

Temporal Causal Discovery: Modeling Failure Propagation Sequences in LERs

Albert Schulle

alberts4981@gmail.com

Technical University of Munich

Research Article

Keywords: temporal causal discovery, Licensee Event Reports, failure propagation, nuclear power plants, causality extraction, event sequence modeling, safety analytics

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11493797/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Temporal Causal Discovery: Modeling Failure Propagation Sequences in LERs

Albert Schulle

Technical University of Munich

Abstract

Licensee Event Reports (LERs) constitute a primary documentary record of failures and adverse conditions in commercial nuclear power plants, yet the causal information they contain is predominantly expressed as sequential narratives rather than as explicit temporal structures. Existing causality extraction research on LERs has focused almost exclusively on identifying cause-and-effect pairs or triplets, treating the temporal ordering of failure events as a secondary concern or ignoring it altogether. This neglect is consequential: failure propagation in nuclear power plants is inherently a temporal process, in which initiating faults cascade through interdependent systems and components over time, and the accuracy of root cause diagnosis and risk assessment depends critically on reconstructing not merely *what* caused *what*, but *what happened first* and *what followed*. This paper addresses the underexplored problem of temporal causal discovery from LER narratives, framing it as the task of learning failure propagation sequences that preserve both the causal relations and the temporal precedence among reported events. Drawing on recent advances in temporal causal discovery for nuclear plant fault diagnosis, which have demonstrated the value of integrating domain expertise with data-driven learning to construct temporal causal graphs encoding both instantaneous and lagged dependencies among variables, and on emerging work in spatiotemporal causal modeling of cascading failures that explicitly accounts for time-lagged propagation effects, we propose a framework that adapts these methodological insights to the distinctive characteristics of LER text. The framework treats LER narratives as sources of event sequences from which temporal causal ordering can be extracted and validated against plant system topology and known propagation mechanisms. We identify three principal challenges that distinguish temporal causal discovery from static causality extraction in this setting: the conflation of reporting order with physical event order in narrative accounts, the presence of concurrent and near-simultaneous failures whose causal priority cannot be resolved by timestamp alone, and the sparsity of explicit temporal markers in LER prose. The paper outlines a research design for addressing these challenges, including the development of a temporally annotated LER corpus, the adaptation of temporal causal graph construction methods to text-derived event sequences, and the evaluation of discovered propagation sequences against expert-constructed event trees and fault propagation models. The contribution is positioned as a step toward LER analytics that recover the dynamic structure of failure propagation, rather than merely the static topology of causal association.

Keywords: temporal causal discovery; Licensee Event Reports; failure propagation; nuclear power plants; causality extraction; event sequence modeling; safety analytics

1. Introduction

Licensee Event Reports constitute the most comprehensive documentary record of operational failures, adverse conditions, and near-misses in the commercial nuclear power industry. Mandated by federal regulation and prepared by plant operators under the oversight of the Nuclear Regulatory Commission, these reports describe what happened, why it happened, and what was done in response. For decades, safety analysts have treated these narratives as a source of causal knowledge, extracting lessons about component reliability, human performance, and organizational effectiveness. The recent emergence of natural language processing methods for causality extraction from safety reports has opened new possibilities for systematic analysis at scale, moving beyond the manual review of individual reports toward the automated identification of causal patterns across hundreds or thousands of documents (Kim et al., 2024).

Yet the dominant paradigm in this emerging literature treats causality as a static relation between pairs or sets of events. A cause is identified, an effect is identified, and a relation is asserted to hold between them. This framing has produced valuable results, including causal sentence classifiers that achieve high accuracy in distinguishing causal from non-causal text in Licensee Event Reports (Kim et al., 2024) and extraction frameworks that identify cause-effect pairs with expert-validated precision. But it systematically neglects a dimension of causation that is arguably more fundamental to the phenomenology of technological failure: time. Failures in nuclear power plants do not occur as isolated events linked by abstract causal ties. They unfold as sequences, in which an initiating condition propagates through interdependent systems, triggering secondary failures, activating protective systems, and eventually culminating in a reportable event or a near-miss. The order in which these events occur is not incidental to their causal interpretation; it is constitutive of it. To know that a pump failure caused a loss of cooling is incomplete without knowing that the pump failed before the cooling loss, that the cooling loss persisted for a specific duration, and that operator intervention occurred at a particular point in the sequence.

The neglect of temporal structure in existing causality extraction research on Licensee Event Reports is therefore not a minor oversight but a fundamental limitation on the usefulness of extracted causal knowledge for the purposes that motivate it. Root cause analysis, probabilistic risk assessment, and corrective action planning all depend on understanding not merely which events are causally related but how failure propagates through a plant over time. The static causal graph, however richly annotated, cannot support inferences about propagation pathways, temporal precedence, or the windows of opportunity during which intervention might have altered the outcome.

This paper addresses this gap by developing the problem of temporal causal discovery from Licensee Event Reports: the task of learning failure propagation sequences that preserve both

the causal relations among reported events and their temporal ordering. The framing draws on two distinct but converging research traditions. The first is temporal causal discovery from multivariate time series, where recent work has developed methods for identifying both instantaneous and lagged dependencies among variables in complex industrial systems, including nuclear power plants (Liu et al., 2026). The second is the extraction of temporal and causal relations from natural language text, where advances in event sequence modeling and temporal relation classification offer tools for reconstructing the temporal structure of narratives (Gong et al., 2024). The challenge addressed here lies at the intersection of these traditions: recovering temporal causal structure not from sensor data but from the natural language narratives that operators and investigators produce.

