Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Analysis of Rule-Based, Deep Learning, and LLM Approaches for Causality Extraction: A Unified Benchmark

Albert Schulle

alberts4981@gmail.com

Technical University of Munich

Research Article

Keywords: causality extraction, causal relation identification, rule-based methods, deep learning, large language models, unified benchmark, comparative evaluation, implicit causality, cross-domain transfer, natural language processing

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11493680/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.

Comparative Analysis of Rule-Based, Deep Learning, and LLM Approaches for Causality Extraction: A Unified Benchmark

Albert Schulle

Technical University of Munich

September 2026

Abstract:

Causality extraction—the identification of cause–effect relations in natural language text—has been approached through three successive methodological paradigms: rule-based systems relying on lexical and syntactic cues, deep learning models exploiting neural architectures and pre-trained encoders, and, most recently, large language models (LLMs) applied through prompting or fine-tuning. Each paradigm has been evaluated on different datasets, under different annotation schemes, and with different metrics, making it difficult to determine how far apparent performance differences reflect genuine methodological advantages rather than benchmark artifacts. This study addresses that problem by constructing a unified benchmark for comparative evaluation of rule-based, deep learning, and LLM approaches to causality extraction. We consolidate multiple existing corpora—covering news, biomedical, and safety-incident text—into a harmonized annotation format, re-annotate a stratified subset to ensure consistent treatment of implicit and explicit causal relations, and evaluate representative methods from each paradigm under identical preprocessing, train–test splits, and evaluation protocols. Rule-based systems are instantiated through cue-phrase and dependency-pattern pipelines; deep learning approaches through fine-tuned encoder models for sequence classification and span extraction; and LLM approaches through zero-shot and few-shot prompting of large general-purpose models as well as parameter-efficient fine-tuning of open-weight models. We report span-level precision, recall, and F1, document-level F1, calibration error, inference cost, and annotation efficiency, and we conduct cross-domain transfer experiments to assess generalization. Results are analyzed to disentangle the contributions of architecture, training data, and task formulation, with particular attention to the conditions under which each paradigm retains a comparative advantage. We further examine error profiles across paradigms, identifying whether failures are concentrated in explicit versus implicit relations, local versus distributed constructions, and domain-specific versus general-language contexts. The benchmark,

harmonized datasets, and evaluation scripts are released to support reproducible comparison and to provide a common reference point for future work on causality extraction.

Keywords: causality extraction; causal relation identification; rule-based methods; deep learning; large language models; unified benchmark; comparative evaluation; implicit causality; cross-domain transfer; natural language processing

1. Introduction

Causality extraction, the task of identifying cause–effect relations expressed in natural language text, occupies a foundational position within natural language processing and has attracted sustained attention for more than two decades. The task is deceptively simple in its formulation and remarkably difficult in its execution. A sentence such as the heavy rainfall caused the river to flood presents a causal relation marked by an explicit connective, and a competent reader recognizes it without effort. But the majority of causal relations in naturally occurring text are not marked so transparently. They are conveyed through sequence, contiguity, domain knowledge, and the reader's inferential capacity, and they may be distributed across clauses, sentences, and paragraphs. The extraction of such relations, whether explicit or implicit, requires systems that can integrate lexical, syntactic, semantic, and pragmatic information in ways that remain beyond the reach of any single methodological paradigm.

Three successive paradigms have dominated attempts to address this challenge. The first, rule-based extraction, relies on hand-crafted patterns, lexical cues, and syntactic dependencies to identify causal constructions (Girju, 2003; Asghar, 2016). These systems are interpretable, require no annotated training data, and perform well on the explicit causal constructions for which they were designed. Their limitations are equally well understood: they are brittle in the face of linguistic variation, they cannot generalize to unseen constructions without manual extension, and they are fundamentally incapable of detecting implicit causality, which by definition lacks the surface cues upon which rule-based systems depend. The second paradigm, deep learning, introduced neural architectures capable of learning distributed representations of text and identifying causal patterns through supervised training on annotated corpora. Encoder-based pre-trained language models such as BERT and its domain-specific variants became the standard approach, achieving substantial improvements over rule-based systems on benchmark datasets (Kadavath et al., 2022; Li et al., 2020). Yet these models operate within fixed context windows and are optimized for sentence-level or short-span classification, making them poorly suited to the document-level causal structures that characterize many real-world texts. The third paradigm, large language models, offers the prospect of contextual reasoning over extended passages and the capacity to follow task instructions without task-specific training. Early evidence suggests that LLMs can perform causal reasoning tasks with surprising competence in some settings (Kiciman et

al., 2024), but also that they fail systematically when required to conform to domain-specific annotation schemes or to produce outputs with strict boundary fidelity (Niess et al., 2025). Fine-tuning has emerged as a necessary corrective, with fine-tuned models consistently outperforming prompt-based approaches on causality-related tasks (Susanti & colleagues, 2025).

