Inferring Causality Between Entities and Events from CTI Reports

CHANG-HONG JIANG, Dept. of Computer Science and Information Engineering, National Chung Cheng University, Minxiong, Taiwan
REN-HUNG HWANG*, College of Artificial Intelligence, National Yang Ming Chiao Tung University, Tainan, Taiwan and Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan
YING-DAR LIN, Computer Science and Information Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
PO-CHING LIN, Department of Computer Science and Information Engineering, National Chung Cheng University, Minxiong, Taiwan
YUAN-CHENG LAI, Information Management, National Taiwan University of Science and Technology, Taipei City, Taiwan
ERIC HSIAO-KUANG WU, National Central University, Zhongli District, Taiwan

Inferring causal relationships between entities and events is essential for accelerating root cause analysis in cybersecurity incidents. This capability is particularly critical in cyber threat intelligence, where identifying which entities trigger events and which are affected enables rapid and effective responses. However, most existing causal inference research primarily focuses on event-to-event causality, overlooking the relationships between entities and individual events. To bridge this gap, we constructed a specialized dataset designed to analyze causal relationships between entities and cyber threat events. We then proposed a causal inference model based on a BERT-based multi-layer stacked architecture (MLSA) and trained it using the constructed dataset. To assess its effectiveness, we compared its performance with that of large language models (LLMs) for causal inference. Experimental results demonstrate that the proposed MLSA model outperforms the state-of-the-art GPT-4o, achieving an average F1-score improvement of approximately 38%. Additionally, MLSA achieves a competitive F1-score in causal inference compared to GPT o3-mini while delivering an 11% improvement in effect relation inference. These findings highlight the effectiveness of our approach, introducing a novel and robust model for event-entity causality analysis in cyber threat intelligence.

CCS Concepts: • **Computing methodologies** → **Information extraction**; **Causal reasoning and diagnostics**; *Neural networks.*

Additional Key Words and Phrases: Cyber threat analysis, Causal inference, Event extraction, Natural language processing

*Corresponding author.

Authors' Contact Information: Chang-Hong Jiang, Dept. of Computer Science and Information Engineering, National Chung Cheng University, Minxiong, Taiwan; e-mail: luck3216ss@gmail.com; Ren-Hung Hwang, College of Artificial Intelligence, National Yang Ming Chiao Tung University, Tainan, Taiwan and Department of Computer Science and Information Engineering, Asia University, Taichung, Taichung City, Taiwan; e-mail: rhhwang@nycu.edu.tw; Ying-Dar Lin, Computer Science and Information Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan Province, Taiwan; e-mail: ydlin@cs.nctu.edu.tw; Po-Ching Lin, Department of Computer Science and Information Engineering, National Chung Cheng University, Minxiong, Taiwan; e-mail: pclin@cs.ccu.edu.tw; Yuan-Cheng Lai, Information Management, National Taiwan University of Science and Technology, Taipei City, Taipei City, Taiwan; e-mail: laiyc@cs.ntust.edu.tw; Eric Hsiao-Kuang Wu, National Central University, Zhongli District, Taoyuan City, Taiwan; e-mail: hsiao@csie.ncu.edu.tw.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 1557-6051/2026/5-ART
https://doi.org/10.1145/3816038

Just Accepted

1 Introduction

Event extraction (EE) has been a prominent research area since its introduction at the 3rd Message Understanding Conference (MUC) in 1991 [20], and was further advanced through integration with named entity recognition (NER) in the Automatic Content Extraction (ACE) program, as detailed by Doddington et al. [4]. This integration enabled the identification of entities and their roles within events. By combining EE and NER, users can more effectively comprehend events and the associated participants. As a critical technique for analyzing both structured and unstructured data, EE continues to evolve, with researchers proposing increasingly sophisticated models to improve its performance.

ACE2005 [22], one of the most widely cited datasets for event extraction, categorizes events into eight major types, each with several subtypes. It is commonly used as a benchmark for evaluating model performance, and recent studies frequently demonstrate improvements over earlier baselines. Beyond ACE2005, domain experts have curated publicly available datasets tailored to specific fields, such as cybersecurity, healthcare, and the legal domain, to extend the applicability of event extraction methods across diverse contexts.

In cybersecurity, EE and NER techniques enable analysts to efficiently process threat intelligence reports by identifying both the perpetrators and the targets of cybersecurity incidents. Inferring causal relationships between events and entities is crucial for industry practitioners, as it facilitates the timely identification of root causes and consequences, ultimately supporting proactive defense strategies. This need is particularly acute in Security Operations Centres (SOCs), where analysts must rapidly triage and investigate large volumes of CTI reports under time pressure. Automating the extraction of event–entity causal relationships can serve as a critical human–AI collaboration mechanism, enabling AI systems to pre-process and structure causal information from raw reports so that SOC analysts can focus their expertise on validation, prioritization, and decision-making. However, most of the existing work on event extraction has focused on improving the accuracy of event and attribute detection, with limited attention paid to associations between events and entities [19]. Our review of recent studies on causal inference [1, 6, 9, 21, 27] shows that prior research primarily addresses causality between events or sentences, rather than exploring causal links between events and entities. This gap underscores the need for more comprehensive approaches to causal inference in cybersecurity contexts.

Earlier research on relationship extraction (RE) typically followed a pipeline approach, where EE or NER was performed first, followed by RE as a separate task. However, this sequential design often led to error propagation and missed opportunities to share contextual features across subtasks. Recent advances have introduced joint models that perform EE, NER, and RE simultaneously, producing promising results [2, 3, 12, 29]. Given the inherent causal connections between events and the entities involved, we posit that a joint extraction framework is well-suited for our research objective.

To address this, we propose a Multi-Layer Stacked Architecture (MLSA) inspired by span-based joint extraction models, with BERT as the foundational encoder. The architecture is composed of four layers, each serving a specific purpose: EE and NER, span-based representation, a multi-head attention mechanism, and RE.

The first layer uses BERT to extract events and named entities, providing contextualized token representations for the entire sentence. In the second layer, these outputs are processed by a multi-head attention mechanism that selectively focuses on key interactions between events and entities, refining the representations. The third layer constructs span-based embeddings to better capture the semantic boundaries of events and entities. Finally, the fourth layer leverages these refined span representations to predict events and their associated parameters. By incorporating event arguments, such as agents and targets, this layer also determines whether an entity caused an event or was affected by it. This layered architecture enables a unified modeling of events, entities, and their causal relationships, thus improving overall performance in joint extraction tasks.

In parallel with the BERT-based MLSA, we also explored a GPT-based inference approach. Leveraging both general-purpose GPT models and inference-optimized variants such as o3-mini, we processed threat intelligence

Just Accepted

reports augmented with event and entity annotations. The models were prompted to generate outputs that describe the causal relationships present in the data.

The experimental results revealed complementary strengths and limitations. General-purpose GPT models demonstrated strong natural language understanding and general causal reasoning capabilities, but struggled to accurately link specific entities to events. In contrast, models optimized for inference like o3-mini produced more precise causal inferences, likely due to improved knowledge integration and contextual understanding. Although GPT models occasionally outperformed MLSA in terms of inference quality, the MLSA model exhibited more consistent performance across varied inputs owing to its structured architecture and self-attention mechanisms, which enhanced its ability to extract detailed event-entity relationships.

Computational efficiency is another important consideration. Large GPT models require substantial computational resources due to their extensive parameter counts and transformer-based design. In comparison, the BERT-based MLSA is more efficient for domain-specific tasks, benefiting from a smaller model size and structured input representations. These findings illustrate the trade-offs between generalization and efficiency, which must be carefully weighed in real-world cybersecurity applications.

In summary, while GPT inference models excel at open-domain reasoning and complex contextual understanding, the BERT-based MLSA model offers a more efficient, stable, and structured approach to extracting causal relationships between events and entities. Together, these two approaches offer complementary roles in a human–AI collaborative SOC workflow: MLSA provides efficient, batch-level causal annotation suitable for integration into automated pipelines, while GPT-based models support interactive, analyst-driven causal queries for complex or ambiguous incidents. These insights provide valuable guidance for selecting and optimizing models in specialized deep learning applications.

This study makes the following key contributions:

- **`Dataset Creation`**: We construct a high-quality, annotated dataset specifically designed to capture causal relationships between cybersecurity events and entities, thereby addressing a previously unaddressed gap in resources for training and evaluating causal inference models in this domain.
- **`MLSA Model Proposal`**: We introduce a BERT-based Multi-Layer Stacked Architecture (MLSA) that jointly models event extraction, entity recognition, and causal relationship inference, demonstrating strong performance and robustness on cybersecurity threat data.
- **`Prompt Design for GPT Models`**: We develop and evaluate task-specific prompting strategies for GPT-based models, revealing their strengths and limitations in performing causal inference over structured threat intelligence inputs.
- **`Novel Research Direction`**: We propose a novel perspective on modeling causal relationships between events and entities, opening new avenues for fine-grained analysis in event-driven domains such as cybersecurity, healthcare, and finance.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes dataset creation and preprocessing, Section 4 details the proposed model architecture, Section 5 presents experimental results and analysis, and Section 6 concludes the paper.

2 Related Work

Event extraction (EE), entity extraction, and causal relationship inference are pivotal research areas in natural language processing (NLP), playing critical roles in event understanding and knowledge construction. These tasks aim to identify events and entities within text and analyze their semantic interactions.

With the advancement of NLP, BERT has become a widely adopted backbone architecture for these tasks due to its ability to model contextualized semantic representations and capture long-range dependencies. Unlike traditional methods such as Word2Vec and GloVe, which generate static word embeddings, BERT employs a

Just Accepted

bidirectional Transformer encoder that dynamically produces contextualized representations. This design has led to significant improvements in tasks such as named entity recognition (NER), question answering, and sentence classification.

