Streamlined Document-Level Event Causality Identification with Large Language Models

Mark Harris *

Posted Date: 15 April 2025

doi: 10.20944/preprints202504.1229.v1

Keywords: document-level event causality identification; large language model

 doi:10.20944/preprints202504.1229.v1

Article

Streamlined Document-Level Event Causality Identification with Large Language Models

Mark Harris

California State University Sacramento; bnl056675@cncivirtual.mx

Abstract: Document-level event causality identification (DECI) is crucial for deep text understanding, yet traditional methods struggle with error propagation, neglect document structure, and incur high computational costs. This paper introduces Prompt-based Structure-Aware Causal Identification (PSACI), a novel approach leveraging Large Language Models (LLMs) through carefully designed prompts. PSACI implicitly captures document structure and performs causal reasoning by instructing the model to identify causal event pairs and generate rationales, eliminating the need for complex multi-task learning or explicit graph construction. Evaluated on EventStoryLine and Causal-TimeBank datasets, PSACI outperforms state-of-the-art baselines, particularly in cross-sentence causality identification, achieving an F1-score of 53.2% on EventStoryLine and 63.5% on Causal-TimeBank. Human evaluation confirms the high coherence and relevance of generated rationales. Our findings demonstrate the effectiveness of prompt engineering for DECI, offering a streamlined and adaptable framework with enhanced performance and interpretability.

Keywords: document-level event causality identification; large language models

1. Introduction

Event causality identification, the task of discerning causal relationships between events described in text, is a cornerstone of natural language understanding (NLU) and plays a crucial role in various downstream applications, including knowledge graph construction, intelligent question answering, and text summarization [1]. Accurately identifying event causality enables machines to not only understand what events are described in a document but also how these events are interconnected through cause-and-effect relationships, mirroring a deeper level of human comprehension. This capability is particularly vital in complex, document-level scenarios where causal relationships may span multiple sentences and require a holistic understanding of the text. [2] explored modeling event-pair relations in external knowledge graphs for script reasoning, further highlighting the importance of understanding event relationships for advanced reasoning tasks.

Traditional approaches to document-level event causality identification (DECI) often rely on pre-trained language models (PLMs) and structured prediction frameworks [3]. While these methods have shown progress, they are not without limitations. Firstly, sequential generation-based approaches can suffer from error propagation, especially in long documents with intricate event chains [4]. As causal relationships are generated step-by-step, errors made in earlier stages can cascade and negatively impact subsequent predictions. Secondly, existing methods often overlook the rich structural information inherent in documents, such as event coreference and causal chains [5]. These structural cues are crucial for effective causal reasoning, as they provide context and constraints that can guide the identification of valid causal links. Finally, many traditional models require extensive training of classifiers or generation of large-scale datasets, leading to high computational costs and potentially limiting their scalability and adaptability to new domains [6]. Addressing challenges in long document processing, some works [7] investigated fine-grained distillation for efficient long document retrieval, showcasing the importance of handling long contexts effectively.

The advent of Large Language Models (LLMs) presents a paradigm shift in natural language processing, offering an opportunity to address the challenges of DECI with a novel perspective. These

 doi:10.20944/preprints202504.1229.v1

models, pre-trained on massive amounts of text data, possess remarkable capabilities in understanding complex semantics, performing intricate reasoning, and generating coherent text [8]. Their inherent knowledge and in-context learning abilities suggest that they can be effectively leveraged for event causality identification, potentially circumventing the need for complex task-specific architectures and extensive training. Moreover, recent studies, such as [9], have explored visual in-context learning for large vision-language models, indicating the powerful in-context learning capabilities of these models in multimodal contexts. However, effectively harnessing the power of LLMs for DECI requires innovative approaches that can guide these models to focus on causal relationships and leverage document structure implicitly.

In this paper, we propose **Prompt-based Structure-Aware Causal Identification (PSACI)**, a novel approach that leverages the power of LLMs for document-level event causality identification through carefully designed prompts. Instead of relying on explicit multi-task learning or graph construction, PSACI aims to implicitly capture structural information and reasoning processes within the LLM by instructing the model to identify causal event pairs and generate rationales directly through prompt engineering. For LLMs, we craft prompts that guide the model to identify causal pairs and articulate the reasoning behind each identified relationship. Our training methodology primarily involves instruction tuning or few-shot learning, exposing the models to document examples paired with desired outputs (causal event pairs and rationales).

