

Quantifying Uncertainty in LLM-Extracted Causal Relations for Safety-Critical Decision Support

Jim Dorsey

jimdorsey4521@gmail.com

Miami University of Science and Technology

Research Article

Keywords: uncertainty quantification, large language models, causality extraction, safety-critical decision support, conformal prediction

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11473816/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.

Quantifying Uncertainty in LLM-Extracted Causal Relations for Safety-Critical Decision Support

Jim Dorsey

Miami University of Science and Technology

Written: September 2026

Abstract

Large language models (LLMs) have demonstrated strong performance in extracting causal relations from safety-critical technical narratives, including nuclear Licensee Event Reports, aviation incident reports, and medical device adverse event reports. However, the outputs of LLM-based causality extraction are typically presented as deterministic cause-effect pairs with no accompanying measure of confidence, leaving downstream decision-makers—risk analysts, regulators, and plant operators—unable to distinguish reliably extracted causal relations from spurious or low-confidence ones. This limitation is particularly consequential in safety-critical domains, where acting on an incorrectly extracted causal relation could lead to misprioritized maintenance, flawed risk model updates, or inappropriate regulatory action. This research develops and evaluates a framework for quantifying uncertainty in LLM-extracted causal relations for safety-critical decision support. The framework comprises three components: (1) a multi-source uncertainty quantification pipeline that combines token-level probability signals, self-consistency sampling across stochastic decoding runs, ensemble disagreement among multiple fine-tuned models, and conformal prediction to produce calibrated confidence estimates for each extracted cause-effect pair; (2) a validation study that assesses whether these uncertainty estimates correlate with expert judgments of extraction validity, using a corpus of nuclear Licensee Event Reports annotated by domain experts; and (3) a decision-theoretic evaluation that examines how uncertainty-aware extraction outputs affect downstream risk-informed decisions, including the flagging of causal relations for expert review and the weighting of extracted pathways

in probabilistic risk model updates. The research further investigates the sources of LLM extraction uncertainty, distinguishing aleatoric uncertainty arising from genuine ambiguity in the source text from epistemic uncertainty arising from model limitations and distributional shift. Results are expected to show that calibrated uncertainty estimates substantially improve the precision of high-confidence extraction subsets and enable cost-effective triage of extracted causal relations for expert review, thereby supporting the trustworthy deployment of LLM-based causality extraction in safety-critical decision contexts.

Keywords: uncertainty quantification; large language models; causality extraction; safety-critical decision support; conformal prediction; calibration; nuclear event reports; expert validation; risk-informed decision-making; trustworthy AI

1. Introduction

Large language models have emerged as powerful tools for extracting structured knowledge from unstructured technical narratives, and their application to safety-critical domains represents one of the most consequential frontiers of contemporary artificial intelligence research. In nuclear safety, aviation incident analysis, medical device surveillance, and chemical process safety, the ability to automatically identify causal relations from event reports holds the promise of dramatically accelerating the incorporation of operating experience into formal risk models and regulatory decision processes (Sohag et al., 2024; Zhang et al., 2026). The extraction of cause-effect pairs from Licensee Event Reports and similar documents has progressed from rule-based methods through deep learning approaches to fine-tuned large language models, with each advance bringing improved accuracy and broader coverage of the linguistic variability characteristic of technical narratives (Sohag et al., 2024; Zhang et al., 2026). Yet a fundamental limitation persists across this entire trajectory of methodological development: the outputs of extraction systems are presented as deterministic assertions of causal relations, with no accompanying measure of confidence, reliability, or uncertainty. A cause-effect pair extracted by an LLM is rendered indistinguishable, in its representational form, from a cause-effect pair that the model has inferred with near-certainty or one that the model has fabricated entirely.