The pre-trained Transformer encoder of BERT, which leverages the self-attention mechanism, is particularly effective in modeling bidirectional context. As a result, it is well-suited for tasks that require a nuanced understanding of semantic relationships across both sentence-level and document-level contexts. Prior research has demonstrated that applying BERT to EE enhances the identification of event trigger words and argument roles, surpassing traditional machine learning and feature-engineered approaches. Similarly, BERT-based entity extraction frameworks have achieved notable success in capturing complex entities and their interactions.

In mainstream EE and NER research, BERT is often used as the foundational model. Some studies apply it directly to NER tasks [15], while others integrate BERT with more complex architectures to boost task-specific performance. Hybrid models that combine BERT with BiLSTM-CRF architectures [16, 23, 28] have demonstrated strong results across various information extraction tasks. In these designs, BiLSTM effectively captures sequential dependencies, while the CRF layer models label transitions and boundary constraints, resulting in improved sequence labeling outcomes. These architectures typically adopt BIO-based labeling schemes.

Notably, augmenting BERT with additional attention mechanisms has also proven effective. Models such as those in [17, 30] incorporate graph-based or task-specific attention modules to dynamically emphasize key contextual information. This refinement enhances the semantic alignment between spans and improves the model's generalization ability across diverse data settings.

This integration of BERT with complementary components, such as attention mechanisms and sequence labeling layers, has led to a dominant "BERT-centric" paradigm in EE and NER research. The success of these architectures underscores the versatility of BERT in adapting to both standalone and hybrid extraction frameworks.

Relation extraction (RE)—particularly for semantic categories such as causal and temporal relations—is another key area that can benefit from BERT's capabilities. BERT-based RE models leverage its contextual embeddings to capture complex dependencies between entities or events. Previous studies [9, 13, 21] have shown that using BERT improves the accuracy of extracting relational links. Recent work [24] also highlights that integrating BERT with convolutional neural networks (CNNs) helps compensate for its limitations in capturing local features. By applying one-dimensional convolutional layers on top of BERT representations, these models can extract additional local semantic information from neighboring tokens, leading to higher precision in challenging scenarios such as overlapping entities or multiple relations. Furthermore, combining BERT with additional attention layers or structured prediction modules can further enhance extraction performance [11, 14, 25].

A notable trend in recent research is the adoption of joint learning frameworks that integrate EE, NER, and RE into unified model architectures. Joint models such as those proposed in [5, 11, 14, 15, 17, 24, 25, 29, 30] are designed to simultaneously learn multiple interrelated tasks. This design facilitates the discovery of latent correlations among subtasks, improving overall extraction accuracy compared to traditional pipeline architectures. The joint approach has proven especially effective for complex scenarios that require holistic modeling of events, entities, and their interdependencies.

In addition to BERT-based models, the Generative Pre-trained Transformer (GPT) family has gained increasing attention for relation extraction tasks, particularly in scenarios that require reasoning or adaptation with limited training data. The study in [26] investigates the use of ChatGPT for temporal relation extraction using three prompting strategies: Zero-shot, Event Ranking, and Chain of Thought. Recent work by Huang et al. [7] (CTIKG) leverages GPT-4 to construct comprehensive knowledge graphs from CTI reports by extracting various entity relationships. While CTIKG focuses on extracting general relationships between entities to construct knowledge graphs, our work specifically targets directional causal relationships between events and entities—distinguishing which entities trigger events (Causal) versus which are affected by them (Effect). This event-entity causal inference is essential for root cause analysis in cybersecurity incidents but is not addressed by CTIKG. Building on the

Just Accepted

strengths of both BERT-based and GPT-based approaches, our work introduces a specialized framework for event-entity causal inference in CTI analysis.

Although earlier versions of GPT exhibited limited capabilities in causal inference, more recent iterations have shown improved performance on complex NLP tasks. For example, [1] proposes a three-step prompting strategy that enables GPT to perform causal reasoning and provide interpretive explanations. Similarly, [6] evaluates GPT's reasoning ability by testing various conditions such as sentence reordering, prompt variation, and one-shot learning. Notably, [27] is among the first to explore causal relationships between events and entities by analyzing how changes in entity states indicate event occurrence. However, its focus is limited to entity-triggered events and does not comprehensively address the full range of causal associations between events and entities.

Collectively, these studies highlight the strengths and evolving capabilities of both BERT- and GPT-based models in extraction and inference tasks. However, most existing research focuses on modeling either event-to-event or entity-to-entity relationships in isolation. The causal interactions that span across events and entities remain relatively underexplored.

As summarized in Table 1, substantial progress has been made in event extraction, entity recognition, and relation inference. However, most existing approaches employing BERT or GPT have focused on modeling relationships of a single type—typically between events or between entities—without fully exploring the causal interactions that span across these types. This narrow focus presents a critical gap, particularly in domains such as cybersecurity, where understanding the causal dynamics between events and the entities involved is essential for timely threat response and analysis. To address this limitation, the present study proposes a unified framework that models and infers causal relationships between events and entities. By targeting this underexplored intersection, our work advances the field of causal inference and enhances its applicability to real-world, high-stakes scenarios.

3 Datasets

This section introduces the datasets used in this study, including the CASIE dataset [16] for extracting cyber threat events and a custom-annotated dataset that establishes causal relationships between events and entities. These datasets form a solid foundation for our work, supporting a more precise analysis of cybersecurity incidents and their associated impacts.

3.1 Network Threat Event Dataset

This study utilizes the CASIE dataset [16], which is designed to annotate and structure cybersecurity event data. Its primary objective is to facilitate the automatic extraction of critical cybersecurity-related information from unstructured news reports. The dataset comprises 1,000 news articles describing cybersecurity incidents and categorizes events into five frequent and high-impact types: Attack.Databreach, Attack.Phishing, Attack.Ransom, Discover.Vulnerability, and Patch.Vulnerability.

Each event is meticulously annotated with both event triggers—keywords or phrases that indicate the occurrence of an event—and event arguments, which represent the entities involved, such as Organization, Person, and Capability. Table 2 and Table 3 present the number of annotated events across the 1,000 news articles, as well as the 21 categories and corresponding counts of event arguments.

In addition to core event and entity labels, the CASIE dataset further enhances annotation granularity by assigning role-specific labels to entities based on event type. For example, in attack-related events, a Person may be labeled as an Attacker or a Victim. These detailed annotations provide a structured framework for interpreting complex cybersecurity scenarios and serve as a valuable resource for advancing event extraction and causal relationship modeling.

Table 4 further illustrates all event roles and their corresponding counts.

Just Accepted

Table 1. Comparison of related works

No.	Year	Event Extraction	Entity Extraction	Relation Extraction	Causal/Effect Inference	Inference object	Model	Joint Model
Tan et al. [21]	2022	✗	✗	✓	✓	Events	Bert	-
Khetan et al. [9]	2022	✗	✗	✓	✓	Events	Bert	
Ban et al. [1]	2023	✗	✗	✓	✓	Sentences	GPT-4	-
Hobbhahn et al. [6]	2023	✗	✗	✓	✓	Sentences	GPT-3.5/4	-
Zhang et al. [27]	2023	✗	✗	✓	✓	Event-Entities	GPT-3	-
Santosh et al. [15]	2021	✗	✓	✓	✗	Entities	Bert	✓
Satyapanich et al. [16]	2020	✓	✓	✗	✗	✗	Bert-Bilstm-CRF	✗
Wei et al. [23]	2023	✓	✓	✗	✗	✗	Bert-bilstm-CRF	✗
Zhang et al [28]	2022	✗	✓	✓	✗	Entities	Bert-bilstm-CRF	✗
She et al. [17]	2022	✓	✗	✗	✗	✗	Bert-GCN-Attention	✓
Eberts et al. [5]	2020	✗	✓	✓	✗	Entities	Bert	✓
Zhu et al. [30]	2023	✗	✓	✓	✗	Entities	Bert-Attention	✓
Park et al. [13]	2022	✗	✓	✓	✗	Entities	Bert	✗
Qiao et al. [14]	2022	✗	✓	✓	✗	Sentences	Bert-bilstm-LSTM	✓
Yao et al. [25]	2022	✓	✗	✓	✗	Events	Bert-GNN	✓
Yuan et al. [26]	2023	✗	✗	✓	✗	Events	ChatGPT	-
Yang et al. [24]	2024	✗	✓	✓	✗	Entities	Bert-CNN	✓
Li et al. [11]	2024	✗	✓	✓	✗	Entities	Bert-CNN-Attention	✓
Zhao et al. [29]	2021	✗	✓	✓	✗	Entities	Bilstm-Attentation	✓
Our	2025	✓	✓	✓	✓	Event-Entities	Bert-MLSA/GPT o3	✓/-

Table 2. Event type and count

Event types	**Counts**
Databreach	1783
Ransom	1585
Phishing	1562
Discover.Vulnerability	2132
Patch.Vulnerability	1417

Just Accepted

Building upon this dataset, the present study further refines and extends its annotations to better align with the research objectives. The information provided by CASIE is also used during model training in the event extraction process.

3.2 Causal Relationship Dataset for Events and Entities

To the best of our knowledge, there is currently no publicly available dataset that captures fine-grained causal relationships between cybersecurity events and entities. To address this gap, we developed a custom-annotated dataset specifically designed to support the training and evaluation of causal reasoning models. Given the high cost of manual annotation, we adopted a semi-automated hybrid approach that combines GPT-based generation with human validation and expert review.

Annotation Procedure:

Table 3. Event argument and count

Event arguments	Counts
Number	260
Person	3832
Data	979
Vulnerability	2554
CVE	317
Capabilities	1742
Malware	484
Time	1402
File	605
Software	845
Organization	3503
Website	325
Patch	771
GPE	167
Device	757
System	1437
PaymentMethod	195
Purpose	555
PII	1071
Version	463
Money	413

Just Accepted

The annotation process consisted of the following stages:

Stage 1 - Pilot Annotation: We manually constructed a pilot dataset of 30 cybersecurity news reports, fully annotated by a domain expert. These reports were selected to ensure coverage of all event types defined in CASIE [16]. In this pilot dataset, we defined two types of directional causal relationships:

- **Causal:** The entity is identified as the cause of an event (entity → event).
- **Effect:** The entity is affected as a result of an event (event → entity).