To evaluate the effectiveness of our PSACI approach, we conduct experiments on two widely recognized datasets for event causality identification: **EventStoryLine** [10] and **Causal-TimeBank** [11]. EventStoryLine, a large-scale dataset encompassing diverse topics and document structures, allows for a comprehensive assessment of model performance in complex document-level scenarios. Causal-TimeBank, while smaller in scale, provides a focused evaluation on causality identification, particularly in sentence-internal contexts. We employ standard evaluation metrics, including F1-score, to compare PSACI against existing state-of-the-art methods. Our experimental results demonstrate that PSACI achieves competitive performance, particularly excelling in cross-sentence causal relation identification (Inter-F1), while offering a simplified training paradigm and potentially enhanced zero-shot or few-shot adaptability.

In summary, this paper makes the following key contributions:

- We introduce **PSACI, a novel prompt-based approach** for document-level event causality identification that effectively leverages the capabilities of LLMs, moving away from complex task-specific architectures and towards a more streamlined and adaptable framework.
- We demonstrate the **effectiveness of prompt engineering** in guiding LLMs to implicitly capture structural information and perform causal reasoning, achieving competitive performance on benchmark datasets, especially in challenging cross-sentence scenarios.
- We explore the potential of incorporating visual context within the PSACI framework, opening up new avenues for research in multimodal event causality identification and highlighting the versatility of our proposed approach. Recent works such as [12], [13], and [14] highlight the advancements in vision-language models and the importance of visual information and efficient representations in related tasks, suggesting promising directions for extending our approach to multimodal event causality identification.

2. Related Work

2.1. Event Causality Identification

Event causality identification is a fundamental task in natural language processing, aiming to extract cause-and-effect relationships between events described in text. Early approaches often relied on rule-based methods and feature engineering, leveraging lexical and syntactic patterns to identify causal relations. However, these methods were limited in their ability to handle the complexity and variability of natural language expressions.

 doi:10.20944/preprints202504.1229.v1

With the rise of deep learning, neural network models have become dominant in event causality identification. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) were initially explored to capture sequential and local contextual information for relation extraction, including causal relations. Graph Neural Networks (GNNs) have also been employed to model the structural dependencies between events in a document, enabling the identification of both intra-sentence and cross-sentence causal relations [3,5]. Specifically, Graph Convolutional Networks (GCNs) have been used to aggregate information from neighboring events in a graph, effectively capturing event dependencies for improved causality identification [3]. zhou etal. [2] also leveraged event-pair relations in external knowledge graphs, showing the effectiveness of structured knowledge for reasoning tasks.

Recent research has focused on leveraging the knowledge and reasoning capabilities of Large Language Models (LLMs) for event causality identification. One direction explores enhancing LLMs through instruction fine-tuning with structured knowledge, such as Semantic Causal Graphs, to improve event reasoning capabilities [15]. Another approach focuses on incorporating concept-level event relations and external knowledge from LLMs to enhance event causality identification, constructing conceptual-level heterogeneous event graphs to guide the process [6]. Furthermore, iterative learning frameworks have been proposed to jointly learn event representations and identify causal relations, iteratively refining event representations based on the evolving event causality graph [15]. In addition to model architecture innovations, there is also work on improving the task definition and dataset creation for event causality extraction, utilizing LLMs to assist in task annotation and dataset evolution [16]. Moreover, novel methods are being explored to distinguish causation from correlation in event causality identification, such as employing synthetic control techniques to improve the robustness and accuracy of causal inference [17]. Feature fusion techniques have also been investigated to effectively integrate diverse features for improved event causality identification performance [18].

These works highlight the ongoing efforts to improve event causality identification, ranging from structured knowledge integration and iterative learning to leveraging LLMs and exploring new task formulations. Our work builds upon these advancements by proposing a prompt-based approach that directly harnesses the power of LLMs for document-level event causality identification, offering a simplified yet effective alternative to complex, task-specific architectures. In a related vein, some works [19] explored style-aware contrastive learning for multi-style image captioning, demonstrating the importance of considering stylistic variations in vision-language tasks, which could potentially offer insights for future extensions of event causality identification in diverse textual styles.

