Evaluating Instructively Generated Statement by Large Language Models for Directional Event Causality Identification

Wei Xiang[1] * **Chuanhong Zhan**[2] * **Qing Zhang**[2] **Bang Wang**[2] †
[1] Faculty of Artificial Intelligence in Education,
Central China Normal University, Wuhan China.
xiangwei@ccnu.edu.cn
[2] School of Electronic Information and Communications,
Huazhong University of Science and Technology, Wuhan, China
{zhanch, Qing Zhang, wangbang}@hust.edu.cn

Abstract

This paper aims to identify directional causal relations between events, including the existence and direction of causality. Previous studies mainly adopt prompt learning paradigm to predict a causal answer word based on a Pre-trained Language Model (PLM) for causality existence identification. However, the indecision in selecting answer words from some synonyms and the confusion of indicating opposite causal directions with the same answer word raise more challenges in directional causality identification. Inspired by the strong capabilities of pre-trained Generative Language Models (GLMs) in generating responses or statements, we propose to instruct a GLM to generate causality statements and identify directional event causality by evaluating the generated statements. Specifically, we propose an *Instructive Generation and Statement Evaluation* method to identify both the existence and direction of causality. We first fine-tune a GLM to instructively generate causality statements based on event description inputs. Then, we evaluate the rationality of the generated statements to determine the existence and direction of event causalities. Experiments on the ESC and MAVEN datasets show that our method significantly outperforms state-of-the-art algorithms, even with fewer training data.

1 Introduction

Event Causality Identification aims to determine whether a causal relation exists between two event mentions in text, which is of great importance for many downstream applications, such as event prediction (Zhou et al., 2022), event-centric knowledge graph construction (Heindorf et al., 2020), event chain mining (Li et al., 2023), etc. Existing research focuses on identifying the existence of causality between two event mentions (Liu et al., 2023; Huang et al., 2024; Yuan et al., 2023). However, the direction of causality is also crucial for understanding the causal relation. This paper aims at identifying directional causal relations between event mentions, which not only recognizes the existence of causalities but also classifies the cause and effect events when a causal relation exists.

The recent prompt learning methods (Shen et al., 2022; Liu et al., 2021; Man et al., 2024) reformulate each event pair into a prompt template as input for a PLM, so as to utilize the PLM to predict a causal answer word or generate a causality label word for causal relation identification. For example, Shen et al. (2022) design a derivative prompt template to jointly predict the causality label word, causal cue word, and causal event mentions using a PLM, and maps them to a causal relation or none relation. Man et al. (2024) introduce a hierarchical optimal transport approach to automatically select important sentences and words from input documents as input for a PLM to generate both causality label word and salient context words.

We argue that the performance of such prompt learning paradigm is heavily dependent on the manually selected causal answer word or the causality label word. On the one hand, some synonyms all have the semantic of causal relations, but they exhibit subtle differences, such as "*because*, *so*, *cause*". This raises more difficulty and indecision for the selection of causal answer words. On the other hand, these causal words themselves contain causality directions, such as "*cause* and *caused by*" each representing an opposite causal direction. As such, we have to use both of them to indicate different causal directions, making it more confusing and challenge to identify causality directions. Meanwhile, the designed prompt templates also have a significant impact on causal relation predictions.

Instead of using a PLM to predict a selected causal answer word or causality label word for

*These authors contributed equally to this work.
†Corresponding author: Bang Wang

Findings of the Association for Computational Linguistics: ACL 2025, pages 779–785