To ensure consistency, we established clear annotation guidelines based on entity roles within events (e.g., Attacker → Causal, Victim → Effect). In cases where entity roles did not clearly indicate causality, contextual analysis of the surrounding text was applied.

Since each news report may describe multiple cybersecurity incidents, we aimed to construct a dataset that adequately represents all event categories while maintaining balance across them. Specifically, we ensured that each event type contained at least 25 instances, and that the variation in sample counts among event types remained within a reasonable range to minimize bias caused by data imbalance. Furthermore, reports containing entities that were more likely to exhibit explicit causal relationships were prioritized, ensuring that GPT could effectively learn and generalize causal patterns across diverse event types during prompt optimization.

Stage 2 - GPT-Assisted Generation: We used GPT-4o with few-shot prompting, providing the 30 manually annotated samples as demonstrations in the system prompt. This enabled the model to learn causal relationship patterns from examples and generate initial causal annotations for the remaining dataset. The prompts included: (1) a system role defining the model's annotation objective, (2) formatted demonstrations of the 30 annotated

Table 4. Event role and count

Event arguments	Counts
Attacker	1671
Time	1402
Trusted-Entity	850
Vulnerability	2554
Number-of-Victim	113
Price	370
Releaser	785
CVE	317
Number-of-Data	147
Patch	771
Vulnerable_System_Version	315
Payment-Method	195
Vulnerable_System	2055
Victim	3473
Tool	1117
Supported_Platform	118
Vulnerable_System_Owner	380
Place	167
Issues-Addressed	135
Damage-Amount	43
Compromised-Data	1999
Capabilities	557
Purpose	555
Attack-Pattern	1050
Discoverer	1390
Vulnerability	2554
Patch-Number	148

Just Accepted

examples, (3) event types and participating entities with their corresponding entity types, (4) the annotation task for participating entities, and (5) an instruction to output results in JSON format. At this stage, we prioritized efficiency over perfection, using temperature = 0.0 for deterministic output. The primary objective was to obtain reasonable initial annotations quickly, domain experts manually corrected all GPT outputs.

Stage 3 - Manual Correction and Expert Review: Since GPT-based inference has not yet reached human-level precision, all GPT-generated annotations underwent manual correction by the primary annotator, who verified and corrected causal relationships based on entity roles and contextual analysis. Subsequently, all annotations were reviewed and validated by a second cybersecurity domain expert. The reviewing expert verified annotation consistency and resolved ambiguous cases through discussion with the primary annotator to reach consensus.

Quality Control Measures:

- Clear annotation guidelines refined through pilot annotation
- Expert review and validation of all annotations by a second specialist

- Consensus-building through discussion for disagreements or ambiguous cases
- Cross-validation against CASIE's entity role labels to ensure logical coherence (e.g., Attacker entities should not be labeled as Effect)

Inter-Annotator Agreement: Our semi-automated pipeline involved sequential annotation (GPT → primary correction → expert review) rather than independent double-annotation, which does not directly yield traditional IAA metrics. To address this, we conducted a prospective reliability study. We selected a stratified random subset of 100 reports containing 2,560 entity-event pairs, ensuring proportional representation of all five event types. An independent annotator with expertise in cybersecurity labeled all causal relationships following the same annotation guidelines without access to the original labels. Cohen's kappa between the independent annotations and the consensus ground truth was $\kappa = 0.85$, indicating almost perfect agreement [10]. The primary source of disagreement involved entities labeled as *No Relation* in the ground truth but assigned a causal role by the independent annotator, suggesting that the original annotations adopt a conservative labeling strategy. Notably, disagreements between *Causal* and *Effect* labels were near zero, confirming that the directional distinction between cause and effect is well-defined in our annotation guidelines.

Dataset Characteristics and Known Limitations:

The dataset inherits CASIE's event type distribution, with Databreach and Vulnerability events being more prevalent than Phishing or Ransom events. The Causal-to-Effect ratio is approximately 1:1.3, reflecting typical cybersecurity scenarios where affected entities outnumber attackers.

We acknowledge several limitations. First, despite the semi-automated approach and expert review, some subjective interpretation remains in ambiguous cases. Specifically, during the expert review stage (Stage 3), the reviewing expert flagged 3,798 out of 25,574 entity-event pairs (14.8%) as requiring discussion before reaching consensus. These cases primarily involved entities serving dual semantic roles (e.g., an organization functioning as both a location modifier and a potential victim) and entities with ambiguous causal directionality. The remaining 85.2% of annotations were accepted without modification, indicating a high degree of initial annotation consistency. Second, the event type imbalance may affect model generalization to underrepresented event types. Third, while the GPT-assisted generation improved efficiency, it may introduce systematic biases based on the GPT model's learned patterns.

The newly annotated dataset [8] has been released publicly through a GitHub repository with detailed annotation guidelines, providing researchers and practitioners with a valuable resource for exploring causal relationships between events and entities. This transparency allows the community to review, refine, or extend the annotations according to their needs and to assess potential biases relevant to their applications.

4 Methodology

This study proposes a Multi-Layer Stacked Architecture (MLSA) for joint event extraction and causal relationship inference. The MLSA framework is inspired by the SpERT model [5], which jointly addresses entity and relation extraction. Building upon SpERT, we extend the framework to simultaneously extract event triggers, event arguments, and their causal relationships. The proposed MLSA framework employs BERT for semantic representation and incorporates a multi-head attention mechanism to enhance the extraction of events and their arguments. Event arguments are classified into specific roles, with these role labels serving as auxiliary features that enrich the contextual understanding of entity-event interactions, thereby facilitating more accurate causal inference.

The methodology includes two distinct approaches for causal inference:

(1) `MLSA-based Event Extraction and Causal Inference` — This approach involves extracting event triggers and arguments using BERT and multi-head attention, followed by establishing causal relationships between events, triggers, and arguments based on contextual and structural features.

Just Accepted

(2) `GPT-based Causal Inference` — This approach leverages GPT to refine causal inference through zero-shot/few-shot prompting and advanced prompt engineering.

Ultimately, we compare the performance of causal inference using the MLSA framework with that of the GPT-based approach, evaluating their effectiveness in capturing and refining causal relationships.

4.1 Event Extraction of MLSA

The event extraction stage involves three components: contextual embedding via BERT, feature enhancement through multi-head attention, and span-level classification.

`BERT Layer`: Input sentences are tokenized using the NLTK toolkit, with special tokens `[CLS]` and `[SEP]` appended to construct a complete input sequence. The sentence is passed into a pre-trained BERT encoder, which generates contextualized token embeddings. These embeddings provide a deep semantic foundation for all subsequent modules.

`Multi-Head Attention Layer`: The BERT-generated embeddings are passed through a multi-head attention layer to refine token-level features by capturing dependencies between distant tokens. This layer produces an attention-weighted representation of each span, which is crucial for modeling variable-length phrases. These span representations are combined with engineered features in the next step (illustrated in the Event Extraction stage of Figure 1).

`Feature Engineering`: To bridge the event extraction and causal inference stages, we extract the following four feature types, which serve as inputs to the classification layer:

Features for Event and Argument Extraction:

- **`Span representation`**: Generated through the multi-head attention mechanism, span representations capture variable-length phrases. For example, for the input ["How", "are", "you"], the span set includes single-token and multi-token spans such as ["How"], ["are"], ["you"], ["How are"], ["are you"], and ["How are you"]. To reduce computational overhead and avoid excessive model complexity, we apply a maximum span length constraint during generation.
- **`Span-Level POS Tagging`**: To enhance linguistic feature representation, part-of-speech (POS) tagging was applied using the NLTK toolkit. Prior studies (e.g., [22]) have shown that incorporating both token-level and span-level POS information can improve model performance in sequence labeling tasks. However, in our experiments, span-level POS tagging alone yielded more substantial performance gains compared to combining it with token-level tags. Therefore, only span-level POS tagging was adopted in this work, with POS tags assigned to each span using NLTK.
- **`Span Length`**: The length of each span is encoded as a learnable feature. During training, this parameter is refined through backpropagation to reflect the typical length distributions of valid event triggers and arguments, which tend to be short. This allows the model to down-weight overly long or semantically irrelevant spans.
- **`CLS Token Representation`**: The `[CLS]` token embedding generated by BERT serves as a global representation of the sentence. This vector captures high-level semantic information and is used to enrich local span representations, especially in ambiguous or complex sentence contexts.

The four feature vectors described above are concatenated and passed into a classification layer. This layer determines whether each span represents an event trigger, an event argument, or neither.

4.2 Causal Relationship Inference of MLSA

Once event triggers and arguments have been extracted, the next step is to infer causal relationships between the identified events and associated entities. As illustrated in the Causal Relationship Inference stage of Figure 1, this process comprises the following steps:

Just Accepted

(1) Pairing event triggers with arguments based on their extracted features.
(2) Incorporating contextual and role-based information to enhance inference.
(3) Classifying relationships between event-argument pairs into `Causal`, `Effect`, or `No Relation`.