This limitation is not merely a matter of presentational convenience. In safety-critical decision contexts, the distinction between a reliably extracted causal relation and a spurious one is precisely the distinction upon which consequential decisions depend. A risk analyst updating a probabilistic risk assessment model must decide which extracted causal pathways to incorporate and which to flag for expert review (Sirola et al., 2020). A regulatory reviewer assessing a licensee's corrective action must

determine whether the causal factors identified in an event report are adequately supported by the narrative evidence (Apostolakis, 2004). A plant operator prioritizing maintenance activities must weigh the likelihood that a reported failure mode is genuinely linked to the observed consequence against the possibility that the apparent linkage reflects coincidence or extraction error. In each of these decisions, the absence of uncertainty information forces decision-makers to treat all extraction outputs as equally trustworthy, or to rely on heuristics and intuition that are neither systematic nor auditable (Kadak and Matsuo, 2007). The result is a systematic underutilization of extraction capability in precisely the contexts where it could be most valuable.

The research literature on uncertainty quantification for large language models has expanded rapidly in response to the recognition that fluency and factual correctness are not coextensive properties of model outputs (Shelmanov et al., 2025; Fadeeva et al., 2023). Methods for detecting hallucinations, estimating semantic entropy, and providing calibrated confidence scores have been developed and benchmarked across a range of natural language processing tasks (Farquhar et al., 2024; Lin et al., 2023). Conformal prediction has emerged as a particularly promising framework for providing statistical guarantees on the correctness of model outputs, with applications ranging from question answering to factuality control in multi-step reasoning (Wang et al., 2026; Su et al., 2024). These developments establish the technical feasibility of attaching reliable uncertainty signals to LLM outputs. However, the application of these methods to structured information extraction tasks, and particularly to causality extraction in safety-critical domains, remains largely unexplored. The uncertainty quantification literature has focused predominantly on free-form generation and question answering, where the unit of analysis is a complete response or a claim within a response (Zhao et al., 2026). The unit of analysis in causality extraction is fundamentally different: it is a relation between two entities, extracted from a span of text, whose validity depends on linguistic evidence, domain knowledge, and the correctness of entity identification and relation classification (Asghar, 2016). Extending uncertainty quantification to this structured extraction setting requires methodological innovation that has not yet been undertaken.

The problem is compounded by the characteristics of causality extraction itself. Recent benchmarking of LLMs on pairwise causal discovery has revealed substantial performance limitations, with the best-performing open-source models achieving mean scores below fifty percent on causal extraction tasks (Anuyah et al., 2026). Performance is particularly poor for implicit causal relations, relations spanning multiple sentences, and texts containing multiple causal pairs, all of which are common in the narrative-dense environment of Licensee Event Reports (Anuyah et al., 2026; Sohag et al., 2024). These findings indicate that extraction errors are not rare edge cases but constitute a substantial proportion of extraction outputs, making the need for uncertainty

quantification correspondingly more urgent. If a model extracts causal relations with an overall accuracy of approximately seventy to eighty percent, as reported in prior work on nuclear event reports (Sohag et al., 2024; Zhang et al., 2026), then between one in five and one in four extracted relations is incorrect. Without uncertainty information, decision-makers have no principled basis for distinguishing the reliable majority from the unreliable minority.

The research described in this paper addresses this gap by developing and evaluating a framework for quantifying uncertainty in LLM-extracted causal relations for safety-critical decision support. The framework comprises three interconnected components. The first component is a multi-source uncertainty quantification pipeline that combines four complementary signals: token-level probability information from the extraction model, self-consistency across multiple stochastic decoding runs, ensemble disagreement among independently fine-tuned models, and conformal prediction to provide calibrated confidence estimates with distribution-free coverage guarantees. These signals are integrated to produce a calibrated uncertainty score for each extracted cause-effect pair, reflecting both aleatoric uncertainty arising from genuine ambiguity in the source text and epistemic uncertainty arising from model limitations and distributional shift (Hüllermeier and Waegeman, 2021). The second component is a validation study that assesses whether the uncertainty estimates produced by the pipeline correlate with expert judgments of extraction validity, using a corpus of nuclear Licensee Event Reports annotated by domain experts. This component addresses the critical question of whether uncertainty estimates are meaningful in the specific domain of application, rather than merely statistically well-calibrated in the abstract. The third component is a decision-theoretic evaluation that examines how uncertainty-aware extraction outputs affect downstream risk-informed decisions, including the flagging of causal relations for expert review and the weighting of extracted pathways in probabilistic risk model updates. This component connects the technical machinery of uncertainty quantification to the practical decision contexts that motivate the research.

