Model	AQA	STF			LTF			MCNC	MCAC	VQA
		P	R	F1	P	R	F1			
BertMultipleChoice				/				91.5	80.3	83.7
Llama2-7b	62.0	35.1	58.8	41.6	47.8	53.8	48.2	30.5	52.5	57.5
Llama2-13b	57.5	41.1	58.5	45.9	46.2	53.9	46.5	42.8	51.6	55.3
Llama2-70b	**63.5**	42.4	66.2	48.8	**48.5**	60.8	**50.7**	47.0	67.8	57.9
Vicuna-13b	44.5	25.0	45.3	29.0	45.4	48.4	44.7	47.3	65.3	51.5
Wizardlm-13b	50.5	33.3	43.5	35.8	37.4	49.5	38.5	45.3	61.7	54.8
Falcon-40b	56.5	29.5	30.3	28.1	21.6	31.0	23.2	41.0	56.5	59.0
Mixtral-8x7b	61.0	52.3	**70.6**	**57.2**	45.2	**61.1**	47.7	55.2	67.7	64.6
Llama3-8b	63.0	**52.4**	70.1	56.8	47.2	59.5	49.5	55.7	67.3	57.0
Llama3-8b-SFT	61.5	27.8	21.2	22.9	27.8	21.2	22.9	**96.1**	**89.2**	**89.3**

Table 2: Experimental results (%) of diverse models on open-ended (AQA, STF, and LTF) and closed-ended (MCNC, MCAC, and VQA) tasks.

Model	Positive			Negative			T/F
	P	R	F1	P	R	F1	
Llama2-7b	54.5	83.2	65.8	66.8	32.8	43.9	3.01
Llama2-13b	52.7	89.9	66.4	69.1	21.9	33.2	5.21
Llama2-70b	54.2	92.2	68.4	76.9	25.0	37.7	5.06
Vicuna-13b	65.5	3.0	5.8	51.2	**98.5**	67.4	0.02
Wizardlm-13b	52.4	89.8	66.2	68.0	21.0	32.1	5.39
Falcon-40b	56.6	71.1	63.0	62.8	47.2	53.9	1.62
Mixtral-8x7b	60.2	82.3	69.6	73.5	47.4	57.7	2.05
Llama3-8b	53.5	**96.5**	68.8	**84.7**	18.8	30.8	7.84
SFT	**78.9**	71.4	**75.0**	74.3	81.3	**77.7**	0.96

Table 3: The significant disparity between positive and negative samples in the VQA task (%). The *T/F* denotes the ratio of the number of samples predicted as positive to those predicted as negative. The ratio of positive and negative samples (label) is 0.97.

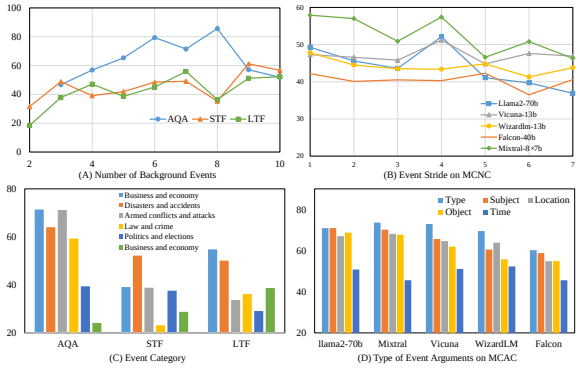

Figure 6: Performances (%) analysis of influencing factors.

tional reactions, etc. (4) **LLMs tend to make pessimistic forecasting.** For instance, when presented with an event background involving protest, impeachment, or border clash, LLMs often make pessimistic predictions such as government collapse, long-term instability, or large-scale wars (as shown in Table 7). (5) **LLMs tend to predict the occurrence of candidate events.** As demonstrated in Table 3, most LLMs achieve much higher recall but lower precision for positive candidates than negative ones and show a notable tendency to predict true rather than false.