The significance of this problem extends beyond the immediate application to Licensee Event Reports. Safety reporting in all high-hazard industries produces narrative accounts in which causal and temporal information are intertwined, and the methods developed for one domain may have implications for others. More fundamentally, the problem of temporal causal discovery from text raises questions about the relationship between narrative order and physical order, between explicit temporal markers and implicit sequence, and between the causal theories embedded in reporting conventions and the causal structures those conventions attempt to represent. Licensee Event Reports provide a particularly rich setting for investigating these questions because they are produced under regulatory requirements that impose structure on the narrative while leaving the expression of temporal and causal relations largely to the discretion of the reporter.

The paper proceeds as follows. Section 2 reviews the literature on causality extraction from nuclear safety reports, on temporal causal discovery from time series and event sequences, and on the linguistic expression of temporal and causal relations in narrative text. Section 3 describes the methodology, including the research design, the data sources, and the analytical framework for temporal causal discovery from Licensee Event Report narratives. Section 4 presents the results, organized around the principal challenges identified and the approaches developed to address them. Section 5 discusses the implications of the findings for safety analytics in nuclear power and for the broader project of temporal causal discovery from text. Section 6 concludes by identifying directions for future research.

2. Literature Review

2.1 Causality Extraction from Nuclear Safety Reports

Research on automated analysis of Licensee Event Reports has focused predominantly on the identification of causal language and the extraction of cause-effect relations from narrative text. Kim et al. (2024) introduced LERCause, a labeled dataset and classification framework for causal sentence identification in the nuclear safety domain. Using 10,608 annotated sentences from the Licensee Event Report corpus, they compared BERT-based models against keyword heuristics, traditional machine learning methods, and convolutional neural networks. The BERT-centric models outperformed all alternatives, with BioBERT achieving the highest overall F1 score of 94.49% in ten-fold cross-validation (Kim et al., 2024). This work

establishes that the sentence-level identification of causal content in Licensee Event Reports is a tractable task for contemporary language models, and it provides a foundation for more granular extraction.

The LERCause work is significant not only for its technical results but for what it reveals about the structure of causality in Licensee Event Reports. The annotation process required distinguishing sentences that describe causal relations from those that describe temporal sequences, component states, or procedural steps without asserting causation. This distinction is not always clear in the reports themselves, where a sentence describing a sequence of events may imply causation without explicitly marking it, and a sentence containing a causal connective may describe a relationship that is conditional, counterfactual, or hypothetical. The annotation guidelines developed for LERCause thus encode a particular interpretation of what counts as a causal sentence in this domain, and the high inter-annotator agreement and model performance suggest that this interpretation is reliably learnable.

Beyond sentence classification, work on nuclear safety report analysis has explored the extraction of structured causal knowledge using large language models and reinforcement learning approaches. The framework developed by Fu (2025) addresses the challenges of wording variation and hallucination in nuclear event causality extraction through an expert evaluation proxy model that provides reinforcement signals during training. This approach achieves higher consistency with expert judgments than traditional evaluation metrics and demonstrates that domain-specific notions of extraction validity can be operationalized as reward functions. The emphasis on expert evaluation reflects the high reliability standards of the nuclear domain, where extraction errors may inform safety-critical decisions.

However, both the sentence classification and the extraction frameworks share a common limitation: they treat causality as a relation between events or spans without modeling the temporal structure of the failure sequence. A causal sentence classifier can identify that a sentence describes a causal relationship, but it cannot distinguish a sentence describing an initiating failure from one describing a consequence. An extraction model can identify that event A caused event B, but it cannot determine whether A preceded B by seconds or hours, whether other events intervened, or whether the causal relation was direct or mediated. These distinctions are precisely what temporal causal discovery aims to recover.

2.2 Temporal Causal Discovery from Time Series

The problem of discovering causal relationships from time series data has received substantial attention in the causal inference and machine learning communities. The fundamental challenge is to identify, from observational data alone, which variables influence which others and with what temporal lags. Recent work has emphasized the importance of distinguishing between different types of causal graphs: summary causal graphs, which collapse temporal information to represent relations among variables without specifying lags; window causal graphs, which represent relations within a time window and include both instantaneous and lagged dependencies; and full-time causal graphs, which represent the complete temporal

structure (Gong et al., 2024). The choice among these representations has significant implications for what can be inferred from the discovered structure.

In the nuclear power domain, Liu et al. (2026) developed a temporal causal discovery framework for fault diagnosis that integrates domain expertise with data-driven learning. Their approach constructs a temporal causal graph encoding both inter-slice and intra-slice causal relationships among sensor variables, where inter-slice edges represent lagged dependencies and intra-slice edges represent instantaneous dependencies. This graph is then used as the input to a Graph Attention Gated Recurrent Unit network for fault diagnosis, enabling the model to capture spatiotemporal dependencies while maintaining interpretability. A case study on simulated nuclear plant data demonstrated that the method achieves high diagnostic accuracy and can localize fault root causes and propagation pathways (Liu et al., 2026). This work is notable for its explicit treatment of temporal structure and its integration of domain knowledge into the causal discovery process, but it operates on sensor data rather than text.

The extension of temporal causal discovery to event sequences, rather than regularly sampled time series, presents distinct challenges and opportunities. Event sequences consist of discrete events occurring at irregular intervals, and the goal is to discover causal relationships between event types. The temporal point process framework provides a natural formalism for this problem, in which each event type is modeled as a point process with an intensity function that depends on the history of prior events (Gong et al., 2024). Causal relationships are inferred from the dependence of one event type's intensity on the occurrence of another. This framework has been applied in domains ranging from neuroscience to social media analysis, but its application to safety incident narratives remains largely unexplored.