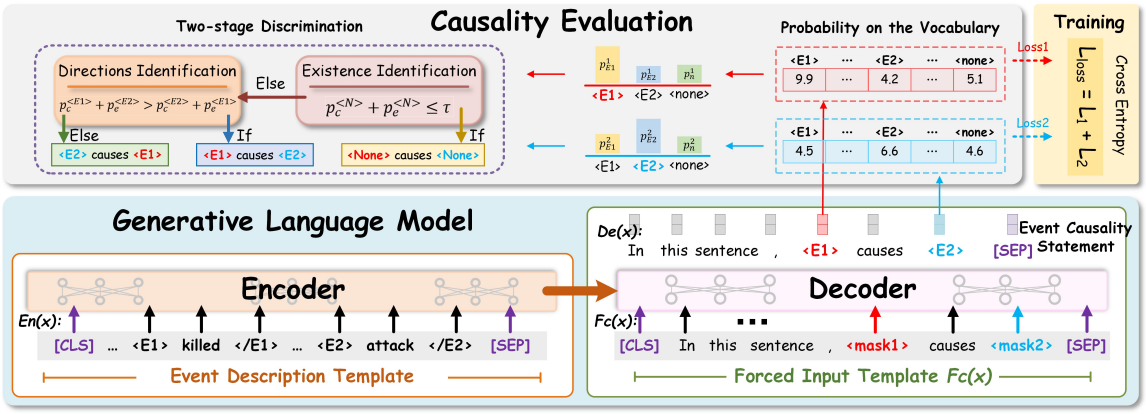

Figure 1: The framework of our proposed directional event causality identification method.

causality identification, we propose to first assume the existence of a causal relation and instruct a pre-trained GLM to generate a causality statement sentence. By evaluating the rationality of the generated statement, we can determine whether the causal relation exists or not, and further identify the causal directions, which neither requires the effort to select causal answer words nor the design of prompt templates. This is also in accordance with the pre-training objective of the GLMs, that is, generating a response or statement given the input text sequence, but not predicting an answer word.

Specifically, we propose an *Instructively Generation and Statement Evaluation* method to identify both the existence and direction of causality. Fig. 1 illustrates the overall framework of our proposed method. We first input the descriptions of each event pair into a GLM and fine-tune it to instructively generate a causality statement in a sequence-to-sequence way. Then we evaluate the rationality of the generated statements to determine the existence of causal relations. When a causal relation exists, we further compare the probabilities of the two event mentions to recognize the cause and effect events for direction identification. Our proposed method achieves the state-of-the-art performance on two public datasets and significant outperforms conventional prompt learning methods. Experiments in low-resource scenario also validates the advantages of our proposed method.

2 Methodology

1) Task Reformulation: Since the event mentions Em are few annotated words within a raw sentence S that includes their full contextual semantics, we concatenate the raw sentences S_i and S_j of an input event pair $x = (Evt_i, Evt_j)$ to form the input event description $En(x)$ for a GLM. If the two input event mentions are within the same sentence, it is directly used as the GLM input. Besides, we insert some virtual event tokens, including <E1> and /<E1>, <E2> and /<E2>, before and after each event mentions Em_i and Em_j, respectively. They are used to represent the two input event mentions, which usually consist of diverse words and varying lengths, for next statement generation and causality evaluation.

We reformulate the directional causality label of each event pair into a simple event causality statement for the instructive generation by a GLM, that is, $De(x)$ = In this sentence, <E1> causes <E2>. Where the event mention denoted by <E1> is the cause, and that of <E2> is the effect. Similarly, if <E2> represents the cause event and <E1> the effect event, we interchange them directly, i.e., "<E2> *causes* <E1>". For the negative samples, i.e., no causal relation between the input event pair, we replace the two virtual event tokens by the virtual token <None>, that is, "<None> *causes* <None>".

2) GLM Fine-tuning: We fine-tune the GLM to generate causality statements in a sequence-to-sequence way, using the event descriptions $En(x)$ as the encoder input and the causality statements $De(x)$ are the target decoder output. To ensure that the GLM can consistently generate causality statements, we adopt the teacher forcing strategy (Goodman et al., 2020), which provides a forced input $Fc(x)$ to the decoder, in addition to the encoder output. The forced input $Fc(x)$ is designed as an event-invisible causality statement, that is, *In this sentence,* <PAD> *causes* <PAD>. Instead of using the predicted previous token as input for the next token generation in GLM decoder, we use the forced input to direct the generation of next token.

Method	*Causality Existence Identification*						*Causality Direction Identification*					
	ESC			MAVEN			ESC			MAVEN		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT (Devlin et al., 2019)	37.2	41.2	39.1	43.3	47.1	45.1	40.7	31.3	35.4	42.4	45.5	43.9
RoBERTa (Liu et al., 2019)	39.7	40.6	40.1	34.6	23.5	28.0	37.3	35.5	36.4	33.8	22.7	27.2
ERGO (Chen et al., 2022)	46.3	50.1	48.1	49.6	62.3	55.2	41.5	43.3	42.4	48.7	60.1	53.8
SENDIR (Yuan et al., 2023)	37.8	82.8	51.9	51.9	52.9	52.4	43.8	43.7	43.7	46.8	43.1	44.9
ChatGLM3-6B (Zeng et al., 2024)	15.3	71.8	25.2	4.3	63.7	8.1	8.4	39.5	13.9	2.5	37.4	4.7
GPT3.5-turbo (Gao et al., 2023)	20.4	58.5	30.2	5.5	38.5	9.6	12.3	35.2	18.2	4.0	28.0	7.0
Our method (BART)	55.8	66.7	60.7	76.8	55.4	64.4	49.0	55.5	52	62.7	56.0	59.1
Our method (LLaMa)	54.6	59.4	57.1	80.6	55.4	65.6	46.2	46.3	46.1	73.3	50.3	59.7