The significance of this research extends across methodological, practical, and regulatory dimensions. From a methodological perspective, it advances the state of the art in uncertainty quantification by extending existing methods from free-form generation to structured extraction tasks, and by integrating multiple sources of uncertainty information in a principled framework. From a practical perspective, it enables the trustworthy deployment of LLM-based causality extraction in safety-critical decision contexts by providing decision-makers with the confidence information they need to use extraction outputs appropriately. From a regulatory perspective, it supports the development of auditable and defensible processes for incorporating machine-

extracted information into risk-informed decision-making, addressing a critical gap in the governance of AI in safety-critical domains (Apostolakis, 2004).

The research is guided by three primary research questions. First, what methods for quantifying uncertainty in LLM-extracted causal relations produce calibrated confidence estimates that reliably distinguish accurate from inaccurate extractions, and how do these methods compare in the context of nuclear event reports? Second, to what extent do the uncertainty estimates produced by these methods correlate with expert judgments of extraction validity, and what factors influence the agreement between automated uncertainty signals and human assessments? Third, how does the availability of uncertainty information affect downstream risk-informed decisions, including the efficiency of expert review triage and the quality of probabilistic risk model updates? These questions are addressed through a methodology that combines uncertainty quantification techniques from natural language processing, expert validation from domain practice, and decision-analytic evaluation.

The scope of the research is deliberately bounded. It focuses on causality extraction from nuclear Licensee Event Reports, building directly on the extraction methods demonstrated by Sohag et al. (2024) and Zhang et al. (2026). While the framework is designed to be adaptable to other safety-critical domains, the empirical evaluation is conducted in the nuclear domain, where the availability of expert annotators and established risk assessment frameworks enables rigorous validation. The research addresses the uncertainty of extraction outputs, not the uncertainty of the underlying causal relations in the world; that is, it quantifies confidence that an extracted relation accurately reflects what the source text states, not confidence that the stated relation is true in an ontological sense (Pearl, 2009). This distinction is important for interpreting the results and for understanding the scope of the contribution.

The remainder of this paper is organized as follows. Section 2 presents a comprehensive literature review covering uncertainty quantification for large language models, causality extraction in safety-critical domains, and the integration of uncertainty information into decision processes. Section 3 describes the methodology in detail, including the data sources, the multi-source uncertainty quantification pipeline, the expert validation protocol, and the decision-theoretic evaluation framework. Section 4 presents the results of the uncertainty quantification experiments, the expert validation study, and the decision evaluation. Section 5 discusses the findings in relation to the research questions, examines implications for practice and research, and identifies limitations and directions for future work. Section 6 concludes the paper by summarizing the principal contributions and their significance.

2. Literature Review

2.1 Uncertainty Quantification for Large Language Models

The recognition that large language models can produce fluent but incorrect outputs has motivated a substantial and rapidly growing body of research on uncertainty quantification (Shelmanov et al., 2025; Zhao et al., 2026). The fundamental challenge is that autoregressive language models generate text token by token, and the probability distribution over tokens at each step does not directly correspond to the probability that the complete generated sequence is correct or reliable (Farquhar et al., 2024). This mismatch between token-level probabilities and sequence-level correctness has driven the development of methods that operate at different levels of analysis, from the internal representations of the model to the semantic content of sampled outputs.