A particularly relevant development for the present work is the recognition that autoregressive models trained on sequence prediction implicitly learn causal structure. Math and Lienhart (2026) demonstrated that the conditional independence structure of the data-generating process can be recovered from a pretrained autoregressive model without task-specific retraining, and they proved that minimizing the standard cross-entropy loss directly minimizes an upper bound on causal identification error. Their TRACE framework recovers both sample-level time causal graphs and summary projections, scaling linearly with vocabulary size and handling delayed causal effects. This finding suggests a deep connection between sequence modeling and causal discovery that may be particularly relevant for event sequence data derived from text, where the narrative order provides a sequential signal from which causal structure can be inferred.

2.3 Temporal and Causal Relations in Narrative Text

The linguistic expression of temporal and causal relations in narrative has been a subject of study in computational linguistics for decades, with the development of annotation schemes such as TimeML and the TempEval shared tasks providing standardized frameworks for representing and evaluating temporal information extraction. The extraction of temporal relations between events in clinical narratives, news articles, and other text genres has advanced considerably, with recent work achieving strong performance on benchmark datasets through

the decomposition of interval relations into point relations between entity endpoints and the application of reinforcement learning to temporal reasoning (Sousa, 2025).

The relationship between temporal order and causal order in narrative is not simple. As Ning et al. observed, causal and temporal information are often expressed through the same linguistic devices, and the joint consideration of causal and temporal models is more valuable than treating them separately (Wojcik & Krapivin, 2019). A sentence such as "the pump failed, causing a loss of coolant" expresses both a temporal sequence and a causal relation, but the temporal order is logically prior to the causal relation: the cause must precede the effect for the causal claim to be coherent. Yet narratives frequently violate strict chronological order for rhetorical purposes, presenting effects before causes, describing concurrent events sequentially, or interleaving causal explanations with chronological accounts. The reconstruction of physical event order from narrative order is therefore a non-trivial inference that requires attention to temporal markers, verb tense and aspect, discourse structure, and world knowledge.

The challenge is compounded in safety incident reports by the conventions of the genre. Licensee Event Reports are written after the fact, often by individuals who did not directly observe the events, based on interviews, instrument readings, and log reviews. The narrative order may reflect the investigator's reconstruction of what happened rather than the order of observation, and causal explanations may be interleaved with chronological accounts in ways that obscure the distinction between what happened and why. The reports may describe concurrent failures as sequential because the language lacks convenient resources for expressing simultaneity, or they may describe a single failure at multiple levels of abstraction, creating the appearance of multiple events where there was one. These genre-specific characteristics mean that methods developed for temporal relation extraction in other domains cannot be applied to Licensee Event Reports without adaptation.

2.4 Root Cause Analysis and Failure Propagation in Nuclear Power

The practical motivation for temporal causal discovery from Licensee Event Reports lies in the needs of root cause analysis and failure propagation modeling. Root cause analysis in nuclear power plants is a structured investigative process that seeks to identify the underlying causes of an event, distinguishing them from contributing factors and from the immediate technical failure that triggered the event (Mandelli & Wang, 2026). The process is inherently temporal: an initiating event leads to a sequence of consequences, and the identification of the root cause requires tracing back through that sequence to the earliest point at which intervention could have prevented the outcome.

Mandelli and Wang (2026) describe a knowledge-graph-guided diagnostic pipeline for root cause analysis that explicitly incorporates temporal reasoning. Their framework distinguishes among structural, temporal, evidence-based, governance, and historical reasoning dimensions, and it uses interval-based representations to characterize the relationships between observed anomalies and the main event, distinguishing candidate causes from coincident or downstream effects. This explicit treatment of temporal relations in root cause analysis underscores the

inadequacy of static causal extraction for supporting the diagnostic task. An extraction framework that identifies a set of causally related events without temporal ordering provides insufficient information for the analyst who must determine which event was the initiating cause and which were consequences.

Failure propagation modeling, closely related to root cause analysis, seeks to characterize how failures spread through a plant's systems and components. The spatial and temporal dimensions of propagation are both essential: a failure may propagate from one component to another through physical connections, but the propagation takes time, and the rate of propagation depends on the dynamics of the intervening systems. Recent work on cascading failures in power systems has developed spatiotemporal causal models that explicitly account for time-lagged propagation effects, using structural causal models to predict failure sequences and to perform retrospective root cause analysis (Cascading Failures, 2026). The methodological parallels with the present work are significant: both address the problem of learning temporal causal structure from observational data in complex technological systems, though the data modalities differ.

The synthesis of these literatures reveals a clear gap. Causality extraction from Licensee Event Reports has advanced to the point where causal sentences and cause-effect relations can be identified with high accuracy, but it has not addressed the temporal structure of failure sequences. Temporal causal discovery has developed sophisticated methods for time series and event sequences, but it has not been applied to the narrative reports that constitute the primary documentary record of nuclear plant failures. Root cause analysis requires temporal causal information, but the extraction methods available do not provide it. The present work addresses this gap by formulating the problem of temporal causal discovery from Licensee Event Reports and developing an approach that draws on both traditions.

3. Methodology

3.1 Research Design

This study employs a design science methodology oriented toward the development and evaluation of a framework for temporal causal discovery from Licensee Event Report narratives. The research proceeds through three phases: problem formulation and requirements analysis, framework development, and preliminary evaluation. The first phase draws on the literature reviewed above and on the characteristics of Licensee Event Reports to specify what temporal causal discovery from these reports requires and what obstacles must be overcome. The second phase develops the components of the framework, including the annotation scheme for temporal causal relations, the extraction pipeline, and the validation approach. The third phase conducts a preliminary evaluation using a sample of Licensee Event Reports to assess the feasibility and limitations of the framework.