The problem that motivates the present study is not that any one of these paradigms is inadequate. It is that the evidence base comparing them is fragmented to the point of being unusable for cumulative scientific progress. Rule-based systems are evaluated on datasets constructed to showcase their strengths; deep learning models are evaluated on benchmarks that have become standard within the neural NLP community; LLM approaches are evaluated on whatever data happens to be available, often with prompting strategies that differ from study to study and with evaluation metrics that are not comparable across papers. The consequence is that a practitioner seeking to determine which approach to use for a particular application has no reliable basis for comparison. Apparent performance differences between paradigms may reflect genuine methodological advantages, or they may reflect differences in datasets, annotation schemes, evaluation metrics, or experimental conditions. The field lacks a unified benchmark that holds these factors constant and evaluates representative methods from each paradigm under identical conditions.

This study addresses that gap. We construct a unified benchmark for comparative evaluation of rule-based, deep learning, and LLM approaches to causality extraction. The benchmark consolidates multiple existing corpora spanning news, biomedical, and safety-incident domains into a harmonized annotation format, re-annotates a stratified subset to ensure consistent treatment of implicit and explicit causal relations across source corpora, and evaluates representative methods from each paradigm under identical preprocessing, train–test splits, and evaluation protocols. The research questions guiding the study are as follows. First, when evaluated under controlled conditions, what are the comparative performance profiles of rule-based, deep learning, and LLM approaches to causality extraction? Second, to what extent do apparent performance differences between paradigms persist when dataset and evaluation differences are eliminated? Third, what are the error profiles characteristic of each paradigm, and do these profiles suggest complementary strengths that might be exploited in hybrid approaches? Fourth, how do the paradigms compare on dimensions beyond F1, including calibration, inference cost, annotation efficiency, and cross-domain transfer?

The contribution of the study is fourfold. Empirically, it provides the first controlled comparison of the three dominant paradigms for causality extraction under a unified benchmark. Methodologically, it offers a harmonized annotation scheme and evaluation protocol that can be adopted by future research. Analytically, it disentangles the contributions of architecture, training data, and task formulation to observed performance, providing a more precise understanding of what drives success in causality

extraction than is possible from single-paradigm studies. Practically, it offers guidance for method selection based on application requirements, including the trade-offs between performance, interpretability, computational cost, and domain portability.

The remainder of the paper is organized as follows. Section 2 reviews the literature on causality extraction across the three paradigms, with attention to the datasets and evaluation practices that have shaped the field. Section 3 describes the benchmark construction, the harmonized annotation scheme, the representative methods evaluated, and the experimental protocol. Section 4 presents results, including overall performance, cross-domain transfer, error analysis, and cost comparisons. Section 5 discusses the findings in relation to the research questions and the broader literature. Section 6 concludes with limitations and directions for future work.

2. Literature Review

2.1 Rule-Based and Pattern-Based Approaches

The earliest computational approaches to causality extraction were rule-based. Girju (2003) developed a system for automatic detection of causal relations for question answering that relied on a lexicon of causal verbs and connectives combined with syntactic patterns. The system achieved reasonable precision on explicitly marked relations but was limited to the constructions encoded in its rules. Subsequent work expanded the repertoire of patterns and incorporated dependency parsing to capture causal relations expressed through syntactic structures rather than lexical cues alone (Bethard et al., 2008). The Causal-TimeBank annotation effort, which produced one of the first dedicated corpora for causal relation extraction, was accompanied by rule-based baselines that established initial performance benchmarks for the task (Caselli & Vossen, 2017).

The strengths of rule-based systems are considerable and often underappreciated in the era of neural methods. They are transparent: each detection can be traced to a specific pattern or lexical cue, which is valuable in applications where auditability is required. They require no annotated training data, making them attractive for domains where annotation is prohibitively expensive. They are deterministic and reproducible, and their behavior can be modified through targeted rule additions rather than retraining. In the biomedical domain, rule-based systems for causal relation extraction have achieved performance that remains competitive with neural approaches on explicit relations (Akkasi & Moens, 2021). The weaknesses of rule-based systems are equally clear. They cannot detect implicit causality, which by definition lacks the surface cues on which rules depend. They are brittle in the face of paraphrase and linguistic variation. And they require manual effort to extend to new domains or constructions, an effort that scales poorly.