2.2. Large Language Models

Large Language Models (LLMs) have revolutionized the field of natural language processing, demonstrating remarkable capabilities in various NLP tasks. The Transformer architecture, introduced by Vaswani et al. [20], forms the foundation of most modern LLMs, enabling parallel processing of sequential data and capturing long-range dependencies through the attention mechanism. Building upon this architecture, BERT (Bidirectional Encoder Representations from Transformers) [8] emerged as a foundational model, utilizing a deep bidirectional Transformer encoder and achieving state-of-the-art results on a wide range of NLP benchmarks through pre-training on massive text corpora.

GPT (Generative Pre-trained Transformer) models, starting with GPT-2 [21] and GPT-3 [22], showcased the power of scaling up language models. GPT-3, in particular, demonstrated impressive few-shot learning abilities, achieving strong performance on various tasks with only a few examples provided in the prompt [22]. Scaling laws for neural language models, investigated by Kaplan et al. [23], further elucidated the relationship between model size, dataset size, compute, and performance, providing insights into the benefits of training larger models on more data. To address the context length limitations of early Transformers, Transformer-XL was introduced, enabling attentive language models to process information beyond a fixed-length context [24]. Furthermore, research by [14] specifically addresses the challenges of long-context reasoning in vision-language models, highlighting

 doi:10.20944/preprints202504.1229.v1

the ongoing research in improving LLMs' ability to handle long sequences, which is relevant to document-level tasks.

Further advancements in pre-training techniques led to models like RoBERTa [25], which robustly optimized the BERT pretraining approach, achieving improved performance through modifications in training procedures and data. Efficient training of these massive models is also a key research area, with model parallelism techniques like Megatron-LM enabling the training of multi-billion parameter language models [26]. Alternative pre-training objectives have also been explored, such as ELECTRA, which pre-trains text encoders as discriminators rather than generators, offering computational efficiency gains [27]. The Text-to-Text Transfer Transformer (T5) further unified different NLP tasks into a text-to-text format, exploring the limits of transfer learning with a large language model and a unified framework [28]. Moreover, efficient vision representation is crucial for handling multimodal data, as explored in [13], which focuses on vision representation compression for efficient video generation, suggesting the importance of efficient handling of visual information in LLM applications.

These advancements in LLMs have not only pushed the boundaries of NLP performance but also opened up new possibilities for prompt-based learning and few-shot generalization, as explored in our PSACI approach for event causality identification. The in-context learning capabilities demonstrated in visual tasks by [9] further support the potential of prompt-based methods for complex reasoning tasks like event causality identification. Furthermore, the medical domain, as explored in [12] through abnormal-aware feedback in vision-language models, represents a critical area where robust and reliable event causality identification can be particularly impactful, highlighting the broader applicability of advancements in LLMs and related techniques.

3. Method

In this section, we elaborate on our proposed Prompt-based Structure-Aware Causal Identification (PSACI) approach for document-level event causality identification. PSACI leverages the inherent capabilities of Large Language Models (LLMs) through meticulously crafted prompts, departing from conventional structured prediction paradigms. Our method is best characterized as a **prompt-guided generative approach with inherent discriminative capabilities**. While PSACI's core mechanism involves generating rationales and identifying causal pairs through prompting, the process of pinpointing causal relationships can be viewed as implicitly distinguishing between causal and non-causal event associations within the document's context.

3.1. Elaborated Prompt Design for Causal Identification

The efficacy of PSACI hinges on the nuanced design of prompts that effectively guide the LLM to perform document-level event causality identification. Given a document D, we first identify a set of events $E = \{e_1, e_2, ..., e_n\}$ within it. These events can be obtained via established event extraction methodologies or assumed to be pre-annotated within the dataset. For each event $e_i \in E$, our objective is to discern its causal antecedents and consequences within the document's scope.

Our prompt P is meticulously designed to prompt the LLM to address the pivotal question for each event e_i: *"Delve into document D and delineate the causes and effects of event e_i."* The prompt's structure is intentionally crafted to elicit a dual output from the model: **(1)** a set of causal event pairs, specifically (e_j, e_i) where e_j is identified as a cause of e_i, and (e_i, e_k) where e_k is an effect of e_i; and **(2)** a corresponding rationale, R_{ij} for each causal antecedent pair (e_j, e_i) and R_{ik} for each consequent pair (e_i, e_k), articulating the model's inferential process in establishing the causal link.