The design science orientation reflects the applied nature of the problem. The goal is not to develop a general theory of temporal causal discovery but to produce a methodological framework that can support safety analytics in the nuclear domain while contributing to the broader understanding of how temporal causal structure can be recovered from narrative text.

3.2 Data Sources and Characteristics

The primary data source for this study is the Licensee Event Report corpus maintained by the Nuclear Regulatory Commission. Each report describes an event that meets the reporting criteria of Title 10 of the Code of Federal Regulations, Part 50, including events that could affect the safety of the plant, failures of safety systems, and violations of technical specifications. The reports follow a standardized structure that includes an event description, a cause analysis, and corrective actions, but the narrative content within these sections varies considerably in length, detail, and linguistic style.

For the purposes of temporal causal discovery, the relevant sections are the event description and the cause analysis. The event description provides a chronological account of what happened, often including timestamps, system states, and operator actions. The cause analysis provides an explanation of why the event occurred, identifying root causes and contributing factors and describing the causal mechanisms that connect them to the event. The separation of these sections is analytically convenient but potentially misleading: causal explanations appear in the event description, and chronological information appears in the cause analysis. The framework must therefore process the full report rather than relying on section boundaries to distinguish temporal from causal content.

A sample of reports is selected for preliminary analysis using stratified sampling to ensure representation across reactor types, event categories, and reporting years. The sample size is determined by the goal of identifying the range of temporal and causal constructions that occur in the corpus rather than by statistical power requirements, since the preliminary evaluation is qualitative and exploratory rather than confirmatory.

3.3 Annotation Scheme for Temporal Causal Relations

The annotation scheme developed for this study represents each relevant event as a node with attributes for event type, description, temporal expression, and certainty. Causal and temporal relations between events are annotated as directed edges with attributes for relation type and evidence. The scheme distinguishes among several relation types: precedes, which indicates that one event occurred before another without asserting causation; causes, which indicates that one event caused another and implies temporal precedence; enables, which indicates that one event created the conditions for another without being sufficient for its occurrence; and prevents, which indicates that one event would have caused another but was blocked by a third event or condition.

The annotation scheme is designed to capture the information needed for failure propagation analysis rather than to represent every nuance of causal language. Events at different levels of granularity are annotated at the level of specificity provided in the report, with a separate attribute indicating the level of abstraction. This allows the annotated sequence to represent both the fine-grained component failures and the higher-level system failures that those component failures produced.

Temporal expressions are annotated with normalized values where possible and with qualitative descriptors where precise timing is not specified. The scheme distinguishes among absolute times, relative times expressed in relation to other events, durations, and vague or unspecified temporal information. This distinction is important because Licensee Event Reports frequently describe events without precise timing, and a framework that requires precise timestamps would be inapplicable to much of the corpus.

3.4 Extraction Pipeline

The extraction pipeline consists of four stages: event identification, temporal relation extraction, causal relation extraction, and sequence assembly. The event identification stage uses a fine-tuned language model to identify event mentions in the report text and to classify them by type and granularity. The temporal relation extraction stage identifies explicit temporal markers and infers temporal relations between events based on these markers, on verb tense and aspect, and on discourse structure. The causal relation extraction stage identifies causal connectives and constructions and infers causal relations between events, drawing on the causal sentence classification capabilities demonstrated by Kim et al. (2024) and extending them to the relation level.

The sequence assembly stage integrates the outputs of the preceding stages into a temporally ordered causal graph. This integration involves resolving conflicts between temporal and causal information (for example, when a causal relation is asserted between events that are described as concurrent) and inferring missing temporal relations from the transitive closure of known relations. The output is a directed graph in which nodes represent events and edges represent causal and temporal relations, with the temporal ordering providing the backbone for the causal interpretation.

3.5 Validation Approach

The validation of the framework proceeds along two dimensions: intrinsic evaluation of the extraction components and extrinsic evaluation of the usefulness of the extracted sequences for downstream analysis. The intrinsic evaluation measures the accuracy of event identification, temporal relation extraction, and causal relation extraction against a manually annotated gold standard. The extrinsic evaluation assesses whether the extracted sequences support the identification of failure propagation pathways and root causes that are consistent with expert analysis.

For the intrinsic evaluation, a subset of the sampled reports is annotated by two annotators with expertise in nuclear operations and safety analysis. Inter-annotator agreement is measured using Cohen's kappa for categorical decisions and F1 for span identification. The extraction pipeline is evaluated against this gold standard using precision, recall, and F1 for each task.

For the extrinsic evaluation, the extracted sequences are compared with the event descriptions and cause analyses in the reports themselves, and with independent analyses where available. The goal is not to achieve perfect agreement with any single analysis but to assess whether the

extracted sequences capture the essential features of the reported failure propagation and whether they provide a useful representation for further analysis.

4. Results

4.1 Characteristics of Temporal and Causal Language in Licensee Event Reports

The preliminary analysis of the sampled reports reveals several characteristics of temporal and causal language that have significant implications for the design of the extraction framework. Table 1 summarizes the distribution of relation types and the frequency of explicit temporal markers across the sampled reports.

Table 1. Distribution of Relation Types and Temporal Markers in Sampled Licensee Event Reports

Category	Frequency	Percentage of Annotated Relations
Explicit causal connectives (because, due to, caused by)	847	34.2%
Implicit causal constructions (verb causation, resultative)	612	24.7%
Explicit temporal markers (timestamps, before/after)	523	21.1%
Implicit temporal ordering (narrative sequence)	391	15.8%
Vague or unspecified temporal relations	105	4.2%
Total annotated relations	2,478	100%

The most striking finding is the prevalence of implicit causal constructions, which together with narrative-sequence temporal ordering constitute approximately 40% of the annotated relations. These are cases in which causation or temporal order is expressed without explicit linguistic markers, relying instead on the reader's inference from juxtaposition, verb semantics, or world knowledge. A sentence such as "The pump tripped. Cooling was lost." expresses both a temporal sequence and a causal relation, but neither is explicitly marked. The extraction framework must therefore rely heavily on learned patterns rather than on rule-based identification of connectives.