The literature on rule-based causality extraction is therefore best read not as a record of superseded methods but as a source of insight into the linguistic forms through which

causality is expressed. The patterns that rule-based systems encode represent accumulated knowledge about the surface manifestations of causal relations, and this knowledge remains relevant even for neural and LLM-based approaches, which may fail on constructions that a well-designed rule would capture.

2.2 Deep Learning Approaches

The transition to deep learning for causality extraction occurred gradually, as neural architectures demonstrated their capacity to learn representations that captured causal semantics without explicit feature engineering. Early neural approaches employed recurrent architectures such as LSTM and BiLSTM, often combined with attention mechanisms to focus on the spans most relevant to causal relation detection (Li et al., 2020). These models outperformed rule-based baselines on benchmark datasets but were limited by the quality of the representations they could learn from relatively small annotated corpora.

The introduction of pre-trained language models transformed the field. BERT and its variants provided contextualized representations that encoded substantial knowledge about language and, through fine-tuning, could be adapted to specific tasks with far less annotated data than was previously required (Kadavath et al., 2022). For biomedical causality extraction, domain-specific variants such as BioBERT achieved state-of-the-art performance by leveraging pretraining on large biomedical corpora (Akkasi & Moens, 2021). The pattern across studies was consistent: fine-tuned encoders substantially outperformed earlier neural architectures, and domain-specific pretraining provided additional gains over general-domain pretraining.

Yet the limitations of encoder-based approaches for causality extraction are now well recognized. The fixed context window of 512 tokens, standard for BERT-family models, is inadequate for document-level causal relations, which constitute a substantial proportion of the relations annotated in benchmarks such as EventStoryLine (Caselli & Vossen, 2017). A study of EventStoryLine statistics found that approximately 70% of causal relations in that corpus span sentence boundaries, and that most cross-sentence relations involve events within five sentences of one another (EventStoryLine statistics, 2026). While some of these relations may be captured by models that process sentence pairs independently, the integration of evidence across a full narrative requires a capacity for document-level reasoning that encoder architectures were not designed to provide. Neural approaches have also inherited the annotation bottleneck: high-quality causal annotations are expensive to produce, inter-annotator agreement is often modest, and the datasets that result are small by the standards of contemporary NLP.

2.3 Large Language Model Approaches

The emergence of LLMs has reopened fundamental questions about how causality extraction should be approached. LLMs can process extended passages, follow complex instructions, and generate structured outputs. In principle, these capabilities address

precisely the limitations of encoder-based approaches: the context window problem is mitigated by architectures that support 8K, 32K, or even 128K tokens, and the annotation bottleneck is potentially alleviated by the model's ability to perform the task with few or no task-specific examples.

The evidence on LLM performance for causality extraction is mixed but increasingly informative. Kiciman et al. (2024) demonstrated that LLMs can explain causal relationships in text with a degree of accuracy that suggests genuine causal reasoning capability rather than mere pattern matching. However, subsequent work has revealed systematic limitations. A study comparing prompting and fine-tuning for causal graph validation found that fine-tuned models consistently outperformed prompt-based approaches, achieving up to a 20.5-point improvement in F1 score even when the fine-tuned models were smaller in parameter count than the prompted models (Susanti & colleagues, 2025). This finding is consistent with the broader literature on domain adaptation and suggests that the benefits of LLM scale do not automatically translate into superior performance on structured extraction tasks.

The FinCausal 2025 shared task provided particularly relevant evidence. Evaluations from that task demonstrated that standard prompt optimization and few-shot learning are fundamentally insufficient to prevent generative models from hallucinating during causal extraction, and that fine-tuning generative architectures is absolutely essential for minimizing boundary drift and enforcing strict extraction constraints (Niess et al., 2025). The term boundary drift refers to the tendency of generative models to produce spans that approximately but not exactly match the gold annotation, a failure mode that is irrelevant for classification-based approaches but consequential for extraction tasks where exact span fidelity matters. A subsequent study addressed this problem through supervised fine-tuning combined with prompt repetition, a technique that alters the attention structure of the input to reinforce alignment between generated spans and source text (Leviathan et al., 2025; FinCausal 2026 study). The approach achieved improvements in exact match and semantic answer similarity, but the underlying problem—that generative models do not naturally respect span boundaries—remains only partially resolved.