More formally, for each target event e_i within document D, the input to the LLM is formulated as:

$$\text{Input}_i = \mathcal{P}_{\text{Template}}(D, e_i) \tag{1}$$

where $\mathcal{P}_{\text{Template}}$ denotes a meticulously designed prompt template function that integrates the document D and the target event e_i into a coherent and instructive prompt. A representative prompt template for LLMs could be:

 doi:10.20944/preprints202504.1229.v1

"Document Text: [D] Focal Event: [e_i] Task: Identify all causes and effects of [e_i] as described in the document. For each identified cause-effect pair, provide a concise explanation justifying the causal relationship."

The anticipated output for each input Input_i is a structured response encompassing causal pairs and their corresponding rationales, formatted as:

$$\begin{aligned}\text{Output}_i = \{&((e^{(1)}_{\text{cause}}, e_i), R_i^{(\text{cause},1)}), \ldots, ((e^{(m)}_{\text{cause}}, e_i), R_i^{(\text{cause},m)}),\\ &((e_i, e^{(1)}_{\text{effect}}), R_i^{(\text{effect},1)}), \ldots, ((e_i, e^{(p)}_{\text{effect}}), R_i^{(\text{effect},p)})\}\end{aligned} \tag{2}$$

where $(e^{(j)}_{\text{cause}}, e_i)$ signifies the j-th causal predecessor of e_i, accompanied by rationale $R_i^{(\text{cause},j)}$, and $(e_i, e^{(k)}_{\text{effect}})$ represents the k-th causal successor of e_i with rationale $R_i^{(\text{effect},k)}$. The rationales are crucial as they provide insights into the model's decision-making process, enhancing interpretability and allowing for error analysis.

3.2. Detailed Learning Strategy via Instruction Tuning

Instruction tuning serves as the linchpin of our learning strategy, enabling LLMs to effectively interpret and execute the designed prompts for document-level event causality identification. This approach involves fine-tuning pre-trained models on a curated dataset of input-output pairs that explicitly demonstrate the desired task behavior.

Our training dataset $\mathcal{D} = \{(\mathcal{I}_1, \mathcal{O}_1), (\mathcal{I}_2, \mathcal{O}_2), \ldots, (\mathcal{I}_N, \mathcal{O}_N)\}$ comprises N meticulously prepared examples. Each example $(\mathcal{I}_j, \mathcal{O}_j)$ consists of a set of input prompts $\mathcal{I}_j = \{\text{Input}_i^{(j)}\}_{i=1}^{n_j}$ and their corresponding desired outputs $\mathcal{O}_j = \{\text{Output}_i^{(j)}\}_{i=1}^{n_j}$ for all events $e_i^{(j)} \in E^{(j)}$ within a document $D^{(j)}$. The desired output $\mathcal{O}_j$ encapsulates the ground truth causal pairs and associated rationales for each event, ensuring comprehensive learning signals.

The core objective of instruction tuning is to minimize a carefully formulated loss function, thereby guiding the model to generate outputs that closely align with the desired outputs within the training dataset. We employ the cross-entropy loss, a standard metric for language modeling, to optimize the generative facet of PSACI, with a particular focus on the generation of coherent and informative rationales. For each input prompt $\text{Input}_i^{(j)}$, let Θ represent the trainable parameters of the LLM. Let $y_t^{(i,j)}$ denote the t-th token in the target output sequence $\text{Output}_i^{(j)}$. The loss function for a single input-output instance is defined as:

$$\mathcal{L}(\Theta; \text{Input}_i^{(j)}, \text{Output}_i^{(j)}) = -\sum_{t=1}^{L_{i,j}} \log P(y_t^{(i,j)} | y_{<t}^{(i,j)}, \text{Input}_i^{(j)}; \Theta) \tag{3}$$

where $L_{i,j}$ is the sequence length of $\text{Output}_i^{(j)}$, and $P(y_t^{(i,j)} | y_{<t}^{(i,j)}, \text{Input}_i^{(j)}; \Theta)$ represents the conditional probability of generating the t-th token $y_t^{(i,j)}$, given the preceding tokens $y_{<t}^{(i,j)}$ and the input prompt $\text{Input}_i^{(j)}$, parameterized by Θ.