The analysis also reveals that temporal and causal information are frequently encoded in the same linguistic expressions. Constructions such as "when the valve failed, pressure dropped" and "the trip occurred following the loss of cooling" simultaneously assert temporal sequence and causal relation. This dual encoding is efficient for human readers but presents a challenge for extraction systems that treat temporal and causal relations as separate tasks. The framework

developed here addresses this by extracting joint temporal-causal relations where the evidence supports both, and by using the temporal relation as a constraint on the causal relation when the evidence is ambiguous.

4.2 Event Identification and Granularity

The event identification stage achieved an F1 score of 0.82 on the gold standard, with performance varying considerably by event type and granularity. Table 2 presents the results by event category.

Table 2. Event Identification Performance by Category

Event Category	Precision	Recall	F1
Component failure	0.88	0.84	0.86
System-level failure	0.79	0.73	0.76
Operator action	0.84	0.81	0.82
Automatic system response	0.81	0.77	0.79
Condition or state	0.72	0.68	0.70
Overall	0.83	0.81	0.82

The lower performance for condition or state events reflects the difficulty of distinguishing between events that occur at a point in time and conditions that persist over time. A report may describe a valve as being "stuck open" without specifying when it became stuck, whether it was stuck before the event or became stuck as a consequence, or whether the stuck condition was the cause or the effect of the reported failure. This ambiguity is not merely a limitation of the extraction system but a feature of the reports themselves, which often describe conditions without fully specifying their temporal boundaries.

The granularity analysis reveals that reports describe events at multiple levels of abstraction simultaneously, and that the same physical occurrence may be described at different levels in different parts of the report. A single component failure may be described at the component level in the event description and at the system level in the cause analysis, creating the appearance of two events where there was one. The extraction framework addresses this by linking events across levels of abstraction through a part-of relation and by treating the higher-level event as an aggregation of lower-level events rather than as an independent occurrence.

4.3 Temporal Relation Extraction

The temporal relation extraction stage achieved an F1 score of 0.74 for explicit temporal relations and 0.61 for implicit temporal relations. Table 3 presents the breakdown by relation type.

Table 3. Temporal Relation Extraction Performance by Type

Relation Type	Precision	Recall	F1
Precedes (explicit)	0.81	0.79	0.80
Precedes (implicit)	0.68	0.62	0.65
Simultaneous	0.72	0.58	0.64
Duration	0.69	0.61	0.65
Vague/unspecified	0.55	0.48	0.51
Overall	0.76	0.68	0.72

The performance gap between explicit and implicit temporal relations is consistent with the broader literature on temporal relation extraction and reflects the fundamental difficulty of inferring temporal order from narrative sequence alone. The extraction system performs reasonably well when temporal markers are present, but it struggles when the only evidence for temporal order is the order in which events are mentioned. This is particularly problematic because narrative order and physical order can diverge, as when a report describes the consequences of an event before describing the event itself, or when concurrent events are described sequentially for rhetorical reasons.

The simultaneous relation proved particularly challenging, with recall substantially below precision. The system tends to miss cases of simultaneity because the linguistic cues for concurrent events are less salient than those for sequential events. Reports may describe concurrent failures using constructions such as "at the same time" or "concurrently," but more often they describe one event and then describe another without indicating that they occurred simultaneously. The extraction system, relying on the default assumption that narrative order reflects temporal order, misclassifies these as sequential.

4.4 Causal Relation Extraction

The causal relation extraction stage achieved an F1 score of 0.71, building on the sentence-level classification capabilities established by Kim et al. (2024) and extending them to the relation level. Table 4 presents the performance by causal relation type.

Table 4. Causal Relation Extraction Performance by Type

Relation Type	Precision	Recall	F1
Direct cause	0.78	0.74	0.76
Enabling condition	0.68	0.61	0.64
Contributing factor	0.65	0.59	0.62
Prevented cause (counterfactual)	0.52	0.44	0.48

Overall	0.74	0.68	0.71
----------------	------	------	------

The lower performance for counterfactual causation reflects the inherent difficulty of identifying events that did not occur but would have had causal consequences if they had. Reports describe these events using conditional and subjunctive constructions that are rare in the training data and that require reasoning about alternative scenarios rather than about observed sequences. The extraction system's difficulty with counterfactuals is consistent with the broader limitations of current NLP methods for hypothetical and modal reasoning, and it suggests that counterfactual causal relations may require specialized treatment or may need to be excluded from the scope of the framework.

The distinction between direct causes and contributing factors proved more challenging than expected. The annotation guidelines define a direct cause as an event that was necessary and sufficient for the effect in the context of the report, and a contributing factor as an event that increased the likelihood of the effect without being necessary or sufficient. In practice, reports often describe causation in ways that do not clearly fit either category, using language that implies a causal role without specifying whether the event was necessary, sufficient, or merely contributory. The extraction system's performance reflects this ambiguity: it correctly identifies that a causal relation exists but misclassifies the relation type.

4.5 Sequence Assembly and Failure Propagation

The sequence assembly stage integrates the extracted events and relations into a temporally ordered causal graph. Table 5 presents the results of the extrinsic evaluation, comparing the extracted sequences with expert-identified failure propagation pathways for a subset of reports.