Table 1: Overall results of causality existence and direction identification on both MAVEN and ESC datasets.

Note that, for each generated token, the GLM decoder estimates a probability p for each word in its vocabulary (including the added virtual token words) and adopts the word with the highest probability as the generated output. We compute the cross-entropy loss between the estimated probabilities $\hat{\mathbf{y}}^{(k)}$ and the ground truth label $\mathbf{y}^{(k)}$ for the two virtual event tokens <E1> and <E2>, respectively (denoted as $\mathcal{L}_1$ and $\mathcal{L}_2$):

$$\mathcal{L} = -\frac{1}{K}\sum_{k=1}^{K}\mathbf{y}^{(k)}\log(\hat{\mathbf{y}}^{(k)}) + \lambda\|\theta\|^2, \quad (1)$$

where λ and θ are the regularization hyper-parameters and model parameter respectively, and k is the training instance number. The overall loss is $\mathcal{L}oss = \mathcal{L}_1 + \mathcal{L}_2$. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with $L2$ regularization for the GLM fine-tuning.

3) Causality Evaluation: After fine-tuning, the GLM can consistently generate a templatized causality statement for each event pair. We use the predicted probability of the two event-invisible tokens for causality evaluation, where P_c and P_e correspond to the preceding and subsequent <PAD> tokens, respectively, representing the potential cause and effect event. We first sum the predicted probabilities of the two <None> tokens to determine the existence of a causal relation. Specifically, if $p_c^{\texttt{<N>}} + p_e^{\texttt{<N>}} > \tau$, it suggests that the two input events are not with a causal relation, where τ is the decision threshold. Otherwise, we accept the existence of a causal relation between the two events and further identify their causality direction.

For causal event pairs, we use the predicted probabilities of the virtual event tokens <E1> and <E2> for identifying causality directions. Here, $p_c^{\texttt{<E1>}}$ and $p_c^{\texttt{<E2>}}$ are the probabilities of <E1> and <E2>, respectively, being the cause event, and $p_e^{\texttt{<E1>}}$ and $p_e^{\texttt{<E2>}}$ are the probabilities as the effect event. Specifically, if $p_c^{\texttt{<E1>}} + p_e^{\texttt{<E2>}} > p_c^{\texttt{<E2>}} + p_e^{\texttt{<E1>}}$, suggesting that <E1> is more likely to be the cause and <E2> the effect, we acknowledge the causality direction as "<E1> *causes* <E2>". Conversely, if $p_c^{\texttt{<E1>}} + p_e^{\texttt{<E2>}} < p_c^{\texttt{<E2>}} + p_e^{\texttt{<E1>}}$, the direction is "<E2> *causes* <E1>".

3 Experiments

We conduct experiments on the widely used EventStoryLine (ESC) and MAVEN datasets, and adopt Precision, Recall, and F1-score as the evaluation metrics. We compare our proposed method with the following competitors: (1) PLMs BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) use the concatenation of two events' contextual representations to identify directed event causalities. We fine-tune these two advanced PLMs and conduct causal relation and direction classification using an MLP. (2) ERGO (Chen et al., 2022) and SENDIR (Yuan et al., 2023) build relational graphs and construct reasoning chains for undirected event causality identification, respectively. We modify these two SOTA causality existence identification models to conduct causality direction identification. (3) Large-scale Language Model (LLMs) GPT3.5-Turbo (Gao et al., 2023) and ChatGLM3-6B (Zeng et al., 2024) design input prompts to conduct zero-shot event causality identification using their official APIs [1]. The details about the datasets, and experiment settings can be found in Appendix A and Appendix B.