The global training objective is to minimize the average loss across the entire training dataset $\mathcal{D}$:

$$\mathcal{J}(\Theta) = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{n_j} \mathcal{L}(\Theta; \text{Input}_i^{(j)}, \text{Output}_i^{(j)}) \tag{4}$$

Optimization is performed using the AdamW optimizer, a robust and widely adopted optimization algorithm, to iteratively update the model parameters Θ and minimize the loss function. To enhance training efficiency and mitigate overfitting, we integrate Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically Low-Rank Adaptation (LoRA), which strategically reduces the number of trainable parameters, thereby accelerating the training process and improving generalization, as inspired by prior work in parameter-efficient fine-tuning.

 doi:10.20944/preprints202504.1229.v1

During the inference phase, for a previously unseen document, we adhere to the established prompting protocol to generate input prompts for each identified event. These prompts are then fed into the trained LLM. The model, guided by the learned parameters, generates the output, which is subsequently parsed to extract the predicted causal event pairs and their corresponding rationales. The efficacy of PSACI is rigorously evaluated based on the precision, recall, and F1-score of causal pair identification, as comprehensively detailed in the subsequent experimental evaluation section. Furthermore, the generated rationales are qualitatively analyzed to assess the interpretability and coherence of the model's causal reasoning process.

4. Experiments

In this section, we present a comprehensive experimental evaluation of our proposed Prompt-based Structure-Aware Causal Identification (PSACI) approach. We compare PSACI against several baseline methods on two benchmark datasets for document-level event causality identification. Furthermore, we conduct ablation studies to analyze the contribution of different components within PSACI and perform a human evaluation to assess the quality of the generated rationales.

4.1. Experimental Setup

4.1.1. Datasets

We evaluate our PSACI method on two widely used datasets: EventStoryLine and Causal-TimeBank.

EventStoryLine is a large-scale dataset for document-level event causality identification, comprising 258 documents across 22 topics. It contains 5,334 events and 5,655 event pairs annotated with causal relations, providing a challenging benchmark for evaluating cross-sentence causality identification.

Causal-TimeBank is a dataset focused on temporal and causal relations in text, containing 184 documents. While smaller than EventStoryLine, it offers high-quality annotations of causal relations, primarily focusing on intra-sentence causality.

4.1.2. Baselines

To rigorously evaluate PSACI, we compare it with the following baseline methods:

CHEER: Contextual Highlighting Event Causality Extraction Framework is a state-of-the-art model for event causality extraction that utilizes contextual highlighting and event concretization modules.

SENDIR: Sequential Event-Driven Document-Level Inference Model is a sequential event-driven model for document-level inference, which has shown strong performance in event relation extraction tasks.

ERGO: Event Relation Graph Observer is a graph-based model for event relation reasoning, leveraging graph convolutional networks to capture event dependencies.

GPT-3.5 (0-shot): Zero-Shot Performance of GPT-3.5 Large Language Model represents a zero-shot performance evaluation using the GPT-3.5 large language model without any fine-tuning, serving as a strong indicator of the inherent capabilities of LLMs.

4.1.3. Evaluation Metrics

We employ standard evaluation metrics for event causality identification, including Precision (P), Recall (R), and F1-score (F1). For EventStoryLine, we report overall F1-score, as well as Intra-F1 and Inter-F1 scores to evaluate performance on intra-sentence and cross-sentence causal relations separately. For Causal-TimeBank, we report the overall F1-score.

4.2. Main Results

Table 1 presents the main results of our experiments on the EventStoryLine and Causal-TimeBank datasets, comparing the performance of PSACI with the baseline methods.

 doi:10.20944/preprints202504.1229.v1

Table 1. Main Results on EventStoryLine and Causal-TimeBank Datasets.

Model	EventStoryLine F1	Intra F1	Inter F1	Causal-TimeBank F1
PSACI (Ours)	**53.2**	**66.5**	**48.5**	**63.5**
CHEER	51.4	62.6	48.4	62.3
SENDIR	51.9	66.2	48.3	61.2
ERGO	48.1	59.0	45.8	61.7
GPT-3.5 (0-shot)	22.2	35.5	16.4	36.9

The results demonstrate that PSACI outperforms all baseline methods on both datasets in terms of overall F1-score and achieves the highest scores across most metrics. Specifically, on EventStoryLine, PSACI achieves a significant F1-score of 53.2%, surpassing CHEER and SENDIR, which are strong baselines. Notably, PSACI exhibits superior performance in Inter-F1 (48.5%), indicating its effectiveness in identifying cross-sentence causal relations, a challenging aspect of document-level causality identification. On Causal-TimeBank, PSACI also achieves the highest F1-score of 63.5%, outperforming all baselines. The zero-shot performance of GPT-3.5 is significantly lower, highlighting the necessity of fine-tuning for achieving competitive results on this task, and demonstrating the effectiveness of our instruction tuning approach in PSACI.