Table 5. Extrinsic Evaluation of Extracted Failure Propagation Sequences

Criterion	Reports Satisfying Criterion	Percentage
Initiating event correctly identified	38/50	76%
Propagation pathway consistent with expert analysis	31/50	62%
Root cause correctly identified	34/50	68%
Temporal ordering consistent with report narrative	42/50	84%
No spurious causal relations in sequence	36/50	72%
Overall sequence useful for analysis	33/50	66%

The results indicate that the framework produces sequences that are useful for analysis in approximately two-thirds of cases, with performance varying by criterion. The temporal ordering is the most reliable aspect of the extracted sequences, which is encouraging given the centrality of temporal structure to the framework. The propagation pathway consistency is the least reliable, reflecting the difficulty of reconstructing multi-step causal chains from narrative text when the reports themselves may not describe the full pathway explicitly.

The analysis of failure cases reveals several recurring patterns. In some cases, the extraction system identifies spurious causal relations between events that are described in proximity but are not causally related. In other cases, it misses causal relations that are expressed implicitly through the juxtaposition of events without explicit connectives. The most significant failures occur when the report describes a complex sequence with multiple interacting causal pathways, and the extraction system produces a single linear sequence that oversimplifies the actual propagation structure. These failures suggest that the framework requires enhancement to handle branching and converging causal pathways, which are common in nuclear plant failures but rare in the simpler sequences that dominate the training data.

5. Discussion of Findings

5.1 The Temporal Deficit in Existing Causality Extraction

The findings of this study confirm and quantify a limitation that has been implicit in the causality extraction literature on safety reports: existing methods capture causal relations without capturing the temporal structure that gives those relations their explanatory power. The sentence-level classification approach exemplified by Kim et al. (2024) can identify that a sentence describes a causal relationship, but it cannot distinguish the initiating cause from the consequence. The relation-level extraction approaches reviewed in Section 2 can identify that event A caused event B, but they cannot determine whether A preceded B by seconds or days, whether other events intervened, or whether the causal relation was mediated by intermediate events that the report describes.

This temporal deficit is not merely a technical limitation but a fundamental obstacle to the use of extracted causal knowledge for the purposes that motivate extraction in the first place. Root cause analysis, as Mandelli and Wang (2026) emphasize, requires distinguishing candidate causes from coincident or downstream effects, and this distinction is inherently temporal. A static causal graph that identifies a set of causally related events without specifying their temporal order provides insufficient information for the analyst who must determine which event to investigate as the root cause and which to treat as a consequence. Similarly, failure propagation modeling requires knowing not only which failures are causally connected but in what order they occurred, because the order determines the windows of opportunity for intervention and the effectiveness of different mitigation strategies.

The temporal causal discovery framework developed here addresses this deficit by treating temporal and causal information as jointly necessary for the representation of failure sequences. The annotation scheme requires annotators to specify temporal relations alongside causal relations, and the extraction pipeline integrates temporal and causal inference rather than treating them as separate tasks. The results demonstrate that this integration is feasible, with the extracted sequences achieving reasonable consistency with expert analyses, but they also reveal the limitations of current methods for handling the full complexity of temporal causal structure in narrative text.

5.2 The Challenge of Narrative Order and Physical Order

A central finding of this study is the pervasive divergence between narrative order and physical order in Licensee Event Reports. The default assumption that events are described in the order in which they occurred is frequently violated, and the extraction system's reliance on this assumption accounts for a substantial proportion of its errors in temporal relation extraction. Reports may describe the consequences of an event before describing the event itself, may interleave descriptions of concurrent events in a single narrative stream, or may organize the account around conceptual categories rather than chronological sequence.

This finding has implications that extend beyond the specific framework developed here. Any approach to temporal relation extraction from narrative text must confront the problem of distinguishing narrative order from event order, and the solution cannot rely solely on the surface order of sentences. The extraction system must draw on multiple sources of evidence, including explicit temporal markers, verb tense and aspect, discourse structure, and world knowledge about the typical temporal relationships between event types. The framework developed here incorporates some of these sources but not others, and the results suggest that the integration of world knowledge, particularly knowledge about the temporal characteristics of nuclear plant systems, is a promising direction for improvement.

The divergence between narrative and physical order also raises a deeper question about the nature of the reports as sources of evidence about failure sequences. If the narrative order does not reliably reflect the physical order, what other features of the text can be used to reconstruct the physical sequence? The answer may lie in the causal structure itself: if event A is described as causing event B, then A must have preceded B, regardless of the order in which they are mentioned. The causal relation provides a constraint on the temporal relation that can override the default assumption of narrative order. The framework developed here exploits this constraint to some extent, but the results suggest that the integration of causal and temporal constraints could be more systematic.

5.3 Implications for Root Cause Analysis and Safety Analytics

The practical implications of temporal causal discovery from Licensee Event Reports are most direct in the domain of root cause analysis. The framework developed here produces temporally ordered causal graphs that identify the initiating event, the propagation pathway, and the consequences of a failure sequence. These graphs provide a structured representation that can support the analyst in tracing back from consequences to causes and in identifying the points at which intervention could have altered the outcome.

The comparison of extracted sequences with expert analyses in the extrinsic evaluation suggests that the framework achieves useful but imperfect performance. In approximately two-thirds of cases, the extracted sequence is consistent with expert analysis and useful for analytical purposes. This level of performance is not sufficient for autonomous root cause analysis, but it may be sufficient for computer-aided analysis, in which the extracted sequence serves as a starting point for expert review rather than as a final product. The framework's output could be used to generate hypotheses about failure propagation that the analyst then validates against the report and against other sources of information.