White-box methods for uncertainty quantification leverage access to the model's internal states, including token-level probabilities, hidden representations, and attention patterns (Shelmanov et al., 2025). Token-level entropy and probability scores provide a direct measure of the model's confidence at each generation step, but they suffer from the limitation that a model can be confidently wrong, assigning high probability to incorrect tokens when its learned representations are systematically biased (Farquhar et al., 2024). Learned probes that use the model's hidden representations to predict correctness have shown promise, but they require access to internal states that are unavailable for many deployed models and may not generalize across tasks (Azaria and Mitchell, 2023). These limitations have motivated the development of black-box methods that operate solely on the model's textual outputs.

Semantic entropy has emerged as a leading black-box approach for uncertainty quantification in free-form generation. The method generates multiple samples from the model for a given input, clusters the samples according to semantic equivalence using a natural language inference model, and computes the entropy of the distribution over semantic clusters (Farquhar et al., 2024). The intuition is that if the model consistently produces semantically equivalent outputs across samples, its underlying belief is likely to be reliable, whereas semantic diversity across samples indicates uncertainty. Semantic entropy has been shown to outperform token-level probability methods for hallucination detection across a range of tasks (Farquhar et al., 2024; Zhao et al., 2026). Extensions of the approach, including trajectory semantic entropy, incorporate uncertainty signals from the reasoning process as well as the final answer (IEEE, 2026). The structural entropy method (SeSE) constructs a hierarchical abstraction of the semantic space to quantify uncertainty with improved precision and interpretability (Zhao et al., 2026).

Consistency-based methods represent a complementary approach that assesses uncertainty by measuring the agreement among multiple model outputs generated through repeated sampling (Vashurin et al., 2025). Self-consistency, originally introduced as a technique for improving reasoning performance, can be repurposed as an uncertainty signal: if the model produces the same answer across multiple samples, that answer is likely to be reliable, whereas disagreement across samples indicates uncertainty (Wang et al., 2023). The CoCoA framework integrates confidence-based and consistency-based signals within a minimum Bayes risk framework, demonstrating that the combination of these complementary sources of information yields improvements over either alone (Vashurin et al., 2025). This integration of multiple uncertainty signals aligns with the multi-source approach adopted in the present research.

Ensemble methods provide a third approach to uncertainty quantification, leveraging disagreement among multiple models as a signal of epistemic uncertainty (Lakshminarayanan et al., 2017). Deep ensembles, which train multiple models with different initializations and aggregate their predictions, have been shown to provide well-calibrated uncertainty estimates in classification and regression tasks. The application of ensemble methods to large language models is complicated by the computational cost of training and running multiple models, but parameter-efficient fine-tuning methods such as low-rank adaptation make ensembles more feasible (Hu et al., 2022). Ensemble disagreement among fine-tuned models provides a direct measure of the extent to which extraction outputs depend on the specific training run, which is a form of epistemic uncertainty that is particularly relevant when annotated data is limited (Hüllermeier and Waegeman, 2021).

Conformal prediction offers a distribution-free framework for providing statistical guarantees on the correctness of model outputs (Angelopoulos and Bates, 2021). The method uses a calibration set of examples with known ground truth to compute a nonconformity score threshold that ensures, with a user-specified probability, that the prediction set for a new example contains the correct answer. Conformal prediction has been applied to large language models for factuality control and selective generation, with methods such as ConU providing correctness coverage guarantees for open-ended generation tasks (Su et al., 2024). The inference-time conformal reasoning framework extends conformal prediction to multi-step reasoning, providing valid coverage guarantees for factuality control during generation (Wang et al., 2026). These developments establish conformal prediction as a rigorous and practical framework for uncertainty quantification, though its application to structured extraction tasks remains unexplored.

The literature on uncertainty quantification for large language models thus provides a rich methodological toolkit. However, the application of these methods to causality extraction presents specific challenges that have not been addressed. The unit of analysis is a relation between two entities rather than a free-form response, requiring methods that can attach uncertainty to structured outputs (Asghar, 2016). The correctness criterion for an extracted causal relation is more complex than for a question-answering response, involving both entity identification and relation classification (Sohag et al., 2024). The safety-critical context demands not only statistical calibration but also interpretability and actionability, since decision-makers must understand what the uncertainty estimate means and how to use it (Apostolakis, 2004). These challenges motivate the framework developed in this research.