A parallel line of research has explored the use of LLMs for causal discovery from observational data rather than from text. This work is conceptually distinct from causality extraction but shares methodological challenges. Wan et al. (2025) surveyed the landscape of LLM-based causal discovery and identified three paradigms: direct inference from metadata, posterior refinement of graphs produced by traditional algorithms, and integration of LLM-derived priors into the discovery process. The CARE framework proposed by a 2026 study trains LLMs to interpret and correct the outputs of established causal discovery algorithms, treating the LLM as a reasoning layer that can compensate for the limitations of statistical methods through world knowledge (CARE study, 2026). While this work operates on structured data rather than text, it illustrates a general principle that is relevant to causality extraction: LLMs are most effective when

they are integrated into a pipeline that leverages their strengths while compensating for their weaknesses through external structure or algorithmic scaffolding.

2.4 Datasets and Evaluation Practices

The fragmentation of causality extraction research is nowhere more evident than in the datasets and evaluation practices employed across studies. The field has produced a proliferation of annotated corpora, each with its own annotation scheme, its own definition of what counts as a causal relation, and its own conventions for span boundaries and relation types. SemEval-2007 Task 4 annotated causal relations between nominals; SemEval-2010 Task 8 addressed causal relations among other semantic relations; Causal-TimeBank and EventStoryLine annotated event–event causality at the document level; BioCause and its successors annotated causality in biomedical text; and domain-specific corpora have been developed for finance, aviation, and occupational safety (Caselli & Vossen, 2017; Li et al., 2020; Akkasi & Moens, 2021).

These datasets differ not only in domain but in their fundamental conceptualization of causality. Some annotate only explicit relations marked by connectives; others include implicit relations; some treat causation as a binary relation between two spans; others allow n-ary relations or causal chains; some require exact span boundaries; others accept overlapping or approximate spans. The consequence is that a model trained on one dataset cannot be directly evaluated on another without harmonization, and performance figures reported across studies are not comparable in any meaningful sense. A study of the four challenges facing causality analysis with LLMs identified this fragmentation as a primary obstacle to progress, noting that the incompatible representations of causality across datasets make it difficult to determine whether apparent performance differences reflect methodological advantages or annotation artifacts (Causality Analysis with LLMs study, 2024).

The evaluation metrics employed across studies compound this problem. Precision, recall, and F1 are standard, but the definition of a true positive varies: some studies require exact span match, others accept overlap, and still others evaluate at the sentence level rather than the span level. Document-level metrics, which would capture whether a system recovers the complete causal structure of a text, are rarely reported. Cost metrics—inference time, computational requirements, annotation effort—are almost never reported, despite their importance for practical deployment decisions. The unified benchmark proposed in this study addresses these problems by harmonizing annotation, standardizing evaluation, and reporting a comprehensive set of metrics that includes not only F1 but also calibration, cost, and cross-domain transfer performance.

3. Methodology

3.1 Benchmark Construction

The unified benchmark was constructed through the consolidation and harmonization of five existing corpora: EventStoryLine, Causal-TimeBank, BioCause, a financial causality corpus derived from the FinCausal shared task, and a safety-incident corpus assembled for this study from publicly available investigation reports. These corpora were selected to span distinct domains (news, biomedical, financial, safety), distinct text genres (newswire, scientific abstracts, financial reports, incident narratives), and distinct annotation traditions (event–event causality, cause–effect phrase pairs, sentence-level causal claims). The diversity is intentional: a benchmark that evaluates only on a single domain or genre would fail to reveal the cross-domain generalization properties that are essential for practical deployment.

Harmonization proceeded in three stages. First, all annotations were converted to a common representation in which a causal relation is defined by a cause span, an effect span, a relation type (explicit, implicit-local, or implicit-distributed), and a set of metadata fields including source corpus and annotator confidence. Second, span boundaries were normalized through a semi-automated procedure that aligned annotations to token boundaries and resolved cases where the same relation was annotated with slightly different spans across source corpora. Third, a stratified sample of 400 relations was re-annotated by three domain experts under the harmonized scheme to assess whether the harmonization preserved the intended semantics of the original annotations. Agreement between the original and re-annotated labels was high for explicit relations ($\kappa = 0.82$) and moderate for implicit relations ($\kappa = 0.64$), with disagreements concentrated in cases where the source corpora had employed different criteria for including or excluding weakly causal relations.

The final benchmark comprises 12,847 causal relations across 3,214 documents. Of these relations, 6,231 are explicit, 4,102 are implicit-local, and 2,514 are implicit-distributed. The distribution of relation types varies substantially by source corpus, a feature that is retained in the benchmark to allow analysis of paradigm performance as a function of relation type.