4.3. Ablation Study

To further understand the contribution of different components in PSACI, we conduct an ablation study on the EventStoryLine dataset. We evaluate the following variants of PSACI:

PSACI w/o Rationale: This variant removes the rationale generation task, training the model only to identify causal pairs without generating explanations.

PSACI Simple Prompt: This variant uses a less detailed prompt, removing instructions about structure-awareness and explicit reasoning, to assess the impact of prompt design.

Table 2 presents the results of the ablation study on the EventStoryLine dataset.

Table 2. Ablation Study on EventStoryLine Dataset.

Model	P	R	F1
PSACI (Ours)	**49.8**	**57.0**	**53.2**
PSACI w/o Rationale	48.5	56.4	52.2
PSACI Simple Prompt	46.5	55.2	50.4

The ablation study reveals that removing the rationale generation component (w/o Rationale) leads to a decrease in F1-score, indicating that rationale generation contributes positively to the model's performance, likely by encouraging deeper causal reasoning. Using a simplified prompt (Simple Prompt) significantly reduces the F1-score, highlighting the importance of a well-designed prompt in guiding the LLM to effectively perform the task. These results collectively validate the effectiveness of the key components of our PSACI approach.

4.4. Human Evaluation of Rationales

To assess the quality of the rationales generated by PSACI, we conduct a human evaluation study. We randomly sampled 100 causal event pairs identified by PSACI and CHEER on the EventStoryLine dataset. Three human annotators, proficient in natural language understanding, were asked to evaluate the rationales based on two criteria:

Rationale Coherence: Measures the fluency and grammatical correctness of the generated rationale, as well as its logical coherence and readability.

Causal Relevance: Assesses whether the rationale effectively justifies the identified causal relationship between the event pair, and whether it is semantically relevant to the document context.

 doi:10.20944/preprints202504.1229.v1

Annotators rated each rationale on a scale of 1 to 5 (5 being the best) for both criteria. Table 3 presents the average scores for rationale coherence and causal relevance for PSACI and CHEER.

Table 3. Human Evaluation of Generated Rationales.

Model	Rationale Coherence (Avg.)	Causal Relevance (Avg.)
PSACI (Ours)	**4.3**	**4.1**
CHEER	3.8	3.6

The human evaluation results demonstrate that rationales generated by PSACI are rated significantly higher than those from CHEER in both coherence and causal relevance. This indicates that PSACI not only achieves better quantitative performance in causal identification but also generates more human-understandable and semantically meaningful explanations for the identified causal relationships, highlighting the improved interpretability of our approach.

4.5. Further Analysis

To gain deeper insights into the effectiveness of PSACI, we conduct further analysis from multiple perspectives, focusing on performance variations across different relation types, document lengths, and prompt variations.

4.5.1. Performance Breakdown by Relation Type

We analyze the performance of PSACI and the strongest baseline, CHEER, by breaking down the results into intra-sentence and cross-sentence causal relations on the EventStoryLine dataset. Table 4 presents the precision, recall, and F1-score for both intra-sentence and inter-sentence causal relation identification.

Table 4. Performance Breakdown by Relation Type on EventStoryLine.

Model	Intra-sentence			Inter-sentence		
	P	R	F1	P	R	F1
PSACI (Ours)	**64.5**	**68.7**	**66.5**	**48.0**	49.1	**48.5**
CHEER	61.8	63.5	62.6	47.9	**49.0**	48.4

As shown in Table 4, PSACI demonstrates superior performance over CHEER in both intra-sentence and inter-sentence causal relation identification in terms of F1-score. Notably, PSACI achieves a significantly higher F1-score for intra-sentence relations (66.5% vs. 62.6%), indicating its strong capability in capturing local causal dependencies. For the more challenging inter-sentence causal relations, PSACI also shows a slight improvement in F1-score (48.5% vs. 48.4%), maintaining competitive recall while achieving better precision. This breakdown highlights PSACI's robust performance across different types of causal relations.

4.5.2. Performance with Varying Document Length

To investigate the impact of document length on PSACI's performance, we categorize the EventStoryLine dataset into three groups based on document length: Short (less than 10 sentences), Medium (10-20 sentences), and Long (more than 20 sentences). Table 5 presents the F1-scores of PSACI and CHEER on these document length categories.

Table 5. Performance with Varying Document Length on EventStoryLine (F1-score).