2.2 Causality Extraction from Technical Narratives

Causality extraction from text is a well-established task in natural language processing, with a substantial literature spanning rule-based, machine learning, and, more recently, large language model approaches (Asghar, 2016). The task involves identifying causal language in text, extracting the entities or events that stand in causal relation to one another, and classifying the type of causal relation (Dunietz et al., 2017). Early approaches relied on explicit causal markers such as “because,” “due to,” and “caused by,” which provide reliable but incomplete coverage of causal language in natural text (Asghar, 2016). Machine learning approaches, particularly deep learning models, enabled the capture of implicit causal relations through learned patterns, improving recall at the cost of some precision (Li et al., 2020).

The application of causality extraction to safety-critical technical narratives has received increasing attention in recent years. Aviation incident reports, medical device adverse event reports, and nuclear Licensee Event Reports share characteristics that make them both valuable and challenging sources of causal information: they are written by domain experts, they describe complex multi-factor events, and they use specialized terminology that differs from general-domain text (Sohag et al., 2024; Zhang et al., 2026). The research by Sohag et al. (2024) and Zhang et al. (2026) represents the most directly relevant work for the present study, as it addresses causality extraction specifically from nuclear Licensee Event Reports and provides the extraction foundation upon which the uncertainty quantification framework is built.

Sohag et al. (2024) developed a hybrid framework that combined deep learning for causal relation detection with rule-based extraction for cause-effect pair identification. Their work compiled a corpus of over twenty thousand text samples from Licensee Event Reports and demonstrated that deep learning could detect causal relations with high accuracy, while rule-based extraction achieved a recall of approximately seventy-two percent (Sohag et al., 2024). Zhang et al. (2026) extended

this work by employing a fine-tuned large language model, achieving improved extraction performance and demonstrating the feasibility of applying large language models to nuclear event narratives. These contributions establish both the feasibility and the limitations of automated causality extraction in the nuclear domain. The reported performance, while strong by the standards of general-domain extraction, implies that a non-trivial proportion of extracted relations are incorrect, underscoring the need for uncertainty quantification.

Benchmarking studies of LLMs on causal discovery tasks have provided additional evidence of extraction limitations. Anuyah et al. (2026) evaluated thirteen open-source LLMs on pairwise causal discovery from text, finding that the best-performing model achieved a mean score of less than fifty percent on causal extraction. Performance was particularly poor for implicit relationships, links spanning multiple sentences, and texts containing multiple causal pairs (Anuyah et al., 2026). These findings are consistent with the error analysis reported by Sohag et al. (2024), which identified implicit causality as a major source of extraction difficulty. The implication is that extraction errors are systematic and predictable, which suggests that uncertainty quantification methods that capture the sources of difficulty can provide meaningful discrimination between reliable and unreliable extractions.

The challenges of causality extraction in safety-critical domains are not solely technical. The interpretation of causal language in technical narratives requires domain knowledge that may not be fully captured by the extraction model, and the validity of an extracted relation depends on the context and purpose of the analysis (Pearl, 2009). A cause-effect pair that is accurately extracted from a Licensee Event Report may nonetheless be irrelevant to a particular risk assessment application, or may be already represented in the risk model, or may be novel and potentially important (Sirola et al., 2020). The uncertainty quantification framework developed in this research focuses on the confidence that an extracted relation accurately reflects what the source text states, leaving questions of relevance and novelty to downstream analysis. This distinction is important for understanding what the uncertainty estimates do and do not measure.