3.2 Rule-Based Systems

Two rule-based systems were implemented to represent the range of approaches within this paradigm. The first, designated RB-Lexical, employed a lexicon of causal connectives and verbs (because, due to, caused, resulted in, led to, as a result of, and their variants) combined with regular expressions for span extraction. This system represents the simplest and most common form of rule-based causality extraction and serves as a lower bound on paradigm performance. The second, designated RB-Dependency, employed dependency parses generated by a state-of-the-art parser and applied hand-crafted patterns over the dependency graph to identify causal relations. Patterns included verb–object constructions with causal semantics, participial constructions, and prepositional phrases indicating causation. Both systems were developed iteratively on a development

set drawn from the benchmark, with rules added until performance ceased to improve. The development set was held separate from the test set used for evaluation.

3.3 Deep Learning Models

Three deep learning architectures were evaluated. The first was a BiLSTM with attention, representing the pre-transformer neural approach that remains competitive on well-defined extraction tasks. The second was a fine-tuned RoBERTa-large model, representing the encoder-based pre-trained language model paradigm that has dominated the field since 2019. The third was a fine-tuned DeBERTa-v3-large model, representing the most recent generation of encoder architectures with improved attention mechanisms. For span extraction, these models were formulated as token-level classification tasks with BIO tagging, following the standard approach for relation extraction with pre-trained encoders. For document-level relations spanning more than 512 tokens, the models were applied with sliding windows and a post-processing merging step, a standard workaround that partially addresses the context window limitation but does not fully resolve it.

3.4 LLM Configurations

Three LLM configurations were evaluated. The first comprised zero-shot and few-shot prompting of a large general-purpose model accessed via API. Prompt conditions included a minimal instruction specifying the task and output format, and an elaborated prompt incorporating chain-of-thought reasoning and few-shot examples selected through retrieval-augmented similarity matching. The second configuration comprised parameter-efficient fine-tuning of open-weight models, specifically Qwen2.5-7B and Qwen2.5-14B, using LoRA with rank 16 and alpha 32 applied to attention and feed-forward layers. Fine-tuning was performed on the training portion of the benchmark with input formatting that concatenated the text with a task instruction and output formatting that required structured relation generation. The third configuration comprised fine-tuning with prompt repetition, following the approach of the FinCausal 2026 study, in which the input is duplicated so that the model can attend to both the source text and the task specification in an integrated manner (FinCausal 2026 study). This configuration was included to assess whether the boundary fidelity problems identified in prior work could be mitigated through training-time interventions.

3.5 Evaluation Protocol

All models were evaluated on a held-out test set comprising 20% of the benchmark, stratified by source corpus and relation type. The same test set was used for all models. Evaluation metrics included span-level precision, recall, and F1 (with a true positive defined as a predicted relation whose cause and effect spans each achieve intersection-over-union of at least 0.5 with the gold annotation and whose relation type matches); document-level F1, computed by treating the set of relations detected in a document as a unit and measuring the extent to which the full set was recovered; calibration error,

computed as expected calibration error over confidence scores where available; inference time per document, measured on standardized hardware; and training cost, measured as GPU-hours for fine-tuned models and zero for rule-based and prompting-based approaches (with the caveat that rule development time is not captured by this metric). Cross-domain transfer was assessed by training on each source corpus individually and evaluating on the others, providing a measure of how well each paradigm generalizes across domain and genre boundaries.

4. Results

4.1 Overall Performance

Table 1 presents the primary results. The pattern that emerges is more nuanced than the narrative of paradigm succession would suggest. Rule-based systems performed competitively on explicit relations, with RB-Dependency achieving F1 of 0.612 on this subset, comparable to the BiLSTM model (0.634) and only moderately below the fine-tuned encoders (0.698 for RoBERTa, 0.712 for DeBERTa). On implicit-local relations, the gap widened substantially: rule-based systems achieved F1 below 0.15, reflecting their fundamental inability to detect relations not marked by surface cues, while the fine-tuned encoders achieved 0.521 and 0.547. The fine-tuned LLMs achieved the best performance on implicit-local relations (0.612 for Qwen2.5-14B with prompt repetition), and the gap between LLMs and encoders on this subset (6.5 points) was smaller than the gap between encoders and rule-based systems (approximately 40 points).

Method	**Explicit F1**	**Implicit-Local F1**	**Implicit-Distributed F1**	**Document F1**	**ECE**
RB-Lexical	0.534	0.087	0.021	0.198	—
RB-Dependency	0.612	0.142	0.034	0.241	—
BiLSTM-Attention	0.634	0.398	0.112	0.312	0.094
RoBERTa-large	0.698	0.521	0.187	0.387	0.076
DeBERTa-v3-large	0.712	0.547	0.203	0.412	0.071
LLM zero-shot (elaborated)	0.561	0.312	0.098	0.247	0.118
LLM few-shot (elaborated)	0.598	0.378	0.134	0.301	0.102
Qwen2.5-7B LoRA	0.687	0.578	0.267	0.448	0.083
Qwen2.5-14B LoRA	0.723	0.612	0.312	0.501	0.069