Overall Results: Table 1 compares the overall results between our proposed method and the competitors. We can first observe that neither

[1] http://platform.openai.com, http://open.bigmodel.cn

ChatGLM3-6B nor GPT3.5-Turbo, which employ zero-shot learning based on LLMs, can outperform the other competitors that are fine-tuned based on minor PLMs. This indicates that although the LLMs are capable of reasoning and answering questions after being pre-trained with advanced techniques and large-scale datasets, they are still not effective for direct application in event causality identification. The second observation is that both ERGO and SENDIR outperform the pure fine-tuned PLMs BERT and RoBERTa. This, however, is not unexpected. As they are built upon a PLM with a well-designed neural network to identify event causalities based on relational graphs or reasoning chains. Nevertheless, they require both expert intelligence and intensive efforts to construct sophisticated neural models.

Finally, our proposed method achieves the best performance in both causality existence and direction identification on the two datasets. This validates the effectiveness of our design objective, that is, identifying directional event causalities by evaluating the statements generated by a GLM. It does not necessitate constructing an elaborate neural network for event representation learning, nor does it require selecting causal answer words and designing prompt templates for PLM-based prompt learning. By simply evaluating the statement generated from a GLM, it can achieve significant performance improvement in event causality identification.

Prompt Ablations: To validate the superiority of our proposed instructively generation and statement evaluation method, we compare it with a conventional prompt learning model, denoted as Prompt. It also uses the event description with virtual tokens <E1> and <E2> as the GLM input and employs a widely used masked prompt template, i.e., "*In this sentence, <E1> <mask> <E2>.*" to identify causal relations. The inserted <mask> token is used to predict one of the virtual answer words in generation, viz. "*cause*", "*caused by*", or "*none*", which is then mapped into a causal relation or a none relation.

Fig. 2 compares the performance of Our method and Prompt on both the ESC and MAVEN datasets for both causality existence and direction identification. We can observe that Our method has achieved significant performance improvement than the Prompt in all ablation experiments. This validates the effectiveness of our proposed method, which evaluates the instructively generated statement for relation determination, rather than the conventional prompt learning of using a <mask> token to predict an answer word and map it to a relation.

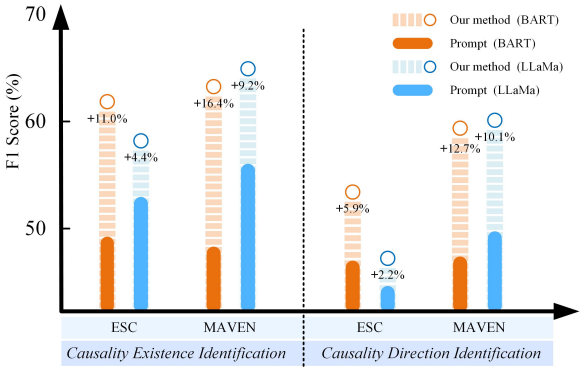

Figure 2: Performance of prompt ablation studies.

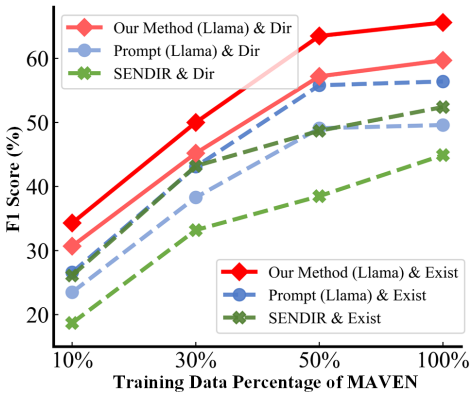

Figure 3: Performance in low-resource scenario.

Low-resource Learning: To examine the performance of our proposed method in low-resource scenario, we conduct experiments on down-sampled ESC and MAVEN datasets with fewer training set; While the development set and test set remain unchanged. The best-performing competitor SENDIR and the conventional prompt model Prompt are used for comparison under the same dataset scale. We randomly down-sample the full training set to construct subsets containing {100%, 50%, 30%, 10%} instances of the full training set. Fig. 3 summarizes the results of both causality existence and direction identification in low-resource scenario on the MAVEN dataset.

It is not unexpected that both Our method and the competitors SENDIR and Prompt suffer from the reduction of training data. However, we can observe that Our method also achieves the best performance in both causality existence and direction identification with fewer training data. This might be attributed to the outstanding capability

of the pre-trained GLM for generating causality statements. Even with fewer training data, it can still generate reliable statements, and Our method identifies event causality by evaluating these generated statements.