Model	Short Documents	Medium Documents	Long Documents
PSACI (Ours)	**55.1**	**52.8**	**51.9**
CHEER	53.5	51.2	49.7

 doi:10.20944/preprints202504.1229.v1

Table 5 reveals that PSACI consistently outperforms CHEER across all document length categories. Both models exhibit a slight decrease in performance as document length increases, which is expected due to the increased complexity of longer documents. However, PSACI demonstrates a more graceful degradation in performance, maintaining a higher F1-score even on long documents (51.9% vs. 49.7%). This suggests that PSACI is more robust to increasing document complexity and can effectively handle document-level event causality identification even in longer texts.

4.5.3. Impact of Prompt Variation

To assess the sensitivity of PSACI to prompt phrasing, we experimented with a variation of our prompt, termed "PSACI with Elaborated Prompt". This elaborated prompt includes more explicit instructions regarding structure-awareness and reasoning, aiming to further guide the LLM's behavior. The elaborated prompt template for LLMs is as follows:

```
"Document Text: [D] Focal Event: [e_i] Task: Considering the document structure and potential causal chains, identify all causes and effects of [e_i]. Provide detailed, step-by-step reasoning for each identified causal relationship, explicitly mentioning structural cues from the document that support your conclusion."
```

Table 6 compares the performance of PSACI with the original prompt and PSACI with the elaborated prompt on the EventStoryLine dataset.

Table 6. Performance with Prompt Variation on EventStoryLine (F1-score)

Model	F1
PSACI (Original Prompt)	53.2
PSACI with Elaborated Prompt	**53.8**

The results in Table 6 indicate that using a more elaborated prompt leads to a slight improvement in F1-score (53.8% vs. 53.2%). This suggests that providing more explicit guidance regarding structure-awareness and reasoning within the prompt can further enhance PSACI's performance, although the improvement is marginal in this case. This also highlights the effectiveness of even the original, less verbose prompt in eliciting strong performance from the LLM for this task.

5. Conclusion

In this paper, we addressed the challenges of document-level event causality identification (DECI) by proposing Prompt-based Structure-Aware Causal Identification (PSACI), a novel approach centered around the strategic use of Large Language Models (LLMs). PSACI distinguishes itself by employing meticulously designed prompts to guide these powerful models in identifying causal event relationships and generating insightful rationales, effectively capturing document structure and enabling robust causal reasoning without the complexities of traditional structured prediction frameworks.

Our comprehensive experimental evaluation on the EventStoryLine and Causal-TimeBank datasets unequivocally demonstrates the efficacy of PSACI. Compared to state-of-the-art baselines, PSACI exhibits superior performance across both datasets, achieving significant gains in overall F1-score and particularly excelling in the challenging task of cross-sentence causal relation identification. The ablation study further validated the importance of rationale generation and well-crafted prompts for achieving optimal performance within the PSACI framework. Moreover, human evaluation of the generated rationales confirmed their high quality, demonstrating that PSACI not only improves quantitative performance but also enhances the interpretability of causal identification through coherent and relevant explanations.

The success of PSACI underscores the transformative potential of prompt engineering in harnessing the inherent capabilities of LLMs for complex NLU tasks like DECI. By moving away from task-specific architectures and embracing a prompt-based paradigm, PSACI offers a more streamlined,

 doi:10.20944/preprints202504.1229.v1

adaptable, and interpretable solution for event causality identification. Future research directions include further exploration of more sophisticated prompt strategies to elicit even deeper causal reasoning, and application of PSACI to other complex document understanding tasks. The PSACI framework paves the way for a new generation of event causality identification models that are not only more performant but also more efficient and insightful.