Influencing Factors for Event Forecasting. Compared to other LLMs, Llama2-70b, Mixtral-8x7b, and Llama3-8b perform better, which aligns with their ranking on the leaderboard, thus confirming the importance of general capabilities for forecasting. We further investigate four influencing factors, as depicted in Figure 6. Due to the limited number of human evaluations on open-ended tasks, the average scores of six LLMs, including Mixtral-8x7b, Llama2-70b, Vicuna-13b, Wizardlm-13b, Falcon-40b, and Llama3-8b, are adopted in Figure 6 (A) and (C). The results in Figure 6 (A) indicate that more event backgrounds might improve prediction accuracy. Figure 6 (B) indicates that larger event strides (the distance between background and target event) are much harder due to increased uncertainty. In Figure 6 (C), the performances of armed conflicts and attacks, international relations, law and crime, and politics and elections are inferior to those of disasters and accidents, because the disaster responses (such as post-disaster relief and material assistance) are quite similar. Figure 6 (D) indicates that the performance distribution of LLMs across argument types remains consistent, with the best results in *Type* while the worst results in *Time*. Given the inherent challenges in time prediction (Zhao, 2022), it may be beneficial

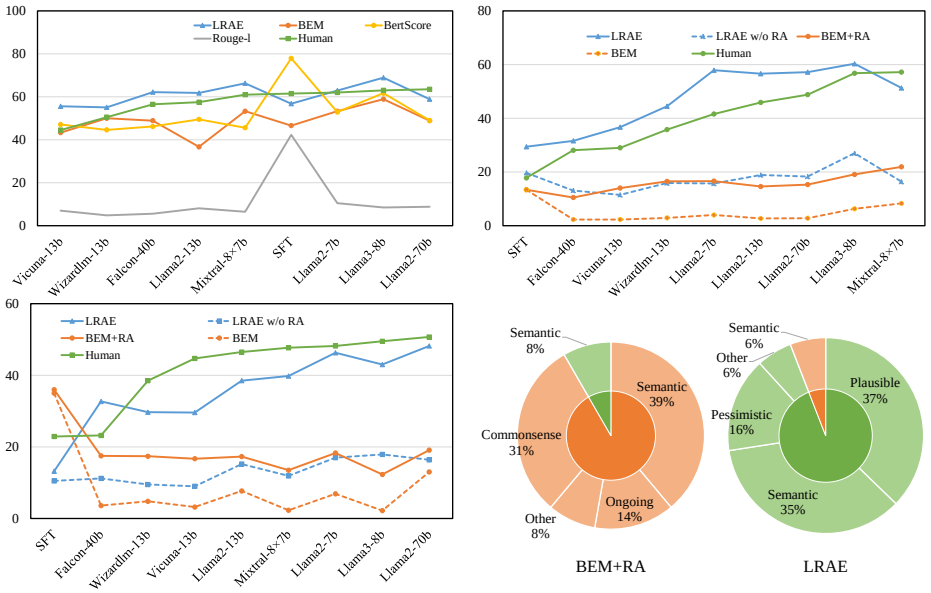

Figure 7: Analysis of open-ended evaluation methods. The upper left, upper right, and lower left figures depict the comparative performance (%) of human and automatic evaluation methods across AQA, STF, and LTF respectively. The lower right graph illustrates the statistic of error types for the BEM+RA and LRAE, with orange segments representing missing errors (positive predictions are misclassified as negative) and green segments representing overestimation errors (negative predictions are misclassified as positive). Detailed definitions for the error types are shown in Table 6.

to exclude time during fine-tuning.

The analysis on SFT and event representation format can be found in Appendix B.4 and B.5.

6.3 Experiments on LRAE

In this section, we evaluate four automatic evaluation methods on open-ended tasks: (1) **Rouge-L** (F1), a lexical matching method; (2) **BertScore** (F1), a token-level semantic matching method; (3) **BEM**[3] (accuracy for AQA and F1 for STF and LTF), a sentence-level semantic matching method; (4) **LRAE**[4] (accuracy for AQA and F1 for STF and LTF), a LLM-based semantic matching method.

As depicted in Figure 7, we sort the LLMs (on the horizontal axis) according to their performance using human evaluations, yielding a progressively ascending line. In the AQA task (upper left), a notable discrepancy is observed between Rouge-L and human evaluations. This is because the answers of LLMs are lengthy while the labels are short, resulting in a low overlap. Additionally, other evaluation methods show good consistency with human evaluation, with our LRAE performing the best, indicating that language models, especially LLMs, can match semantically equivalent answers despite significant textual differences.