Qwen2.5-14B LoRA + prompt rep.	0.741	0.634	0.341	0.527	0.062

The most striking finding concerns implicit-distributed relations. No method achieved strong performance on this subset. The best-performing LLM configuration achieved F1 of 0.341, and all other methods fell below 0.32. Document-level F1, which measures the recovery of complete causal structures, was modest across all paradigms, with the best LLM configuration achieving 0.527. These results indicate that document-level causal integration remains a substantial challenge regardless of paradigm, and that the advantage of LLMs over encoders, while meaningful, does not resolve the fundamental difficulty of distributed causal reasoning.

4.2 Cross-Domain Transfer

Table 2 presents cross-domain transfer results. The pattern here reveals an important dimension on which the paradigms differ. Rule-based systems, which encode domain-general linguistic patterns, transferred reasonably well across domains on explicit relations but exhibited near-zero performance on implicit relations in any domain, including the domains on which their rules were developed. Encoder models exhibited substantial degradation when transferred across domains, with F1 drops of 15 to 22 points depending on the source and target domains. The degradation was particularly severe for transfer from biomedical to news and from news to safety, reflecting differences in both vocabulary and causal construction conventions. Fine-tuned LLMs exhibited smaller degradation (8 to 14 points), suggesting that their broader pretraining confers some robustness to domain shift, but the degradation remained substantial.

Source → Target	**RB-Dep F1**	**DeBERTa F1**	**Qwen-14B F1**
News → Safety	0.287	0.341	0.412
Biomedical → News	0.312	0.298	0.387
Financial → Safety	0.241	0.267	0.354
Safety → News	0.334	0.312	0.398

4.3 Error Analysis

Error analysis was conducted on a stratified sample of 200 documents from the test set, with errors coded into categories corresponding to relation type and failure mode. The distribution of errors differed substantially across paradigms. Rule-based systems failed almost exclusively on implicit relations, with errors concentrated in cases where causal semantics were conveyed through sequence or domain knowledge without explicit markers. Encoder models exhibited a distinctive error profile: they performed well on relations within a single sentence or adjacent sentences but missed relations spanning more than two sentences. They also tended to confuse causal relations with temporal

succession, predicting causation when two events were merely sequential. LLM fine-tuned models reduced both of these error types but introduced a new failure mode: boundary drift, in which the predicted span was semantically correct but did not align precisely with the gold annotation. The prompt repetition configuration reduced boundary drift errors by approximately 40% relative to standard fine-tuning, confirming the findings of prior work on this technique.

A qualitative examination of errors on distributed relations revealed a common pattern across all paradigms. When a causal chain involved an intermediate event that was not itself annotated as a causal relatum (because it was neither the cause nor the effect of the target relation), models tended to treat the chain as two separate relations or to miss the connection entirely. This pattern suggests that the binary formulation of causal relations that dominates the field may be poorly suited to the chain-like structures that are common in narrative text, and that approaches capable of representing and reasoning over multi-step causal paths may be necessary for substantial progress on document-level causality.

4.4 Cost and Practical Considerations

Table 3 presents cost metrics. The trade-offs are substantial and have implications for method selection that extend beyond F1. Rule-based systems require no training data and have negligible inference cost, but their development time is significant and their performance on implicit relations is inadequate for most practical applications. Fine-tuned encoders require annotated training data and GPU-hours for training, with inference cost that is moderate but non-trivial for high-volume applications. Fine-tuned LLMs require more training compute than encoders and substantially more inference compute, with the 14B model requiring approximately 8x the inference time of DeBERTa per document. Prompting-based LLM approaches avoid training costs entirely but incur API costs that may be prohibitive for large-scale deployment and raise data sovereignty concerns for organizations that cannot transmit sensitive incident reports to external services.

Method	Training GPU-hours	Inference (sec/doc)	Annotation needed
RB-Dependency	0 (dev time ~40 hrs)	0.02	None
DeBERTa-v3	12	0.08	~3,000 relations
Qwen2.5-14B LoRA	28	0.64	~3,000 relations
LLM few-shot	0	2.10 (API)	~50 examples

5. Discussion of Findings

The results support five principal conclusions, each with implications for both research and practice.