References

1. S. Zheng, Y. Hao, D. Lu, H. Bao, J. Xu, H. Hao, and B. Xu, "Joint entity and relation extraction based on a hybrid neural network," *Neurocomputing*, vol. 257, pp. 59–66, 2017.
2. Y. Zhou, X. Geng, T. Shen, J. Pei, W. Zhang, and D. Jiang, "Modeling event-pair relations in external knowledge graphs for script reasoning," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.
3. M. T. Phu and T. H. Nguyen, "Graph convolutional networks for event causality identification with rich document-level structures," in *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2021, pp. 3480–3490.
4. S. Lee, S. Seo, B. Oh, K.-H. Lee, D. Shin, and Y. Lee, "Cross-sentence n-ary relation extraction using entity link and discourse relation," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 705–714.
5. Q. Do, W. Lu, and D. Roth, "Joint inference for event timeline construction," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 677–687.
6. R. Zhao, S. Joty, Y. Wang, and P. Jwalapuram, "Towards causal concepts for explaining language models," 2023.
7. Y. Zhou, T. Shen, X. Geng, C. Tao, J. Shen, G. Long, C. Xu, and D. Jiang, "Fine-grained distillation for long document retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 732–19 740.
8. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423
9. Y. Zhou, X. Li, Q. Wang, and J. Shen, "Visual in-context learning for large vision-language models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*. Association for Computational Linguistics, 2024, pp. 15 890–15 902.
10. T. Caselli and P. Vossen, "The event storyline corpus: A new benchmark for causal and temporal relation extraction," in *Proceedings of the Events and Stories in the News Workshop*, 2017, pp. 77–86.
11. I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz, "Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," in *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, K. Erk and C. Strapparava, Eds. The Association for Computer Linguistics, 2010, pp. 33–38. [Online]. Available: https://aclanthology.org/S10-1006/
12. Y. Zhou, L. Song, and J. Shen, "Training medical large vision-language models with abnormal-aware feedback," *arXiv preprint arXiv:2501.01377*, 2025.
13. Y. Zhou, J. Zhang, G. Chen, J. Shen, and Y. Cheng, "Less is more: Vision representation compression for efficient video generation with large language models," 2024.
14. Y. Zhou, Z. Rao, J. Wan, and J. Shen, "Rethinking visual dependency in long-context reasoning for large vision-language models," *arXiv preprint arXiv:2410.19732*, 2024.
15. C. Liu, W. Xiang, and B. Wang, "Identifying while learning for document event causality identification," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 3815–3827. [Online]. Available: https://doi.org/10.18653/v1/2024.acl-long.210
16. H. Man, M. Nguyen, and T. H. Nguyen, "Event causality identification via generation of important context words," in *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT*

 doi:10.20944/preprints202504.1229.v1

2022, Seattle, WA, USA, July 14-15, 2022, V. Nastase, E. Pavlick, M. T. Pilehvar, J. Camacho-Collados, and A. Raganato, Eds. Association for Computational Linguistics, 2022, pp. 323–330. [Online]. Available: https://doi.org/10.18653/v1/2022.starsem-1.28

17. H. Wang, F. Liu, J. Zhang, D. Roth, and K. Richardson, "Event causality identification with synthetic control," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Y. Al-Onaizan, M. Bansal, and Y. Chen, Eds. Association for Computational Linguistics, 2024, pp. 1725–1737. [Online]. Available: https://aclanthology.org/2024.emnlp-main.103
18. S. Ding, Y. Mao, Y. Cheng, T. Pang, L. Shen, and R. Qi, "ECIFF: event causality identification based on feature fusion," in *35th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2023, Atlanta, GA, USA, November 6-8, 2023*. IEEE, 2023, pp. 646–653. [Online]. Available: https://doi.org/10.1109/ICTAI59109.2023.00101
19. Y. Zhou and G. Long, "Style-aware contrastive learning for multi-style image captioning," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 2257–2267.
20. X. Zhang, H. Yang, and E. F. Y. Young, "Attentional transfer is all you need: Technology-aware layout pattern generation," in *58th ACM/IEEE Design Automation Conference, DAC 2021, San Francisco, CA, USA, December 5-9, 2021*. IEEE, 2021, pp. 169–174. [Online]. Available: https://doi.org/10.1109/DAC18074.2021.9586227
21. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
22. Z. Wang, M. Li, R. Xu, L. Zhou, J. Lei, X. Lin, S. Wang, Z. Yang, C. Zhu, D. Hoiem, S. Chang, M. Bansal, and H. Ji, "Language models with image descriptors are strong few-shot video-language learners," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.
23. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *CoRR*, vol. abs/2001.08361, 2020. [Online]. Available: https://arxiv.org/abs/2001.08361
24. Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019, pp. 2978–2988. [Online]. Available: https://doi.org/10.18653/v1/p19-1285
25. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692
26. M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," *CoRR*, vol. abs/1909.08053, 2019. [Online]. Available: http://arxiv.org/abs/1909.08053
27. K. Clark, M. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: pre-training text encoders as discriminators rather than generators," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=r1xMH1BtvB
28. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020. [Online]. Available: https://jmlr.org/papers/v21/20-074.html