For the many-to-many matching tasks (STF: upper right, LTF: lower left), LRAE and BEM are selected[5]. However, the gaps between them and human evaluations become larger. Unlike AQA (argument-level one-to-one matching), STF and LTF are required to match multiple atomic events, each with multiple event arguments (event-level many-to-many matching), thereby significantly increasing the difficulties. Notably, LRAE exhibits the best consistency with human evaluation, particularly on STF, due to the strong language understanding ability of LLMs and the retrieval augmentation module (RA). After removing RA, there is a significant performance decline for LRAE w/o RA, demonstrating the effectiveness of RA. The detailed results of LRAE and BEM are presented in Table 8, and 9. However, LRAE exhibits minor ranking discrepancies compared to human evaluation, indicating the need for further refinement.

Furthermore, on STF and LTF, we collect samples with the absolute F1-score differences between

[3]For AQA, BEM conducts once matching between prediction and label. For STF and LTF, BEM conducts multiple one-to-one matching for each atomic prediction-label pair.

[4]For AQA, we remove the retrieval augmentation from LRAE and employ once matching using LLM.

[5]For an atomic prediction-label pair, BertScore and Rouge-L provide continuous outputs, rather than binary 'Yes' or 'No' classifications, thus are not applicable to many-to-many matching.

human and automatic evaluation exceeding 0.3. We then investigate the reasons for these failures of BEM+RA and LRAE, as depicted in Figure 7 (lower right). For BEM+RA, 92% of errors are matching missing (positive predictions are judged as negative). This issue stems primarily from suboptimal semantic modeling ability. Additionally, BEM lacks commonsense reasoning ability and fails in handling ongoing events (as shown in Table 6). In contrast, LRAE alleviates the matching missing issues but introduces minor overestimation errors (negative predictions are judged as positive). Among these errors, LRAE often regards plausible and pessimistic predictions as correct, even in the absence of relevant content supporting the prediction. These failures indicate that LLM-based evaluation may suffer from hallucinations and thus needs further optimization of LLMs.

7 Conclusion

To promote event forecasting from the closed-ended paradigm to the open-ended paradigm, we introduce OpenForecast, an open-ended event forecasting dataset characterized by defining three open-ended tasks and automatically annotating a large-scale dataset from Wikipedia and news. Additionally, we introduce LRAE for the automatic evaluation of open-ended tasks. Using human evaluations and experiments, we demonstrate the quality and challenges of OpenForecast, as well as LRAE's superior consistency with human evaluations. Future work will focus on exploring advanced fine-tuning methods for open-ended tasks and increasing consistency with human evaluation.

Acknowledgement

This research is sponsored by the Xinjiang Uygur Autonomous Region "Tianshan Talents" Scientific and Technological Innovation Leading Talent Project (2022TSYCLJ0035), the Youth Talents Support Project of Xinjiang Uyghur Autonomous Region (2023TSYCQNTJ0037), the "Tianshan Elite" Science and Technology Topnotch Youth Talents Program (2022TSYCCX0059), Tianshan Talent Training Program (2023TSYCCX0041), the Outstanding Member Program of the Youth Innovation Promotion Association of Chinese Academy of Sciences (Y2021112, Y2023118), and the Natural Science Foundation of Xinjiang Uyghur Autonomous Region (2022D01D04, 2022D01B207).

Limitations

In this section, we discuss several limitations in our work. First, the construction of OpenForecast relies on the performance of LLMs and necessitates substantial computational resources. Second, despite demonstrating superior consistency, our LRAE exhibits minor ranking discrepancies when compared to human evaluation. Further research is needed to enhance its robustness.

Ethics Statement

In our study, OpenForecast was developed using open-source projects, including Wikipedia and WCEP. These resources have been widely employed in other studies, ensuring that no ethical standards were compromised.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes LLM better reasoner. *CoRR*, abs/2310.20689.

Matin N. Ashtiani and Bijan Raahemi. 2023. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Syst. Appl.*, 217:119509.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3676–3687. Association for Computational Linguistics.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 291–305. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 77–86. Association for Computational Linguistics.

David Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15607–15631. Association for Computational Linguistics.