4 Conclusion

In this paper, we propose to evaluate the generated statements by a GLM for both causality existence and direction identification. We propose an *Instructive Generation and Statement Evaluation* method to identify directional causalities. Our proposed method first fine-tunes a GLM to instructively generate causality statements and then evaluates the rationality of the generated statements for event causality identification. Experiments on the ESC and MAVEN datasets have validated that our method can significantly outperform state-of-the-art algorithms, even with fewer training data.

Limitations

• Considering the input length limitation of the GLM, we only input sentence-level event descriptions to the GLM for causality statement generation, but this lacks document-level semantics and information.
• Fine-tuning GLMs is computationally demanding; therefore, we only use two smaller-scale GLMs, BART and LLaMA, in our experiments, and we abandon the use of the most advanced large-scale language models, such as GPT4 and Gemini.
• Although our proposed method outperforms the competitors in a low-resource scenario, its performance gap compared to full training data fine-tuning is still evident.

Acknowledgements

This work is supported in part by The Hubei Provincial Natural Science Foundation of China (Grant No: 2025AFD762) and The National Natural Science Foundation of China (Grant No: 62172167).

References

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*, pages 77–86, Vancouver, Canada.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: event relational graph transformer for document-level event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128, Gyeongju, Korea.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, MN, USA.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1808–1817, Minneapolis, MN, USA.

Sebastian Goodman, Nan Ding, and Radu Soricut. 2020. Teaforn: Teacher-forcing with n-grams. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 8704–8717, Online.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *The 29th ACM International Conference on Information and Knowledge Management*, pages 3023–3030, Virtual.

Peixin Huang, Xiang Zhao, Minghao Hu, Zhen Tan, and Weidong Xiao. 2024. Distill, fuse, pre-train: Towards effective event causality identification with commonsense-aware pre-trained model. In *Proceedings of the 31th International Conference on Computational Linguistics*, pages 5029–5040, Torino, Italy.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 7871–7880, Virtual.

Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023. Open-domain hierarchical event schema induction by incremental prompting and verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, page 5677–5697, Toronto, Canada.

Jian Liu, Yubo Chen, and Jun Zhao. 2021. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3608–3614, Virtual, Japan.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint*, arXiv:1907.11692(1):1–13.

Zhenyu Liu, Baotian Hu, Zhenran Xu, and Min Zhang. 2023. PPAT: progressive graph pairwise attention network for event causality identification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5150–5158, Macao, China.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, pages 1–18, New Orleans.

Hieu Man, Chien Van Nguyen, Nghia Trung Ngo, Linh Ngo, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Hierarchical selection of important context for generative event causality identification with optimal transports. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics*, pages 8122–8132, Torino, Italy.

Shirong Shen, Heng Zhou, Tongtong Wu, and Guilin Qi. 2022. Event causality identification via derivative prompt joint learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2288–2299, Gyeongju, Republic of Korea.

Zhengwei Tao, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Chengfeng Dou, Yongqiang Zhao, Fang Wang, and Chongyang Tao. 2023. Seag: Structure-aware event causality generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4631–4644, Toronto, Canada.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates.

Changsen Yuan, Heyan Huang, Yixin Cao, and Yonggang Wen. 2023. Discriminative reasoning with sparse event representation for document-level event-event relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, page 16222–16234, Toronto, Canada.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, and et al. 2024. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint*, arXiv:2406.12793(1):1–19.

Pengpeng Zhou, Bin Wu, Caiyong Wang, Hao Peng, Juwei Yue, and Song Xiao. 2022. What happens next? combining enhanced multilevel script learning and dual fusion strategies for script event prediction. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 10001–10040, Virtual.

A Experimental Settings

Datasets: We conduct experiments on the widely used EventStoryLine (ESC) dataset (Caselli and Vossen, 2017) and the MAVEN dataset (Wang et al., 2022). The ESC contains 22 topics, 258 documents, 5,334 event mentions, and 5,625 causal event pairs. Following (Gao et al., 2019), we use the last two topics as the development set and conduct 5-fold cross-validation on the remaining 20 topics. The MAVEN contains 4,480 documents, 103,193 event mentions, and 57,992 causal event pairs. As it does not release the test set, following (Tao et al., 2023), we use the original development set as the test set and sample 10% of the data from the training set to form the development set.