2.3 Expert Validation and Human-in-the-Loop Evaluation

The evaluation of uncertainty quantification methods presents distinctive challenges that are not fully addressed by standard automated metrics. A method can produce statistically well-calibrated uncertainty estimates, in the sense that the predicted probabilities match observed frequencies, while nonetheless failing to provide useful discrimination between reliable and unreliable outputs in a specific application context (Hüllermeier and Waegeman, 2021). Expert validation provides a complementary perspective, assessing whether uncertainty estimates align with the

judgments of domain experts who are the ultimate arbiters of extraction validity in safety-critical contexts.

The use of expert judgment in the evaluation of natural language processing systems has a long history, particularly in domains where automated ground truth is difficult to establish (Creswell and Plano Clark, 2018). Expert annotation provides the gold standard for evaluating extraction accuracy, and inter-annotator agreement measures provide a means of assessing the reliability of expert judgments. However, expert annotation is expensive and time-consuming, which limits the scale at which it can be applied. The research by Zhang et al. (2026) employed expert validation to assess the quality of LLM-extracted causal relations, finding generally positive assessments of accuracy and usefulness while identifying specific areas for improvement. This research builds on that approach by focusing specifically on the relationship between automated uncertainty estimates and expert judgments, a question that has not been systematically addressed in the causality extraction literature.

The human-in-the-loop paradigm provides a framework for integrating expert judgment into the extraction pipeline (Kadak and Matsuo, 2007). Rather than treating the model as an autonomous extractor whose outputs are either accepted or rejected in their entirety, the human-in-the-loop approach uses the model to identify and prioritize candidates for expert review, focusing scarce expert attention on the cases where the model is least confident (Zhang et al., 2026). This approach has the potential to dramatically improve the efficiency of expert review while maintaining or improving the quality of the final extracted knowledge base. The effectiveness of the approach depends critically on the quality of the uncertainty estimates: if the model’s confidence is well-calibrated, low-confidence cases will be disproportionately likely to contain extraction errors, and expert review will be correspondingly efficient (Su et al., 2024). The research described in this paper evaluates this premise directly.

2.4 Decision-Theoretic Integration of Uncertainty

The ultimate purpose of uncertainty quantification in safety-critical contexts is to support better decisions. The decision-theoretic framework provides a principled basis for understanding how uncertainty information should affect decision-making, distinguishing between the value of information, the cost of errors, and the trade-offs between different courses of action (Pearl, 2009). In the context of causality extraction for risk assessment, decisions include which extracted relations to incorporate into risk models, which to flag for expert review, and which to exclude or treat as unreliable. The availability of calibrated uncertainty estimates enables these decisions to be made on a rational basis, weighing the expected benefits of incorporating a relation against the expected costs of acting on a false extraction (Sirola et al., 2020).

The application of decision theory to the evaluation of uncertainty quantification methods has been relatively limited in the natural language processing literature, which has focused predominantly on statistical calibration and discrimination metrics (Shelmanov et al., 2025). However, the practical value of uncertainty quantification depends on its consequences for decisions, not merely on its statistical properties (Apostolakis, 2004). A method that produces well-calibrated probabilities but does not improve decision outcomes is of limited practical utility. Conversely, a method with imperfect calibration that nonetheless provides useful discrimination between high-stakes and low-stakes cases can be valuable. The decision-theoretic evaluation component of this research addresses this gap by assessing the downstream consequences of uncertainty-aware extraction for expert review triage and risk model updating.

The risk-informed decision-making framework that has become central to nuclear safety regulation provides a natural context for this evaluation (Apostolakis, 2004). In this framework, decisions are informed by probabilistic risk assessments that quantify the likelihood and consequences of adverse events, and the goal is to allocate resources and attention in proportion to risk. The incorporation of extracted causality into risk models can improve the fidelity of the models, but only if the extracted relations are accurate (Sirola et al., 2020). The uncertainty quantification framework enables a risk-informed approach to the use of extracted causality itself, directing expert review and modeling effort toward the relations whose accuracy is most uncertain and whose potential impact is greatest.