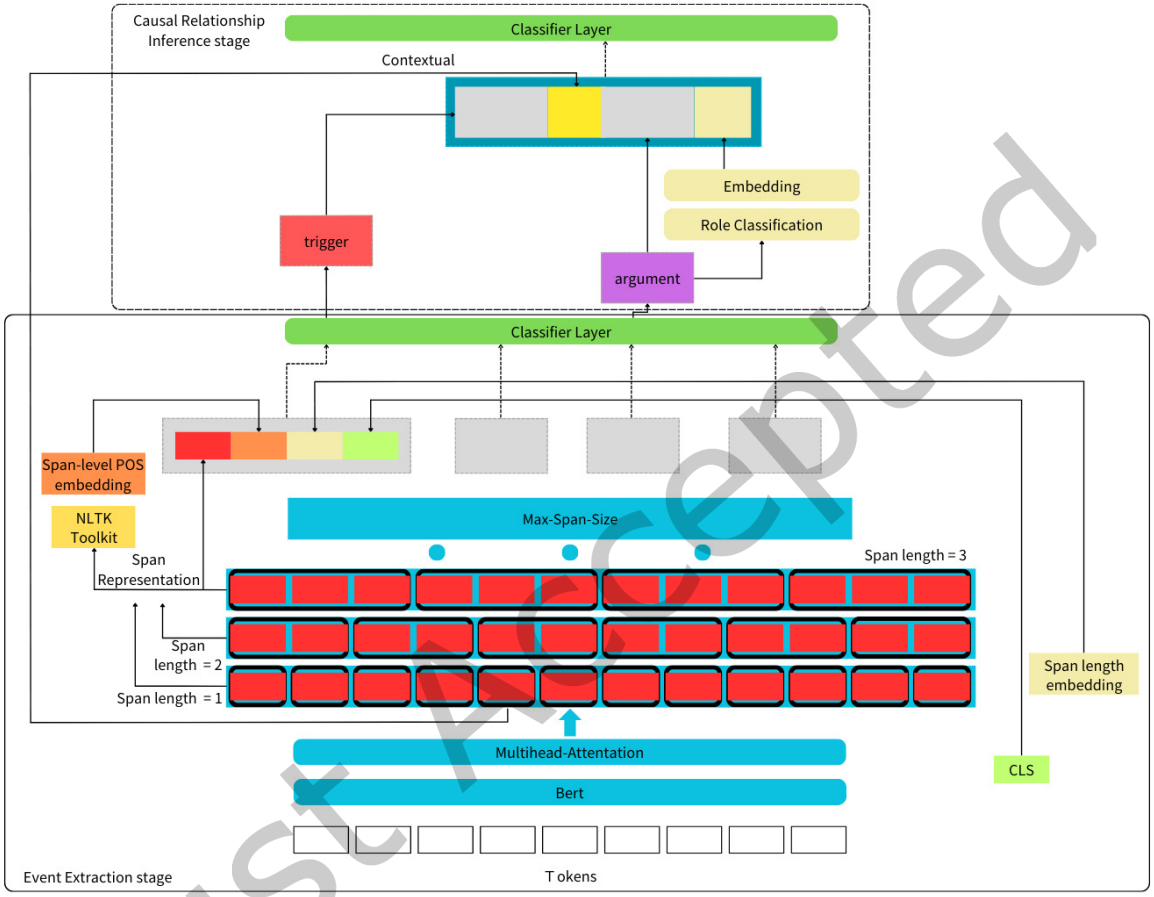

Fig. 1. Overall Research Framework

Just Accepted

Features for Causal Relationship Inference:

- **`Contextual information`**: Contextual information derived from event triggers and the surrounding entities plays a critical role in supporting causal relationship inference. This background knowledge provides the model with a broader semantic understanding of how events unfold within a sentence. By analyzing the local context around trigger words and nearby entities, the model can identify interaction patterns that reflect the underlying causal structure.
- **`Event Role Classification`**: Entities involved in an event are assigned specific semantic roles to support causal structure inference. For example, an entity of type `Person` may be classified as an `Attacker` or a `Victim`, depending on its function within the event context. These roles provide intuitive cues regarding the

entity's causal position in the event. An `Attacker` is typically associated with the cause of an event, whereas a `Victim` is more often linked to the outcome or effect. The role classification results are incorporated as input features, thereby enhancing the model's ability to learn and infer causal relationships between events and entities.

These features are combined with the previously described features of the trigger–participating entity pair. The integrated representation is then passed into a classification layer, which determines whether the given pair corresponds to a `Causal`, `Effect`, or `No Relation` relationship.

Model summary. Table 5 presents the model summary of **MLSA**. With approximately **110 million** trainable parameters, MLSA is substantially smaller than large proprietary models such as GPT-4o. Owing to its compact architecture, the model achieves higher computational efficiency and more effective utilization of resources.

Table 5. Model Summary for MLSA

Layer (type:depth-idx)	Output Shape	Param #
MLSA	[8, 51, 28]	–
BertModel: 1-1	[8, 768]	–
BertEmbeddings: 2-1	[8, 52, 768]	–
Embedding: 3-1	[8, 52, 768]	22,268,928
Embedding: 3-2	[8, 52, 768]	1,536
Embedding: 3-3	[8, 52, 768]	393,216
LayerNorm: 3-4	[8, 52, 768]	1,536
Dropout: 3-5	[8, 52, 768]	–
BertEncoder: 2-2	[8, 52, 768]	–
ModuLeList: 3-6	–	85,054,464
BertPooler: 2-3	[8, 768]	–
Linear: 3-7	[8, 768]	590,592
Tanh: 3-8	[8, 768]	–
MultiheadAttention: 1-2	[8, 52, 768]	2,362,368
Embedding: 1-3	[8, 51, 25]	625
Embedding: 1-4	[8, 51, 64]	2,944
Dropout: 1-5	[8, 51, 1625]	–
Linear: 1-6	[8, 18, 28]	45,528
Embedding: 1-7	[8, 18, 64]	1,920
Dropout: 1-8	[8, 18, 2546]	–
Linear: 1-9	[8, 18, 3]	7,641
Total params:		110,731,298
Trainable params:		110,731,298
Non-trainable params:		0
Total mult-adds (M):		864.20

Just Accepted

4.3 GPT-Based Causal Inference

Recent advancements in GPT-based causal inference have attracted significant attention in the research community. This study investigates whether GPT can generate causal inference results that align with the intended research objectives. To address this, we employed GPT-4o to conduct causal inference experiments through the GPT API.

Zero-Shot Approach: GPT's performance was initially evaluated in a zero-shot setting, in which no annotated examples were provided during inference. Event descriptions and relevant entity information were directly supplied to GPT using prompts that instructed it to infer causal relationships. While the model demonstrated logical reasoning capabilities, its outputs exhibited inconsistencies in formatting and occasionally included unexpected or irrelevant results.

Few-Shot Learning and Prompt Optimization: To improve GPT's accuracy and consistency, we adopted a hybrid strategy involving few-shot demonstrations and iterative prompt design.

- **Few-Shot Demonstration Strategy**: We used the 30 manually annotated cybersecurity reports (described in Section 3.2) as few-shot demonstrations for GPT during the causal inference stage. These demonstrations were provided in the system prompt, showing GPT examples of how to identify `Causal` and `Effect` relationships between events and entities. Each demonstration included: (1) event trigger words and their positions, (2) participating entities with type and position information, and (3) the correct causal relationship labels. This approach enabled the model to learn the task structure and apply consistent reasoning patterns to new, unseen data.
  To facilitate systematic analysis and evaluation, GPT was instructed to generate responses in a predefined JSON format. This structured output ensured a clear distinction between `Causal` and `Effect` entities, as illustrated in Figure 2.
- **Iterative Prompt Engineering**: To enable GPT to produce more accurate and consistent causal inferences, we began with the most basic prompt—'Please generate a causal analysis based on the events that occurred and the entities involved and send it back to me'—observed GPT's outputs, and repeatedly refined the prompt based on the errors we observed, so that GPT would perform causal analysis only on the participating entities and no overly obvious errors would remain.


```
"Causal": [
[],
[{"start": 249, "end": 250, "text": ["The", "hacker"], "type": "Person"}],
[],
[{"start": 269, "end": 269, "text": ["hacker"], "type": "Person"}]],

"Effect": [
[{"start": 260, "end": 262, "text": ["highly", "sensitive", "data"], "type": "Data"}],
[{"start": 257, "end": 257, "text": ["server"], "type": "System"}],
[{"start": 279, "end": 280, "text": ["personal", "details"], "type": "PII"}, {"start": 282, "end": 283, "text": ["police", "office"], "type":
"Person"}],
[{"start": 274, "end": 274, "text": ["server"], "type": "System"}]]
```


Fig. 2. Example of GPT output

In addition to GPT-4o, OpenAI has released inference-optimized variants, namely `o1` and `o3-mini`. To comprehensively evaluate their performance on causal inference tasks, we applied the same few-shot learning approach used with GPT-4o. These models were provided with the same 30 annotated examples as demonstrations to align their causal reasoning with our research objectives. However, unlike GPT-4o, which required extensive prompt engineering to address issues such as inconsistent outputs and limited domain alignment, the inference-optimized models demonstrated stronger instruction-following capabilities. Nevertheless, we initiated a new round of prompt engineering for each model, developing prompts from scratch and refining them iteratively to ensure that the generated inferences aligned more closely with our expectations.

Just Accepted

By integrating BERT-based event extraction, structured feature engineering, and GPT-assisted causal inference, the proposed methodology establishes a comprehensive framework for event and causality extraction. This approach combines the strengths of traditional machine learning models with those of advanced generative AI techniques, resulting in improved accuracy and robustness in modeling causal relationships.

5 Evaluation

All experiments were conducted on NVIDIA H100 GPU (80 GB VRAM) with CUDA 11.8 and PyTorch 2.6.0+cu124. The model was trained for 50 epochs (best performance achieved at epoch 45) using the AdamW optimizer with a learning rate of 5e-5.

In addition to conventional hyperparameters, two task-specific negative sampling parameters were introduced: neg_entity_count = 150 and neg_relation_count = 100. These parameters control the number of negative entity and relation samples generated during training, which is critical for maintaining a balanced ratio between positive and negative examples and ensuring robust contrastive learning.