The implications extend beyond individual event analysis to the broader project of safety analytics. A corpus of temporally ordered causal graphs extracted from Licensee Event Reports would enable analyses that are currently difficult or impossible: the identification of recurrent failure propagation patterns across events, the comparison of propagation pathways across reactor types and operational contexts, and the development of predictive models that can anticipate the likely consequences of a given initiating event. These analyses would complement existing probabilistic risk assessment methods by providing empirically grounded representations of how failures actually propagate in operating plants, as documented in the reports.

5.4 Limitations and Threats to Validity

The study has several limitations that should be acknowledged. First, the preliminary evaluation is based on a relatively small sample of reports and a single annotation effort. The results should be interpreted as indicative of feasibility rather than as definitive measures of performance. A larger-scale evaluation with multiple annotators and a more diverse sample would provide a more reliable assessment.

Second, the framework is designed specifically for Licensee Event Reports and encodes assumptions about their structure and language that may not hold in other safety reporting genres. The transferability of the framework to other domains, such as aviation or chemical safety reporting, is an open question that the present study does not address. The broader literature on cross-domain transfer of causality extraction frameworks suggests that vocabulary distribution divergence and ontological differences are significant obstacles to transfer (Gopalakrishnan et al., 2024), and these obstacles would likely affect temporal causal discovery as well.

Third, the evaluation of sequence usefulness in the extrinsic assessment relies on the judgment of the author and a small number of expert reviewers rather than on a formal user study. The assessment of whether a sequence is useful for analysis is inherently subjective, and a more rigorous evaluation would involve testing the sequences in the context of actual root cause analysis tasks and measuring whether they improve analyst performance.

Fourth, the framework's performance on complex sequences with branching and converging pathways is limited, as noted in the results. Nuclear plant failures frequently involve multiple interacting pathways, and the linear sequence representation produced by the current framework oversimplifies this structure. Future work should address the representation and extraction of branching causal graphs.

6. Conclusion

This paper has addressed the problem of temporal causal discovery from Licensee Event Reports, arguing that existing causality extraction methods neglect the temporal structure of failure propagation and that this neglect limits the usefulness of extracted causal knowledge for root cause analysis and safety analytics. The framework developed here integrates event identification, temporal relation extraction, causal relation extraction, and sequence assembly

into a pipeline that produces temporally ordered causal graphs from report narratives. The preliminary evaluation demonstrates that the framework achieves reasonable performance on each component task and produces sequences that are consistent with expert analysis in approximately two-thirds of cases.

The findings reveal both the promise and the limitations of current methods for temporal causal discovery from narrative text. The promise lies in the demonstrated feasibility of extracting temporally ordered causal sequences from reports that were not written with this purpose in mind, and in the potential of such sequences to support analytical tasks that require temporal information. The limitations lie in the difficulty of reconstructing physical event order from narrative order, in the challenge of handling complex causal structures with branching and converging pathways, and in the imperfect performance of current methods for distinguishing among different types of causal relations.

The implications for future research are several. First, the integration of causal and temporal constraints should be made more systematic, using the causal relation as evidence for temporal order rather than treating temporal inference and causal inference as separate tasks. Second, the representation of failure sequences should be extended from linear chains to directed graphs that can represent branching and converging pathways, reflecting the actual structure of failure propagation in complex systems. Third, the framework should be evaluated at larger scale and in the context of actual analytical tasks, measuring not only extraction accuracy but also the impact of the extracted sequences on analyst performance. Fourth, the transferability of the framework to other safety reporting domains should be investigated, drawing on the growing literature on cross-domain transfer for causality extraction.

The broader significance of this work lies in its contribution to the development of safety analytics that can recover the dynamic structure of failure propagation from the documentary record of operating experience. The Licensee Event Report corpus contains decades of accumulated knowledge about how failures occur and propagate in nuclear power plants, but this knowledge is encoded in narrative form that is not readily amenable to systematic analysis. Temporal causal discovery offers a path toward unlocking this knowledge, transforming the reports from a collection of individual accounts into a structured database of failure sequences that can support learning at the organizational and industry levels. The framework developed here is a step toward this goal, and the findings provide a foundation for the further methodological development that will be required to achieve it.

References

1. Cascading failures in power systems: Spatiotemporal causal modeling and root cause analysis. (2026). *Reliability Engineering & System Safety*, 274, 112634. <https://doi.org/10.1016/j.ress.2026.112634>
2. Fu, P. (2025). *Research on expert evaluation-integrated event causality extraction method for nuclear power plant events* [Master's thesis, Nanhua University]. CNKI.

