context: An Entailment Approach for News Stakeholder Classification. In *Proceedings of the SIGIR*. 2426–2430.

[11] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. 591–600.

[12] Srivatsan Laxman, Vikram Tankasali, and Ryen W. White. 2008. Stream prediction using a generative model based on frequent episodes in event sequences. In *Proceedings of the SIGKDD*. 453–461.

[13] Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021. The Future is not One-dimensional: Complex Event Schema Induction by Graph Modeling for Event Prediction. In *Proceedings of the EMNLP*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). 5203–5215.

[14] Xiaoyan Liu, Jiarui Zhao, Ran Liu, and Kai Liu. 2022. Event history analysis of the duration of online public opinions regarding major health emergencies. *Frontiers in Psychology* 13 (2022), 954559.

[15] Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, and Tat-Seng Chua. 2023. Context-aware Event Forecasting via Graph Disentanglement. In *Proceedings of the SIGKDD*. 1643–1652.

[16] Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, and Tat-Seng Chua. 2023. Context-aware Event Forecasting via Graph Disentanglement. In *Proceedings of the SIGKDD*. 1643–1652.

[17] Yaswanth Narsupalli, Abhranil Chandra, Sreevatsa Muppirala, Manish Gupta, and Pawan Goyal. 2024. Review-Feedback-Reason (ReFeR): A Novel Framework for NLG Evaluation and Reasoning. arXiv:2407.12877 [cs.CL]

[18] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 https://arxiv.org/pdf/2303.08774.pdf

[19] Sarah Pratt, Seth Blumberg, Pietro Kreitlon Carolino, and Meredith Ringel Morris. 2024. Can Language Models Use Forecasting Strategies? arXiv:2406.04446 [cs.LG]

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the EMNLP-IJCNLP*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). 3982–3992.

[21] Shakila Khan Rumi, Ke Deng, and Flora D. Salim. 2018. Theft prediction with individual risk factor of visitors. In *Proceedings of the SIGSPATIAL*. 552–555.

[22] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. arXiv:2401.18059 [cs.CL]

[23] Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock. 2024. Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy. arXiv:2402.19379 [cs.CY]

[24] Smriti Sharma, Rajesh Kumar, Pawan Bhadana, and Sumita Gupta. 2013. News event extraction using 5W1H approach & its analysis. *International Journal of Scientific & Engineering Research* 4, 5 (2013), 2064–2068.

[25] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[26] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *Proceedings of the ICLR*.

[27] Qi Yan, Raihan Seraj, Jiawei He, Lili Meng, and Tristan Sylvain. 2024. AutoCast++: Enhancing World Event Prediction with Zero-shot Ranking-based Context Retrieval. arXiv:2310.01880 [cs.LG]

[28] Haoyan Yang, Yixuan Wang, Xingyin Xu, Hanyuan Zhang, and Yirong Bian. 2024. Can We Trust LLMs? Mitigate Overconfidence Bias in LLMs through Knowledge Transfer. arXiv:2405.16856 [cs.CL]

[29] Yiying Yang, Zhongyu Wei, Qin Chen, and Libo Wu. 2019. Using External Knowledge for Financial Event Prediction Based on Graph Neural Networks. In *Proceedings of the CIKM*. 2161–2164.

[30] Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. 2024. MIRAI: Evaluating LLM Agents for Event Forecasting. arXiv:2407.01231 [cs.CL]

[31] Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives. arXiv:2401.02009 [cs.CL]