SpERT [5] indicates that setting these values too low leads to slow convergence and poor recognition of entity span boundaries, as the model lacks sufficient negative evidence to distinguish entity edges. Conversely, excessively large values cause the model to converge too rapidly in the early training phase, failing to adequately learn from positive examples while also resulting in excessive GPU memory consumption.

To assess the performance of our model, we employ three standard classification metrics: Precision, Recall, and F1-score. These metrics provide a comprehensive evaluation of the model's effectiveness across various application scenarios.

Given the following definitions:

- `TP (True Positive)`: The number of correctly predicted positive instances.
- `FP (False Positive)`: The number of incorrectly predicted positive instances.
- `FN (False Negative)`: The number of actually positive instances incorrectly classified as negative.

The three evaluation metrics are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

By leveraging these metrics, we systematically analyze the classification performance of our model and make informed adjustments to enhance its accuracy and robustness.

5.1 Comparison of Event Extraction Performance

The accuracy of event and event parameter extraction is critical to the effectiveness of relational inference. Low extraction accuracy can directly degrade the overall performance of causal reasoning. To assess the impact of our approach, we compare its performance against CASIE [16] in both event and event parameter extraction. The results of this comparison are presented in Table 6 and Table 7.

Table 6 presents a comparative analysis of CASIE and our method across different event categories. In most cases, our model achieves a higher F1-score than CASIE, with particularly notable improvements in the `Databreach`, `Phishing`, and `DiscoverVulnerability` categories. Furthermore, the Micro-average (`Micro avg`) score indicates that our method exhibits greater stability across event types, reinforcing its robustness in multi-class event classification.

Just Accepted

Table 6. Comparison of Event Extraction Results

Label	Precision(%)		Recall(%)		F1-score(%)	
	CASIE	Ours	CASIE	Ours	CASIE	Ours
Databreach	75.2	78.2	82.2	82.4	78.5	80.2
Phishing	81.5	79.6	77.3	80.3	79.4	79.9
Ransom	84.7	80.9	80.1	78.8	82.3	79.8
DiscoverVulnerability	76.0	80.5	81.7	81.8	78.7	81.1
PatchVulnerability	84.6	85.6	79.1	80.7	81.8	83.0
Micro avg	79.6	81.4	80.2	80.6	79.8	81.1

However, CASIE achieves a higher F1-score in the `Ransom` category (82.3% vs. 79.8%). This is because ransom events tend to employ a limited set of stereotypical trigger words (e.g., "ransom," "encrypt," "decrypt"). CASIE's BiLSTM-CRF architecture is particularly effective at capturing such sequential patterns with well-defined label transitions, whereas our span-based approach may introduce additional false positives when trigger expressions are short and formulaic.

As shown in Table 7, our method also demonstrates superior performance across most event argument categories. In key categories such as `CVE`, `Patch`, `Organization`, and `Time`, the proposed model outperforms CASIE in terms of F1-score, highlighting its effectiveness in capturing structured information with high accuracy. Additionally, improvements in categories such as `Capability`, `Purpose`, and `Malware` suggest that our method enhances recognition in more challenging argument extraction tasks. Accurate identification of entities likely to serve as causal or effect factors is essential for downstream causal inference, further validating the practical significance of these improvements.

Nevertheless, CASIE outperforms our method in three argument categories: `Vulnerability` (83.5% vs. 80.0%), `PII` (81.1% vs. 76.2%), and `Person` (80.7% vs. 79.7%). These categories share a common characteristic: their entity boundaries are relatively well-defined and consistent across documents. `Vulnerability` entities typically follow standardized naming conventions, `PII` entities are often short noun phrases, and `Person` entities are proper nouns with clear boundaries. CASIE's BIO-based sequence labeling with CRF boundary constraints is well-suited for such entities. In contrast, our span-based approach faces challenges when multiple overlapping candidate spans compete, occasionally leading to boundary misalignment—consistent with the entity segmentation errors analyzed in Section 5.5.1 (Cases 1 and 4). Conversely, our method significantly outperforms CASIE in categories involving longer or more semantically diverse entities, such as `Capability` (+13.9%), `Purpose` (+17.0%), and `GPE` (+21.8%), where the span-based representation and multi-head attention mechanism provide clear advantages for capturing complete semantic meaning.

5.2 Comparison of Event Role Recognition

Since event roles are a core feature in causal inference, the model's ability to accurately identify these roles directly affects the quality of final inference outcomes. To evaluate this impact, we compare the performance of our method with CASIE in event role recognition, as shown in Table 8.

The results indicate that our model outperforms CASIE in most event role categories, with particularly strong performance in `Attacker`, `Discover`, `Victim`, and `Trusted-Entity`. Notably, recall in categories such as `Patch-number`, `Tool`, and `Vulnerable-System-Version` has significantly improved. This suggests that our method reduces false negatives and enhances the model's ability to capture fine-grained role-related information.

Just Accepted

Table 7. Comparison of Event Argument Extraction Results

Label	Precision (%)		Recall (%)		F1-score (%)	
	CASIE	Ours	CASIE	Ours	CASIE	Ours
Money	93.9	95.1	92.8	94.0	93.3	94.5
CVE	85.5	93.4	86.3	95.5	85.9	94.4
Patch	81.7	87.5	80.4	86.9	81.0	87.2
GPE	76.0	73.5	31.6	60.8	44.7	66.5
Capability	45.2	64.4	59.0	65.9	51.2	65.1
Purpose	65.2	79.7	53.6	69.3	58.8	75.8
Website	54.9	68.5	67.9	70.3	60.7	69.3
Data	77.6	79.5	67.6	72.2	72.2	75.6
Device	65.1	64.7	63.9	64.2	64.5	64.5
Organization	75.0	78.2	84.2	83.6	79.3	80.8
PaymentMethod	78.3	81.6	83.7	83.9	80.9	82.1
Vulnerability	86.1	78.6	81.1	81.3	83.5	80.0
File	78.9	79.4	74.7	75.8	76.8	77.5
Number	87.3	85.8	83.5	86.1	85.4	85.9
PII	78.7	73.9	83.6	78.7	81.1	76.2
Malware	68.6	82.8	51.3	60.8	58.7	70.1
Version	69.8	75.7	66.5	78.9	68.1	77.3
Time	76.4	78.5	85.1	89.3	80.5	83.4
Person	76.7	79.2	85.1	80.3	80.7	79.7
Micro avg	72.8	78.5	76.7	77.4	74.7	77.9

Just Accepted

Such improvements are crucial for constructing a robust causal relationship inference framework, as accurate role identification enhances both the precision and reliability of causal modeling.

5.3 Comparison of Causal Inference Performance

To facilitate a deeper understanding of the generative logic and application context behind the results produced by the GPT model and the MLSA framework in this study, a concrete example is provided (see Figure 3 for the original text). This example enables readers to grasp the operational mechanism of the subsequent analytical process more effectively, thereby enhancing their overall comprehension of the research methodology and its underlying principles.

5.3.1 Experiments on the Latest GPT Model, GPT-4o.

In the early stages of our experiments, GPT-4o demonstrated promising capabilities in causal inference. Even when trained on a small dataset with simple prompts, it was able to produce reasonable results. However, some limitations were observed in the output, including the following issues:

(1) `Conflicting Labels`:
In some cases, the same entity was labeled as both `Causal` and `Effect`, which is logically inconsistent. To resolve this, we refined the prompts to explicitly enforce that an entity can belong to only one category or none. Initial prompt → "Please generate a causal analysis based on the events that occurred and the

Table 8. Comparison of Event Role Recognition Results

Label	Precision (%)		Recall (%)		F1-score (%)	
	CASIE	Ours	CASIE	Ours	CASIE	Ours
Attacker	83.9	85.3	77.6	81.7	80.6	83.5
Discover	83.5	90.1	87.5	88.0	85.5	89.1
Number-of-Data	100.0	95.8	100.0	97.4	100.0	96.5
Number-of-Victim	100.0	96.3	100.0	97.9	100.0	97.0
Patch-number	72.0	73.6	45.0	60.1	55.3	66.1
Ransom-Price	100.0	96.5	98.5	98.3	99.2	97.4
Releaser	81.6	92.7	87.5	90.5	84.5	91.6
Trusted-Entity	75.5	80.4	77.5	78.1	76.4	79.2
Vulnerable-System-Version	54.0	71.8	75.0	75.9	62.7	73.6
Vulnerable-System-Owner	74.0	75.7	77.0	75.2	75.4	75.4
Vulnerable-System	94.0	92.9	96.0	95.4	94.9	94.1
Tool	71.0	85.4	92.0	93.1	80.1	89.0
Victim	87.1	91.1	89.4	90.6	88.2	90.8
Micro avg	84.3	85.9	85.2	86.4	84.8	86.0

POLICE have launched an investigation after a number of patients have had their personal details accessed by a member of staff at Crosshouse Hospital. Those affected, believed to be mainly women, were notified that information, including their phone numbers and addresses, had been accessed and that the member of staff who breached their privacy had ‘no legitimate clinical or administrative requirement’. The breach came from within information held in the Radiology Information System and patients affected had all recently been referred for an X-ray. A member of staff has been ‘excluded from work’ following the breach which took place between April and September of this year. Dr Alison Graham, Medical Director for the NHS Ayrshire & Arran said: “NHS Ayrshire & Arran has been made aware of a member of staff inappropriately accessing patient records. This individual is currently excluded from work." We are currently investigating and are contacting a number of patients to ascertain the extent of this breach. We wish to apologise to anyone affected by this. We take patient confidentially extremely serious and will ensure a full investigation is conducted. "We are working closely with Police Scotland and the Information Commissioner’s Office. As this is an ongoing police investigation, we are not able to confirm any further details."

Fig. 3. Original text example.

entities involved. Remember, each entity can only be one of the following: "Causal", "Effect", or not related to the event."