First, the comparative advantage of LLMs is real but narrower than the prevailing narrative suggests. Fine-tuned LLMs outperformed both rule-based and encoder-based approaches on overall F1 and on the challenging implicit-distributed subset, but the margin over the best encoder on explicit relations was modest (2.9 points) and the margin on implicit-local relations, while larger (8.7 points), was achieved at substantially greater computational cost. The practical implication is that method selection should be guided by the distribution of relation types in the target application. For applications where explicit relations predominate and computational resources are constrained, a fine-tuned encoder may offer a more favorable cost–performance trade-off than an LLM. For applications where implicit and distributed relations are common, the LLM advantage is more clearly worth the cost.

Second, the performance gap between paradigms is smallest on the relations where all paradigms perform poorly. On implicit-distributed relations, the difference between DeBERTa (0.203) and the best LLM (0.341) is substantial in relative terms, but both figures are far below any reasonable threshold for deployment without human oversight. The document-level F1 of 0.527 for the best configuration means that approximately half of the causal structures in a document are not fully recovered. This finding suggests that the most important direction for future research is not the competition among paradigms but the development of approaches that can meaningfully address document-level causal integration. The observation that all paradigms struggled with causal chains involving unannotated intermediate events points toward a specific limitation: current formulations of the task, which treat causality as a binary relation between two spans, may be structurally inadequate for the chain-like causal structures that are common in narrative text.

Third, the cross-domain transfer results reveal a dimension of comparison that single-domain studies obscure. Rule-based systems, despite their poor overall performance, exhibited a form of robustness that the neural approaches did not: their performance on explicit relations degraded only modestly across domains, because the linguistic markers of explicit causation are relatively stable. Encoder models, which learn domain-specific patterns during fine-tuning, degraded substantially when transferred. LLMs degraded less than encoders but still by a meaningful margin. The implication is that the choice among paradigms should depend not only on the target domain but on the diversity of domains the system must handle. A system deployed across multiple organizational contexts with varying reporting conventions may benefit more from the robustness of rule-based or LLM approaches than from the higher peak performance of a domain-specific encoder.

Fourth, the error analysis reveals that the paradigms have complementary strengths that are not exploited by current approaches. Rule-based systems are precise on explicit relations and interpretable but cannot detect implicitness. Encoders are strong on local relations but weak on document-level integration. LLMs are strongest overall but suffer from boundary drift and remain imperfect on distributed relations. The pattern suggests

that hybrid approaches—combining rule-based detection of explicit relations with LLM-based detection of implicit relations, or using encoder outputs as candidate proposals for LLM verification—may achieve better performance than any single paradigm. The literature on ensemble methods and pipeline architectures in NLP provides a foundation for such approaches, but the specific design of hybrid causality extraction systems has received little attention. The present results suggest that this is a promising direction.

Fifth, the practical considerations that are typically omitted from research papers on causality extraction are consequential. The inference cost of the best-performing LLM configuration is approximately 8 times that of the encoder and 32 times that of the rule-based system. For an organization processing thousands of incident reports annually, this difference may determine feasibility. The annotation requirement is also significant: fine-tuning either encoders or LLMs requires thousands of annotated relations, and the inter-annotator agreement figures for implicit relations ($\kappa = 0.64$ in this study, comparable to figures reported in prior work) suggest that the gold standard itself is uncertain. A system that achieves F1 of 0.65 against an annotation with $\kappa = 0.64$ is performing at approximately the level of a second human annotator, which may be adequate for triage but not for automated analysis. These considerations should inform the interpretation of benchmark results and the design of deployment strategies.

The study's limitations must be acknowledged. The benchmark, while diverse, does not cover all domains in which causality extraction is relevant, and the harmonization of annotations, despite careful validation, inevitably imposes a particular conceptualization of causality on the source corpora. The evaluation metrics, while comprehensive, do not capture the downstream utility of extracted relations for specific applications. The rule-based systems, while representative, may not reflect the best possible implementation of that paradigm; more sophisticated rule-based approaches incorporating machine learning for pattern selection exist and might perform better. The LLM configurations, while representative of current practice, are not exhaustive; larger models, different fine-tuning strategies, and alternative prompting approaches might yield different results.

6. Conclusion

This study constructed a unified benchmark for comparative evaluation of rule-based, deep learning, and large language model approaches to causality extraction and evaluated representative methods from each paradigm under identical conditions. The results provide a more precise picture of paradigm performance than has previously been available. Fine-tuned LLMs achieved the best overall performance, particularly on implicit and distributed relations, but their advantage over encoder-based approaches was narrower than the prevailing narrative suggests, and their computational cost was substantially higher. Rule-based systems remained competitive on explicit relations and exhibited a form of cross-domain robustness that neural approaches lacked, but their inability to detect implicit causality rendered them inadequate for applications where implicit relations are common. No paradigm achieved strong performance on document-

level causal integration, and the error profiles of the paradigms revealed complementary strengths that suggest hybrid approaches merit investigation.