2.5 Research Gap and Contribution

The literature review reveals a significant research gap at the intersection of uncertainty quantification for large language models, causality extraction in safety-critical domains, and decision-theoretic evaluation. The uncertainty quantification literature has developed sophisticated methods for attaching confidence signals to free-form generation, but has not systematically addressed structured extraction tasks (Shelmanov et al., 2025; Zhao et al., 2026). The causality extraction literature has demonstrated the feasibility of extracting causal relations from nuclear event reports, but has presented extraction outputs as deterministic assertions without uncertainty information (Sohag et al., 2024; Zhang et al., 2026). The decision-theoretic evaluation literature has established frameworks for assessing the value of information, but has not been applied to the evaluation of uncertainty-aware extraction in risk assessment contexts (Apostolakis, 2004). The research described in this paper addresses this gap by developing a multi-source uncertainty quantification framework specifically for LLM-extracted causal relations, validating the framework against expert judgments, and evaluating its consequences for risk-informed decisions.

The contributions of the research are fourfold. First, it extends uncertainty quantification methods from free-form generation to structured relation extraction, developing a multi-source pipeline that integrates token-level probabilities, self-consistency, ensemble disagreement, and conformal prediction. Second, it provides the first systematic evaluation of uncertainty quantification for causality extraction in a safety-critical domain, using a corpus of nuclear Licensee Event Reports annotated by domain experts. Third, it establishes the relationship between automated uncertainty estimates and expert judgments of extraction validity, identifying the extent to which uncertainty estimates can serve as a proxy for expert review. Fourth, it evaluates the decision-theoretic consequences of uncertainty-aware extraction, assessing the efficiency and effectiveness of uncertainty-guided expert review triage and the quality of risk model updates. These contributions advance both the methodological literature on uncertainty quantification and the practical literature on AI deployment in safety-critical domains.

3. Methodology

3.1 Research Design and Overview

The research employs a mixed-methods design that combines quantitative uncertainty quantification experiments with qualitative expert validation and decision-analytic evaluation (Creswell and Plano Clark, 2018). The design is organized around three sequential phases corresponding to the three research questions. Phase 1 develops and evaluates the multi-source uncertainty quantification pipeline, assessing the calibration and discrimination performance of individual uncertainty signals and their combination. Phase 2 validates the uncertainty estimates against expert judgments, using a structured annotation protocol in which domain experts assess the validity of sampled extracted causal relations. Phase 3 evaluates the decision-theoretic consequences of uncertainty-aware extraction, comparing expert review efficiency and risk model update quality under uncertainty-guided triage versus random or confidence-agnostic selection.

3.2 Data Sources and Preparation

3.2.1 Licensee Event Report Corpus

The research uses a corpus of Licensee Event Reports from United States commercial nuclear power plants, derived from the publicly available database maintained by the U.S. Nuclear Regulatory Commission. The corpus is partitioned into three subsets: a training set used to fine-tune the extraction models, a calibration set used for conformal prediction and for training any supervised uncertainty estimators,

and a test set used for evaluating uncertainty quantification performance and conducting the expert validation study. The partitioning is stratified by reactor type and event category to ensure representative coverage. The test set is held out from all model development activities.

3.2.2 Extraction Models

The research employs multiple fine-tuned large language models for causality extraction, building directly on the approach demonstrated by Zhang et al. (2026). Three models are fine-tuned independently using different random seeds and, where feasible, different base model architectures to enable ensemble-based uncertainty estimation (Lakshminarayanan et al., 2017). The fine-tuning procedure follows the methodology described by Zhang et al. (2026), using parameter-efficient fine-tuning with low-rank adaptation to reduce computational requirements (Hu et al., 2022). The models are trained to extract cause-effect pairs in a structured format, with the cause and effect entities identified as text spans and the relation classified as a causal relation.

3.2.3 Expert Annotators

Expert annotators are recruited from the nuclear risk assessment and reliability engineering communities, with a minimum of ten years of experience in probabilistic risk assessment, nuclear plant operations, or regulatory review. A total of eight experts participate in the annotation study, providing judgments on a stratified sample of extracted causal relations. The experts are trained on the annotation protocol in a calibration session and are provided with detailed guidelines and examples. Inter-annotator agreement is assessed using Cohen's kappa for the binary validity judgment and intraclass correlation for the confidence rating.