(2) `Cross-Event Entity Confusion`: When inputting the full threat intelligence passage, GPT sometimes included entities from unrelated events in its causal reasoning. This was due to the model being exposed to excessive context. To mitigate this, we constrained the model to focus only on the relevant event and associated entities during causal analysis. Prompt: "Please generate a causal analysis based on the events that occurred and the entities I provided. Do not include or consider any entities that are not among those provided. Remember, each entity can only be one of the following: "Causal", "Effect", or not related to the event."

(3) `Misclassification of Time Entities`: GPT occasionally classified temporal expressions such as `yesterday` or `March` as `Causal` or `Effect`, which are semantically inappropriate. To address this, we incorporated entity type information into the training data and explicitly instructed the model to consider

Just Accepted

Table 9. Comparison of GPT 4o Before and After Optimization

	Precision (%)	Recall (%)	F1-score (%)
Causal (before Optimization)	51.1	46.1	48.4
Causal (after Optimization)	61.0	55.3	58.0
Effect (before Optimization)	63.6	46.8	53.9
Effect (after Optimization)	67.4	53.4	59.6

entity types during classification. Prompt: "Please generate a causal analysis based on the events that occurred and the entities I provided, taking into account the types of entities to assist your judgment. Do not consider any other entities, and note that temporal-type entities should not be classified as “Causal” or “Effect.” Remember, each entity can only be one of the following: "Causal", "Effect", or not related to the event."

Although GPT was capable of performing causal inference with basic inputs, achieving consistent and accurate results required multiple rounds of iterative prompt design and testing. The final optimized prompt, which significantly improved output quality, was as follows:

“Please generate a causal analysis based on the events that occurred and the entities I provided. Consider both the types of entities and their positions within the context to assist your judgment. Do not include or consider any entities that are not among those provided, and note that temporal-type entities should not be classified as ’Causal’ or ’Effect.’ Remember, each entity must be assigned to only one of the following categories: ’Causal’, ’Effect’, or not related to the event. The output format must match the format used in the learning dataset.”

After this optimization, the model’s output quality improved substantially, as shown in Table 9.

5.3.2 Inference Model Experiments.

In comparative evaluations, the `o1` model exhibited superior reasoning capabilities compared to GPT-4o, even without extensive prompt tuning. It demonstrated a stronger understanding of the few-shot demonstration strategy and prompt structure, offering two notable advantages.

First, `o1` required fewer iterations of prompt refinement, thereby reducing the cost of manual intervention. Second, unlike standard GPT models that often rely on verbose prompts to achieve accurate inference, `o1` delivered comparable or better results using concise instructions. This efficiency translated to lower token usage and reduced API call costs.

Building on this, `o3-mini`, a refined version of `o1`, further improved performance. Using the same few-shot demonstration dataset and prompt framework, `o3-mini` achieved higher inference accuracy, greater output stability, and a more accurate understanding of causal structures. These results suggest that inference-optimized models such as `o3-mini` can perform reliably with less manual tuning and more consistent alignment with expected outputs, making them increasingly suitable for real-world causal inference applications.

Figures 4–7 illustrate the comparative outputs of GPT-4o, o1, and o3-mini alongside ground truth labels for the same example.

As shown in Figure 4, while GPT-4o successfully identified some causal relationships, its overall performance exhibited a notable number of misclassifications. For instance, entities that should have been labeled as `Effect` were incorrectly classified as `Causal`, and some causally related events were not recognized. These errors highlight the limitations of GPT-4o in accurately modeling event-entity relationships in complex contexts. Although large language models (LLMs) possess strong generalization capabilities, their causal reasoning remains constrained and may lead to inconsistent outputs.

Just Accepted

Table 10. Comparison of GPT-4o, o1 and o3-mini

	Precision (%)	Recall (%)	F1-score (%)
Causal (GPT-4o)	61.0	55.3	58.0
Causal (GPT o1)	83.4	77.3	80.2
Causal (GPT o3-mini)	85.5	77.3	81.2
Effect (GPT-4o)	67.4	53.4	59.6
Effect (GPT o1)	72.5	58.7	64.9
Effect (GPT o3-mini)	83.8	69.2	75.8

In comparison, GPT o1 demonstrated significantly improved alignment with the ground truth annotations. As shown in Figure 5, although some misclassifications persist, the overall error rate is lower. Moreover, o1 achieved this level of performance using only the few-shot demonstration strategy and minimal prompt engineering, indicating better instruction following and reasoning capability.

Finally, Figure 6 illustrates the output of o3-mini, the most advanced inference-optimized GPT model evaluated in this study. In this example, o3-mini's results were entirely consistent with the ground truth (Figure 7), representing the ideal outcome. This finding suggests that as LLM training and inference optimization techniques continue to advance, newer models are becoming increasingly capable of reliably distinguishing causal relationship types.

Throughout this study, we conducted a detailed comparison of GPT-4o, o1, and o3-mini on causal reasoning tasks, evaluating their respective strengths and limitations in identifying event-entity relationships. A quantitative summary of their performance is presented in Table 10.

Fig. 4. GPT-4o response for a representative causal inference example.

5.3.3 *Comparison with MLSA Extraction Model.*

We also evaluated the performance of our MLSA extraction model on the same test instance. As illustrated in Figure 8, the MLSA model achieves high causal inference accuracy. However, it occasionally suffers from errors in entity segmentation. For example, the phrase `["patient", "records"]` should have been treated as a

Just Accepted

```
"Causal": [
[],
[{"start": 51, "end": 54, "text": ["the", "member", "of", "staff"], "type": "Person"}],
[{"start": 17, "end": 20, "text": ["a", "member", "of", "staff"], "type": "Person"}],
[{"start": 77, "end": 80, "text": ["the", "Radiology", "Information", "System"], "type": "System"}],
[{"start": 93, "end": 96, "text": ["A", "member", "of", "staff"], "type": "Person"}],
[{"start": 142, "end": 145, "text": ["a", "member", "of", "staff"], "type": "Person"}]],

"Effect": [
[{"start": 44, "end": 44, "text": ["addresses"], "type": "PII"}, {"start": 40, "end": 42, "text": ["their", "phone", "numbers"], "type":
"PII"}, {"start": 37, "end": 37, "text": ["information"], "type": "Data"}, {"start": 32, "end": 32, "text": ["women"], "type":
"Person"}],
[],
[{"start": 12, "end": 14, "text": ["their", "personal", "details"], "type": "PII"}, {"start": 9, "end": 9, "text": ["patients"], "type":
"Person"}],
[{"start": 74, "end": 74, "text": ["information"], "type": "Data"}, {"start": 82, "end": 82, "text": ["patients"], "type": "Person"}],
[],
[{"start": 148, "end": 149, "text": ["patient", "records"], "type": "PII"}]]
```


The entity should be classified as "Effect" but was mistakenly identified as "Causal."

Fig. 5. GPT o1 response for the same causal inference example.


```
"Causal": [
[],
[{"start": 51, "end": 54, "text": ["the", "member", "of", "staff"], "type": "Person"}],
[{"start": 17, "end": 20, "text": ["a", "member", "of", "staff"], "type": "Person"}],
[],
[{"start": 93, "end": 96, "text": ["A", "member", "of", "staff"], "type": "Person"}],
[{"start": 142, "end": 145, "text": ["a", "member", "of", "staff"], "type": "Person"}]],

"Effect": [
[{"start": 44, "end": 44, "text": ["addresses"], "type": "PII"}, {"start": 40, "end": 42, "text": ["their", "phone", "numbers"], "type":
"PII"}, {"start": 37, "end": 37, "text": ["information"], "type": "Data"}, {"start": 32, "end": 32, "text": ["women"], "type":
"Person"}],
[],
[{"start": 12, "end": 14, "text": ["their", "personal", "details"], "type": "PII"}, {"start": 9, "end": 9, "text": ["patients"], "type":
"Person"}],
[{"start": 74, "end": 74, "text": ["information"], "type": "Data"}, {"start": 77, "end": 80, "text": ["the", "Radiology", "Information",
"System"], "type": "System"}, {"start": 82, "end": 82, "text": ["patients"], "type": "Person"}],
[],
[{"start": 148, "end": 149, "text": ["patient", "records"], "type": "PII"}]]
```


Fig. 6. GPT o3-mini response for the same causal inference example.

Just Accepted

single entity but was incorrectly split into two. Although both fragments were correctly labeled as `Effect`, this segmentation error could hinder the precision of downstream causal inference.

5.4 Computational Efficiency Analysis

Having established MLSA's performance advantages, we now examine its computational requirements for practical deployment considerations.

```
"Causal": [
[],
[{"start": 51, "end": 54, "text": ["the", "member", "of", "staff"]}],
[{"start": 17, "end": 20, "text": ["a", "member", "of", "staff"]}],
[],
[{"start": 93, "end": 96, "text": ["A", "member", "of", "staff"]}],
[{"start": 142, "end": 145, "text": ["a", "member", "of", "staff"]}]],

"Effect": [
[{"start": 44, "end": 44, "text": ["addresses"]}, {"start": 40, "end": 42, "text": ["their", "phone", "numbers"]}, {"start": 37, "end":
37, "text": ["information"]}, {"start": 32, "end": 32, "text": ["women"]}],
[],
[{"start": 12, "end": 14, "text": ["their", "personal", "details"]}, {"start": 9, "end": 9, "text": ["patients"]}],
[{"start": 74, "end": 74, "text": ["information"]}, {"start": 77, "end": 80, "text": ["the", "Radiology", "Information", "System"]},
{"start": 82, "end": 82, "text": ["patients"]}],
[],
[{"start": 148, "end": 149, "text": ["patient", "records"]}]]
```


Fig. 7. Ground truth labels for the causal inference example.


```
"Causal": [
[],
[{"start": 51, "end": 54, "text": ["the", "member", "of", "staff"]}],
[{"start": 17, "end": 20, "text": ["a", "member", "of", "staff"]}],
[],
[{"start": 93, "end": 96, "text": ["A", "member", "of", "staff"]}],
[{"start": 142, "end": 145, "text": ["a", "member", "of", "staff"]}]],

"Effect": [
[{"start": 44, "end": 44, "text": ["addresses"]}, {"start": 40, "end": 42, "text": ["their", "phone", "numbers"]}, {"start": 37, "end":
37, "text": ["information"]}, {"start": 32, "end": 32, "text": ["women"]}],
[],
[{"start": 12, "end": 14, "text": ["their", "personal", "details"]}, {"start": 9, "end": 9, "text": ["patients"]}],
[{"start": 74, "end": 74, "text": ["information"]}, {"start": 77, "end": 80, "text": ["the", "Radiology", "Information", "System"]},
{"start": 82, "end": 82, "text": ["patients"]}],
[],
[{"start": 148, "end": 148, "text": ["patient"]}, {"start": 149, "end": 149, "text": ["records"]}]]
```


"Patient records" were mistakenly classified as two separate entities.