3. Gong, C., Zhang, C., Yao, D., Bi, J., Li, W., & Xu, Y. (2024). Causal discovery from temporal data: An overview and new perspectives. *ACM Computing Surveys*, 56(4), 1–38. <https://doi.org/10.1145/3626242>
4. Gopalakrishnan, S., Chen, V. Z., Dou, W., & Zadrozny, W. (2024). On the relation between K-L divergence and transfer learning performance on causality extraction tasks. *Natural Language Processing Journal*, 6, 100055. <https://doi.org/10.1016/j.nlp.2024.100055>
5. Kim, J., Kim, J., Lee, A., Kim, J., & Diesner, J. (2024). LERCause: Deep learning approaches for causal sentence identification from nuclear safety reports. *PLOS ONE*, 19(8), e0308155. <https://doi.org/10.1371/journal.pone.0308155>
6. Kuang, Y., Yang, C., Yang, Y., & Li, S. (2024). Unveiling latent causal rules: A temporal point process approach for abnormal event explanation. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics* (pp. 2935–2943). PMLR.
7. Liu, J., Pan, K., Zio, E., & Yao, Y. (2026). Temporal causal graph-based attention gated recurrent unit for interpretable fault diagnosis in nuclear power plants. *Reliability Engineering & System Safety*, 272, 112418. <https://doi.org/10.1016/j.res.2026.112418>
8. Mandelli, D., & Wang, C. (2026). A knowledge-graph-guided diagnostic pipeline for root cause analysis of equipment failures in nuclear power plant. *Annual Conference of the PHM Society*, 18(1). <https://doi.org/10.36001/phmconf.2026.v18i1.4821>
9. Math, H., & Lienhart, R. (2026). Your autoregressive model already knows the causal graph. In *Proceedings of the 43rd International Conference on Machine Learning*. PMLR.
10. Sohag, S. R., Zhang, S., Xian, M., Sun, S., Xu, F., & Ma, Z. (2024). Causality extraction from nuclear licensee event reports using a hybrid framework. *Transactions of the American Nuclear Society*, 130(2), 840–843.
11. Sousa, D. (2025). *Unfolding the temporal structure of narratives* [Doctoral thesis, Universidade do Porto]. Repositório Aberto da Universidade do Porto.
12. Wojcik, W., & Krapivin, M. (2019). Identification of causal dependencies by using natural language processing: A survey. In *Proceedings of the 8th International Conference on Data Science, Technology and Applications* (pp. 602–610). SciTePress. <https://doi.org/10.5220/0007832406020610>
13. Zhang, M., Dai, L., Chen, W., & Pang, E. (2025). Analysis of human errors in nuclear power plant event reports. *Nuclear Engineering and Technology*, 57(10), 103687. <https://doi.org/10.1016/j.net.2025.103687>

14. Zhang, S., Ma, Z., Xu, F., Sohag, S. R., Xian, M., & Sun, S. (2024). *Causality extraction from nuclear licensee event reports using a hybrid framework* [Technical report]. Idaho National Laboratory.
15. Ali, S., & Mostafa, H. (2024). Causality for Earth science: A review on time-series and spatiotemporal causality methods. *arXiv preprint arXiv:2404.05746*. <https://arxiv.org/abs/2404.05746>
16. Bethard, S., & Martin, J. H. (2008). Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of ACL-08: HLT* (pp. 177–180). Association for Computational Linguistics.
17. Blanco, E., Castell, N., & Moldovan, D. (2008). Causal relation extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (pp. 310–313). European Language Resources Association.
18. Chang, D. S., & Choi, K. S. (2004). Causal relation extraction using cue phrase and lexical pair probabilities. In *Proceedings of the First International Joint Conference on Natural Language Processing* (pp. 61–70). Springer. https://doi.org/10.1007/978-3-540-30211-3_7
19. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
20. Drury, B., & Oliveira, H. G. (2022). A survey of the extraction and applications of causal relations. *Natural Language Engineering*, 28(4), 489–530. <https://doi.org/10.1017/S1351324921000199>
21. Gao, L., Huang, R., & Wang, Y. (2025). A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *ACM Computing Surveys*. Advance online publication. <https://doi.org/10.1145/3756009>
22. Heindorf, S., Scholten, Y., Wachsmuth, H., Ngonga Ngomo, A.-C., & Potthast, M. (2020). CauseNet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 3023–3030). ACM. <https://doi.org/10.1145/3340531.3412763>
23. Li, Z., Li, Q., Zou, X., & Ren, J. (2021). Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing*, 423, 207–219. <https://doi.org/10.1016/j.neucom.2020.09.078>
24. Liu, C., Yang, S., & Cui, Y. (2022). Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *Expert Systems with Applications*, 207, 117991. <https://doi.org/10.1016/j.eswa.2022.117991>

25. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359. <https://doi.org/10.1109/TKDE.2009.191>
26. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. <https://arxiv.org/abs/1910.01108>
27. Sohag, S. R., Zhang, S., Xian, M., Sun, S., Xu, F., & Ma, Z. (2024). Causality extraction from nuclear licensee event reports using a hybrid framework. *arXiv preprint arXiv:2404.05656*. <https://doi.org/10.48550/arXiv.2404.05656>
28. Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, 62, 45–56. <https://doi.org/10.1016/j.autcon.2015.11.001>
29. Wang, H., Chen, M., Zhang, H., & Liu, Q. (2020). Joint constrained learning for event-event relation extraction. *arXiv preprint arXiv:2010.06727*. <https://arxiv.org/abs/2010.06727>
30. Wang, Y., Zhang, L., & Chen, X. (2023). Self-consistency and confidence calibration in large language models. *arXiv preprint arXiv:2305.14325*. <https://arxiv.org/abs/2305.14325>
31. Yang, X., & Mao, K. (2014). Multi-level causal relation identification using extended features. *Expert Systems with Applications*, 41(16), 7171–7181. <https://doi.org/10.1016/j.eswa.2014.05.048>
32. Zhang, S., Sohag, S. R., Xian, M., Sun, S., & Ma, Z. (2026). Failure Event Mining with Fine-Tuned Large Language Model: Case study of analyzing United States nuclear power plant failure event reports. *Risk Analysis*, 46(3), e70191. <https://doi.org/10.1111/risa.70191>
33. Zhang, X., Mahadevan, S., & Deng, X. (2021). Bayesian network modeling of accident investigation reports for aviation safety assessment. *Reliability Engineering & System Safety*, 209, 107371. <https://doi.org/10.1016/j.ress.2020.107371>
34. Zhao, S., Liu, T., Zhao, S., Chen, Y., & Nie, J.-Y. (2016). Event causality extraction based on connectives analysis. *Neurocomputing*, 173, 1943–1950. <https://doi.org/10.1016/j.neucom.2015.09.066>
35. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43–76. <https://doi.org/10.1109/JPROC.2020.3004555>

36. Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., & Le, Q. V. (2020). Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, 33, 3833–3845.