Settings: Our method is implemented based on the pre-trained Generative Language Models, namely BART-base (Lewis et al., 2020) and Llama-160M (Liu et al., 2019), and executed using the PyTorch framework with CUDA acceleration on an NVIDIA GTX 3090 GPU. We optimize our model using AdamW and set the learning rate l_{tr} for the GLM to 5e-6, the weight decay to 1e-2 respectively. The batch size is set to 16 and the decision threshold τ is set to 1.0. All trainable parameters are randomly initialized from normal distributions. All models have undergone parameter tuning to select the best-performing parameters as the final results.

B LLM Prompt

Figure 4 illustrates the demonstration of the LLM reasoning process in zero-shot settings. We employ a two-stage query approach to evaluate the directional causality identification performance of the LLM on the ESC and MAVEN datasets. For causality existence identification, we concatenate the event sentences as the input to the LLM. Notably, if the sentences for two events are identical, we input only one of them to avoid redundancy.

Then we query the LLM: Is there a causal relationship between "*marked*" and "*link*", and identify event causality based on the LLM's output. Specifically, we organize the LLM's input using the following template.

> **Input**: {*Sentences*}
> **Question**: Is there a causal relationship between "marked" and "link" ? Let's think step by step. Answer yes or no first. And give the corresponding reasons.
> **Answer**: Yes, {*Response-1*}

In the causality direction identification, we again concatenate the event sentences as the input, and additionally include the LLM's inference process from the previous stage as part of the input. Subsequently, we further query the LLM to determine the specific directionality of the causal event pair. We organize the LLM's input using the following template.

> **Input**: {*Sentences*} {*Response-1*}
> **Question**: According to the *Response-1*, is the causal relationship between "*marked*" and "*link*" from "*marked*" to "*link*" ? Let's think step by step. Answer yes or no first. And give the corresponding reasons.
> **Answer**: Yes, {*Response-2*}

C Low-resource Learning

Fig. 5 summarizes the results of both causality existence and direction identification in low-resource scenario on the ESC dataset.

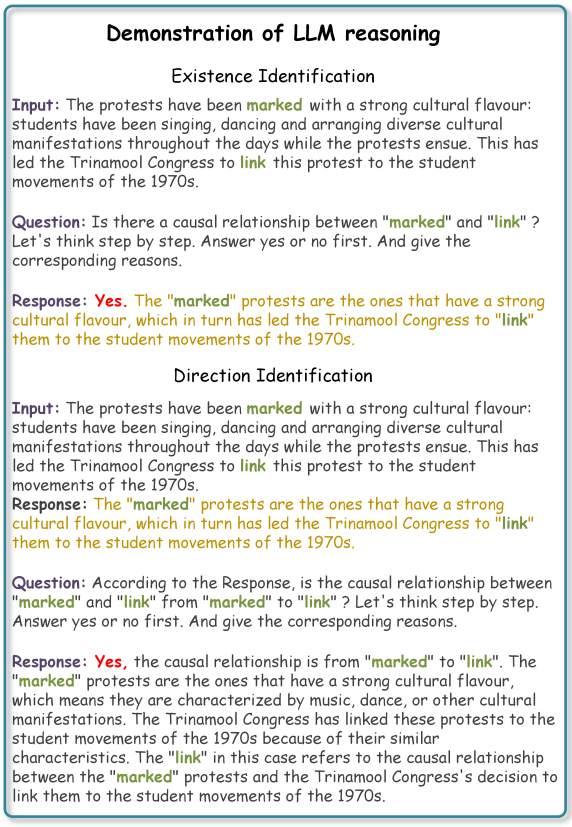

Figure 4: A demonstration of LLM reasoning ECI task.

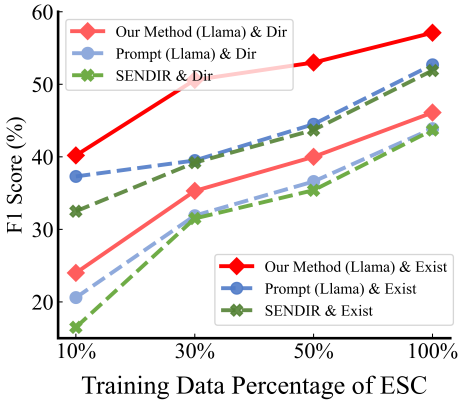

Figure 5: Performance in low-resource scenario.