Fig. 8. The response generated by the proposed MLSA framework.

Just Accepted

Training efficiency. Table 11 summarizes MLSA's runtime footprint. With a batch size of 8, the system processes reports with an average training time of **1.9 s** per report, corresponding to a throughput of **31.2** samples per minute. Peak GPU memory usage is modest at **5.3 GB**, leaving substantial headroom for larger batches or concurrent workloads on the same device. End-to-end, the full training-set (800 CTI news reports) pass completes in **6.59 min**, indicating that MLSA can be iterated rapidly and deployed efficiently.

Inference efficiency. Table 12 summarizes MLSA's inference efficiency on our test set (200 CTI news reports). On a single NVIDIA H100 (80 GB), MLSA achieves an average inference time of **5.08 s** per report, with peak GPU

Table 11. MLSA Training Efficiency

Metric	Value
Hardware	NVIDIA H100 (80GB)
Batch Size	8
Avg. Training Time per Report	1.9 s
Throughput	31.2 samples/min
Peak GPU Memory Usage	5.3 GB
Total Train Set Processing Time	6.59 min

Table 12. MLSA Inference Efficiency

Metric	Value
Hardware	NVIDIA H100 (80GB)
Batch Size	32
Avg. Inference Time per Report	5.08 s
Throughput	12 samples/min
Peak GPU Memory Usage	21.7 GB
Total Test Set Processing Time	16.56 min
Cost Comparison (200 reports):	
GPT-4o (API)	$8.29 total

memory of **21.7 GB** at batch size 32, yielding a throughput of **12** samples/minute and a total processing time of **16.56** minutes for the full test set.

For GPT-based inference, API-level metrics are heavily influenced by network conditions and rate limits, making them unreliable for deployment planning. We instead report token costs: processing the same 200-report set with GPT-4o costs approximately **$8.29** (2,994,001 input tokens; 81,849 output tokens). This cost-efficiency advantage makes MLSA particularly attractive for high-volume production deployments.

5.5 Failure Analysis and Ablation Studies

5.5.1 Error Pattern Analysis.

Next, as shown in Table 7, although our model outperforms CASIE [16] in overall event argument extraction, certain entity categories that are often closely associated with causal reasoning, such as Capability, Purpose, and Malware, still demonstrate relatively low recall. Since these entities are crucial for subsequent causal analysis, failures in their extraction can significantly impair the completeness of causal reasoning. Upon further examination of the error samples, we observed that most entities are relatively short (fewer than five tokens), whereas entities belonging to Capability or Purpose often exceed ten or even fifteen tokens. Long spans tend to increase the likelihood of extraction errors in sequence labeling tasks, and our experiments confirm this finding—the model faces greater difficulty recognizing long entities, leading to lower recall for these categories. In addition, Malware entities present unique challenges. Many malware names are proper nouns absent from BERT's pretraining corpus, meaning the model lacks sufficient prior knowledge to interpret them correctly. This vocabulary coverage gap limits its generalization capability and results in lower performance for Malware entity extraction.

To illustrate these error patterns, we provide concrete failure examples:

Case 1 - Entity Segmentation Error (Figure 8): The phrase "patient records" was incorrectly segmented into ["patient"] and ["records"]. Although both fragments were correctly labeled as Effect, the incomplete entity

Just Accepted

extraction prevented accurate causal relationship modeling. This demonstrates that boundary detection remains a critical challenge.

Case 2 - Role Misclassification (Figure 9): In the sentence "So far at least, almost all phishing attacks impacting this industry have involved only Google and DropBox.", the gold labels annotate "this industry," "Google," and "DropBox" as Effect. However, "Google" was misclassified as a Vulnerable-System rather than a Trusted-Entity, leading the model to mistakenly infer it as a Causal entity and thus reversing the causal direction. Table 8 shows that certain roles (such as Trusted-Entity) achieve only 79.2% F1-score, and such misclassifications directly result in causal inference failures.

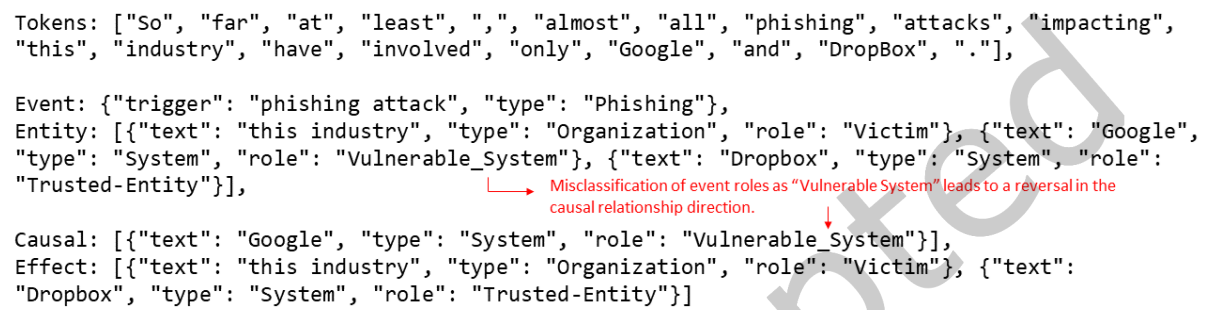

Fig. 9. Role Misclassification.

Case 3 - Unseen Malware Name (OOV) and NER Failure: When encountering previously unseen malware names such as "GoldenEye" or "WannaCry," BERT fails to recognize these entities due to out-of-vocabulary tokenization, leading to missed entity tags. Without successful entity recognition, no causal relationship can be inferred—even when the text explicitly indicates causality. This illustrates a core limitation of BERT-based methods in cybersecurity domains with rapidly evolving terminology. Figure 10 shows an error case where "GoldenEye" was not correctly recognized.

Case 4 – Failure to Extract Long Entities: When an entity span is excessively long, the model tends to struggle in determining its boundaries and core semantics. Such entities often contain multiple modifiers or embedded sub-entities, causing semantic dispersion and reducing extraction accuracy. In addition, overly long spans may exceed the model's effective attention range, leading to partial recognition or unintended segmentation into multiple entities, ultimately resulting in extraction failure or misclassification. Figure 11 illustrates this type of error.

5.5.2 Ablation Experiments.

Beyond these cases, we further analyzed other potential error sources that may affect causal inference performance. We conducted ablation experiments on the individual components and semantic features of the MLSA architecture to evaluate each module's contribution. Specifically, we tested the following four configurations:

Removing the Multi-Head Attention Layer: When this layer was removed, the model relied solely on BERT outputs. The average F1-score for causal relation classification dropped by about 8.9%, with the Causal class showing the largest decline. This demonstrates that the attention layer significantly enhances the model's ability to focus on critical causal cues.

Restricting Span Length to 1: This configuration simulates token-level judgment, preventing the model from recognizing multi-word entities. The results show a dramatic drop in recall to about 31.8%, with F1-scores of 13.7% for Causal and 10.5% for Effect. These results confirm that multi-span entities are essential for capturing complete semantic meaning in causal analysis.

Just Accepted

Fig. 10. Failure to identify malware.

Removing POS Features: Part-of-speech features provide syntactic information that helps the model distinguish event triggers from ordinary nouns. Without POS features, performance declines noticeably, indicating that syntactic cues contribute valuable support for semantic reasoning.

Removing Role Features: Event-role features enable the model to clearly determine the functional relationship of each entity. When these features are excluded, precision and related metrics drop sharply, confirming that role-level semantic information greatly enhances causal relation classification.

The causal inference results of these ablation experiments are summarized in Table 13.

Next, we further compared the results of GPT o3-mini and the MLSA extraction model, as shown in Table 13. In terms of precision, the GPT model performed comparably to the MLSA extraction model in identifying both causal and effect relationships, and even slightly outperformed it in certain cases. However, with respect to recall, the GPT model showed noticeably lower performance. This indicates that, although causal inference has been significantly improved through reinforcement and few-shot learning, GPT still faces limitations in accurately identifying the complete set of relevant entity relationships. In contrast, the MLSA extraction model, built upon the BERT architecture, maintains advantages in both the accuracy and completeness of causal relationship identification. Its balanced performance across precision and recall demonstrates higher overall effectiveness in structured inference tasks.

In summary, while GPT models demonstrate exceptional generalization capabilities and flexible reasoning across a wide range of scenarios, there remains room for improvement in the extraction and identification of specific causal relationships. By contrast, the advantages of the MLSA extraction model primarily stem from its deep learning of contextual information and strong annotation capabilities, particularly its stability in multi-level semantic analysis and structured information extraction. However, BERT-based models such as MLSA may still

Just Accepted

Fig. 11. The overlong entity was incorrectly segmented.

face challenges in generalization when applied to cross-domain or highly unstructured texts. Therefore, within the current research context, integrating pre-trained language models with MLSA extraction methods may lead to more effective and reliable structured causal inference. Table 13 shows that o3-mini achieves better results than the MLSA extraction model on the precision metric. This suggests that o3-mini is more effective at determining whether candidate entities form a Causal/Effect relation. To further improve the outputs of the MLSA extraction model, we applied o3-mini to the extracted results for an additional round of causal inference. The results confirm that this combined strategy leads to a slight improvement in overall performance.