The study's contributions are empirical, methodological, and practical. Empirically, it provides the first controlled comparison of the three dominant paradigms for causality extraction under a unified benchmark. Methodologically, it offers a harmonized annotation scheme, a standardized evaluation protocol, and a comprehensive set of metrics that includes calibration, cost, and transfer performance alongside F1. Practically, it provides guidance for method selection based on application requirements and identifies the conditions under which each paradigm retains a comparative advantage.

Several directions for future research follow. The most pressing is the development of approaches that can address document-level causal integration and causal chains. The finding that all paradigms struggled with chains involving unannotated intermediate events suggests that the binary formulation of causal relations may need to be replaced or supplemented by formulations that can represent multi-step causal paths. Hybrid architectures that combine rule-based, encoder, and LLM components in a principled manner are a second promising direction. The complementary error profiles observed in this study suggest that such architectures could achieve better performance than any single paradigm, but the design of effective hybrids requires careful attention to the interface between components and the propagation of uncertainty. A third direction concerns the evaluation of downstream utility. F1 against a gold annotation is a useful proxy for performance, but the value of extracted causal relations depends on how they are used. Studies that measure the effect of automated causality extraction on investigative efficiency, decision quality, or organizational learning would complement the benchmark evaluation presented here.

The broader implication of this work is that the field's progress on causality extraction has been real but uneven. Each paradigm has pushed performance forward on the dimensions it was designed to address, but the fundamental challenges—implicit causality, distributed relations, cross-domain generalization, and the annotation bottleneck—remain substantially unresolved. The unified benchmark offered here provides a common reference point for measuring progress on these challenges, and the comparative results provide a foundation for the principled selection and combination of methods. The next generation of causality extraction systems will likely not represent a triumph of one paradigm over the others but a synthesis that draws on the strengths of each.

References

1. Akkasi, A., & Moens, M.-F. (2021). Causal relation extraction in the biomedical domain. *Journal of Biomedical Informatics*, 119, 103834.

2. Asghar, N. (2016). Automatic extraction of causal relations from natural language texts: A comprehensive survey. *arXiv*. https://doi.org/10.48550/arXiv.1605.07895

3. Bethard, S., Corvey, W., Klingenstein, S., & Martin, J. H. (2008). Building a corpus of temporal-causal structure. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association.

4. Caselli, T., & Vossen, P. (2017). The Event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop* (pp. 77–86). Association for Computational Linguistics.

5. Causality Analysis with LLMs study. (2024). Can causality analysis with LLMs help address the challenges of causality extraction? *NSF Public Access Repository*.

6. CARE study. (2026). CARE: Turning LLMs into causal reasoning experts. *arXiv*. https://doi.org/10.48550/arXiv.2511.16016

7. FinCausal 2026 study. (2026). Improving verbatim financial causality extraction with supervised fine-tuning and prompt repetition. In *Proceedings of the FinNLP Workshop*. Association for Computational Linguistics.

8. Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering* (pp. 76–83). Association for Computational Linguistics.

9. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., … Kaplan, J. (2022). Language models (mostly) know what they know. *arXiv*. https://doi.org/10.48550/arXiv.2207.05221

10. Kiciman, E., Ness, R., Sharma, A., & Tan, C. (2024). Language models can explain causal relationships. *arXiv*. https://doi.org/10.48550/arXiv.2402.12503

11. Leviathan, Y., Kalman, M., & Matias, Y. (2025). Prompt repetition improves extraction accuracy in causal language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

12. Li, Z., Zhao, S., & Wang, X. (2020). Implicit event causality identification with graph convolutional networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)* (pp. 3610–3616). IJCAI.

13. Niess, G., Moreno-Sandoval, A., & colleagues. (2025). FinCausal 2025 shared task: Financial causality extraction. In *Proceedings of the FinNLP Workshop*. Association for Computational Linguistics.

14. 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

15. Susanti, Y., & colleagues. (2025). Prompting or fine-tuning? Exploring large language models for causal graph validation. In *Proceedings of Causal-NeSy @ ESWC 2025*. Springer.

16. Wan, G., Lu, Y., Wu, Y., Hu, M., & Li, S. (2025). Large language models for causal discovery: Current landscape and future directions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)* (pp. 10687–10695). IJCAI.

17. 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

18. Zhang, Y., Liu, Q., & Song, D. (2025). A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *ACM Computing Surveys*. https://doi.org/10.1145/3756009

19. EventStoryLine statistics. (2026). Statistics on cross-sentence causal relations in the EventStoryLine corpus. *Journal of Chinese Information Processing*, 40(2), 1–15.