We applied our developed event-entity causal relationship dataset to Trimf [18], a representative joint entity and relation extraction method, and conducted a comparative analysis as presented in Table 13. Trimf was selected as the baseline because it employs a trigger-sense memory flow framework that jointly addresses entity and relation extraction, making it the most closely aligned publicly available method for our task. The results indicate that, although Trimf was not specifically designed for causal relationship identification, it is still capable of producing reasonable inference outcomes to a certain extent. However, its overall performance remains substantially lower than that of the model proposed in this study. This finding further supports the effectiveness of our approach, particularly in its use of multi-head attention mechanisms combined with event role features. This integration enables the model to more precisely capture fine-grained semantic and structural information, thereby improving the accuracy and reliability of causal relationship identification.

6 Conclusion

In this paper, we introduced a novel research direction in the causal relationship analysis of Cyber Threat Intelligence (CTI) reports, emphasizing the importance of exploring causal links between events and entities. A major challenge in this area is the lack of publicly available datasets to support related research. To address this limitation, we constructed a causal relationship dataset based on cybersecurity events and entities annotated

Just Accepted

Table 13. Comparison of Causal Inference Results Between MLSA Extraction Model and GPT

		Precision (%)	Recall (%)	F1-score (%)
Causal	MLSA Model (Without Mutihead attention)	73.1	69.6	71.3
	MLSA Model (Span length set 1)	14.1	13.3	13.7
	MLSA Model ((Without POS)	76.8	75.6	76.2
	MLSA Model (Without Event Role)	76.1	73.7	74.9
	MLSA Model	80.0	78.1	79.0
	GPT o3-mini	85.5	77.3	81.2
	MLSA + GPT o3-mini	82.1	78.1	80.0
	Trimf[18]	71.6	70.7	71.2
Effect	MLSA Model (Without Mutihead attention)	73.6	74.2	73.8
	MLSA Model (Span length set 1)	10.8	10.2	10.5
	MLSA Model ((Without POS)	79.7	80.4	80.0
	MLSA Model (Without Event Role)	78.4	75.2	76.8
	MLSA Model	83.2	84.7	83.9
	GPT o3-mini	83.8	69.2	75.8
	MLSA + GPT o3-mini	83.9	84.7	84.4
	Trimf[18]	68.0	71.3	69.6

in CASIE [16]. This dataset provides a valuable resource for training and evaluating models, and serves as a foundation for future research on causal inference in cybersecurity contexts.

To achieve our research objectives, we proposed the MLSA framework, which is capable of simultaneously performing event extraction, event argument identification, and causal relationship inference. In addition, we incorporated event argument role features, which significantly enhance the model's ability to identify causal relationships. This integrated framework enables more efficient analysis of event causes and their affected entities, thereby improving the interpretation and response to cybersecurity incidents.

We further evaluated the causal reasoning capabilities of GPT-4o and inference-optimized models (`o1` and `o3-mini`) by providing few-shot demonstrations from our dataset and designing specialized prompts to guide the inference process. Through iterative refinements, we optimized response quality. Experimental results demonstrate that, even with few-shot learning and prompt engineering, GPT-4o's causal reasoning performance remains inferior to that of the professionally trained MLSA framework. However, `o1` and `o3-mini`, designed specifically for inference optimization, achieved performance that was comparable to or exceeded that of the MLSA model in certain evaluation metrics.

Although `o1` and `o3-mini` still fall short in some aspects compared to the MLSA framework, their overall performance suggests that, as large-scale language models continue to evolve, their potential in causal reasoning tasks is becoming increasingly evident. Moreover, compared to traditional BERT-based architectures, GPT models exhibit stronger generalization capabilities, making them more suitable for relationship extraction across diverse domains.

Just Accepted

From a practical deployment perspective, the proposed approaches naturally support human–AI collaboration in Security Operations Centres (SOCs). The MLSA framework, with its computational efficiency (Section 5.4) and structured output format, can be integrated into SIEM/SOAR pipelines to automatically generate event–entity causal annotations from incoming CTI reports, providing SOC analysts with pre-structured causal summaries for rapid review and validation rather than requiring them to manually parse lengthy reports. Meanwhile, GPT-based inference offers complementary capabilities for on-demand, interactive causal analysis, where analysts can query specific incidents and receive explanations of causal chains. In both cases, the human analyst retains the final decision-making authority, while the AI component accelerates the most time-consuming aspects of root cause analysis, thereby reducing analyst workload and response time.

While our study focuses on cybersecurity incidents, the same research direction can be extended to other domains where understanding event–entity causal relationships is critical, such as healthcare and finance. We hope our work inspires researchers in related fields to contribute to the development and enrichment of such resources.

References

[1] Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902* (2023).

[2] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. *arXiv preprint arXiv:1808.06876* (2018).

[3] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications* 114 (2018), 34–45.

[4] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, Vol. 2. Lisbon, 837–840.

[5] Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*. IOS Press, 2006–2013.

[6] Marius Hobbhahn, Tom Lieberum, and David Seiler. 2022. Investigating causal understanding in LLMs. In *NeurIPS ML safety workshop*.

[7] Liangyi Huang and Xusheng Xiao. 2024. Ctikg: Llm-powered knowledge graph construction from cyber threat intelligence. In *First Conference on Language Modeling*.

[8] ChangHomg Jiang. 2025. *Event Entities Causal*. https://github.com/aiden0828/Event_Entities_Causal Accessed: March 1, 2025.

[9] Vivek Khetan, Roshni Ramnani, Mayuresh Anand, Subhashis Sengupta, and Andrew E Fano. 2022. Causal bert: Language models for causality detection between events expressed in text. In *Intelligent Computing: Proceedings of the 2021 Computing Conference, Volume 1*. Springer, 965–980.

[10] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[11] Song Li, Yuxin Yang, and Liping Zhang. 2024. Bert-based with Short-range Dependency Enhancement for Relation Extraction. *Available at SSRN 4774199* (2024).

[12] Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1858–1869.

[13] Yeon-Ji Park, Min-a Lee, Geun-Je Yang, Soo Jun Park, and Chae-Bong Sohn. 2023. Web interface of NER and RE with BERT for biomedical text mining. *Applied Sciences* 13, 8 (2023), 5163.

[14] Bo Qiao, Zhuoyang Zou, Yu Huang, Kui Fang, Xinghui Zhu, and Yiming Chen. 2022. A joint model for entity and relation extraction based on BERT. *Neural Computing and Applications* (2022), 1–11.

[15] Tokala Yaswanth Sri Sai Santosh, Prantika Chakraborty, Sudakshina Dutta, Debarshi Kumar Sanyal, and Partha Pratim Das. 2021. Joint Entity and Relation Extraction from Scientific Documents: Role of Linguistic Information and Entity Types. *EEKE@ JCDL* 21 (2021), 15–19.

[16] Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8749–8757.

[17] Xiangrong She, Jianpeng Chen, and Gang Chen. 2022. Joint learning with BERT-GCN and multi-attention for event text classification and event assignment. *IEEE Access* 10 (2022), 27031–27040.

[18] Yongliang Shen, Xinyin Ma, Yechun Tang, and Weiming Lu. 2021. A Trigger-Sense Memory Flow Framework for Joint Entity and Relation Extraction. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1704–1715. doi:10.1145/3442381.3449895

Just Accepted

[19] Nan Sun, Ming Ding, Jiaojiao Jiang, Weikang Xu, Xiaoxing Mo, Yonghang Tai, and Jun Zhang. 2023. Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. *IEEE Communications Surveys & Tutorials* 25, 3 (2023), 1748–1774.
[20] Beth M Sundheim. 1991. Overview of the third message understanding evaluation and conference. In *Third Message Understanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991*.
[21] Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The causal news corpus: Annotating causal relations in event sentences from news. *arXiv preprint arXiv:2204.11714* (2022).
[22] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. Linguistic Data Consortium, LDC2006T06. Available from https://catalog.ldc.upenn.edu/LDC2006T06.
[23] Ying Wei, Lili Bo, Xiaobing Sun, Bin Li, Tao Zhang, and Chuanqi Tao. 2023. Automated event extraction of CVE descriptions. *Information and Software Technology* 158 (2023), 107178.
[24] Yongqing Yang and Siyuan Li. 2024. Entity overlapping relation extracting algorithm based on CNN and BERT. *IEEE Access* (2024).
[25] Hao-Ren Yao, Luke Breitfeller, Aakanksha Naik, Chunxiao Zhou, and Carolyn Rose. 2024. Distilling Multi-Scale Knowledge for Event Temporal Relation Extraction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2971–2980.
[26] Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt. *arXiv preprint arXiv:2304.05454* (2023).
[27] Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. 2023. Causal reasoning of entities and events in procedural texts. *arXiv preprint arXiv:2301.10896* (2023).
[28] Wenhao Zhang, Chunshan Wang, Huarui Wu, Chunjiang Zhao, Guifa Teng, Sufang Huang, and Zhen Liu. 2022. Research on the chinese named-entity–relation-extraction method for crop diseases based on bert. *Agronomy* 12, 9 (2022), 2130.
[29] Shan Zhao, Minghao Hu, Zhiping Cai, and Fang Liu. 2021. Modeling dense cross-modal interactions for joint entity-relation extraction. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 4032–4038.
[30] Yaqin Zhu, Xuhang Li, Zijian Wang, Jiayong Li, Cairong Yan, and Yanting Zhang. 2023. ER-LAC: Span-Based Joint Entity and Relation Extraction Model with Multi-Level Lexical and Attention on Context Features. *Applied Sciences* 13, 18 (2023), 10538.

Received 20 April 2025; revised 4 April 2026; accepted 6 May 2026

Just Accepted