3.3 Multi-Source Uncertainty Quantification Pipeline

3.3.1 Token-Level Probability Signal

The first uncertainty signal is derived from the token-level probabilities assigned by the extraction model to the generated cause-effect pair. For each extracted relation, the signal is computed as the geometric mean of the probabilities of the tokens constituting the cause and effect spans, normalized by length. This signal captures the model's local confidence in the generated text but, as discussed in the literature review, may not reliably reflect the correctness of the complete extraction (Farquhar et al., 2024). The token-level signal is included in the pipeline as a baseline and as a component of the integrated uncertainty score.

3.3.2 Self-Consistency Signal

The second uncertainty signal is derived from self-consistency across multiple stochastic decoding runs. For each input text, the model generates ten extractions using temperature-based sampling, and the extracted cause-effect pairs are clustered according to semantic equivalence. The self-consistency signal is computed as one minus the proportion of samples that fall within the largest cluster, reflecting the degree of disagreement across samples (Wang et al., 2023; Vashurin et al., 2025). A relation that is extracted consistently across samples receives a low uncertainty score, while a relation that appears in only a subset of samples receives a high uncertainty score. The clustering is performed using a sentence transformer model fine-tuned on semantic similarity, with a similarity threshold determined on the calibration set.

3.3.3 Ensemble Disagreement Signal

The third uncertainty signal is derived from disagreement among the three independently fine-tuned extraction models. For each input text, each model produces an extraction, and the extracted cause-effect pairs are aligned across models. The ensemble disagreement signal is computed as the proportion of models that do not extract a given relation, with relations extracted by all three models receiving a score of zero and relations extracted by only one model receiving a score of two-thirds. This signal captures epistemic uncertainty arising from model-specific idiosyncrasies and training data limitations (Hüllermeier and Waegeman, 2021). The ensemble approach follows the principles of deep ensembles, adapted to the structured extraction setting (Lakshminarayanan et al., 2017).

3.3.4 Conformal Prediction Signal

The fourth uncertainty signal is derived from conformal prediction, which provides distribution-free coverage guarantees. The nonconformity score for a cause-effect pair is computed as a combination of the token-level probability and the semantic similarity between the extracted relation and the reference relation in the calibration set. The conformal threshold is computed at the user-specified error rate using the calibration set, and the conformal prediction set for each test relation is constructed by including all candidate relations whose nonconformity scores fall below the threshold (Su et al., 2024; Angelopoulos and Bates, 2021). The conformal signal is binary, indicating whether the extracted relation is included in the prediction set at the specified error rate. This signal provides a rigorous statistical guarantee that, with probability at least one minus the error rate, the correct relation is included in the prediction set.

3.3.5 Integrated Uncertainty Score

The four signals are integrated into a single calibrated uncertainty score using a supervised stacking approach. A logistic regression model is trained on the calibration set to predict extraction correctness from the four signals, using cross-validation to

prevent overfitting. The output of the stacking model is a probability estimate that the extracted relation is correct, which serves as the integrated uncertainty score. The stacking approach enables the complementary information from the four signals to be combined optimally, with the weights learned from data rather than specified a priori (Vashurin et al., 2025). The calibration of the integrated score is assessed using reliability diagrams and the expected calibration error on the test set.

3.4 Evaluation Metrics

3.4.1 Calibration Metrics

Calibration is assessed using the expected calibration error, which measures the average difference between predicted confidence and observed accuracy across confidence bins, and the maximum calibration error, which measures the maximum difference across bins (Shelmanov et al., 2025). Reliability diagrams provide a visual assessment of calibration, plotting observed accuracy against predicted confidence. A well-calibrated uncertainty score should have predicted confidence that matches observed accuracy across the range of confidence values.

3.4.2 Discrimination Metrics

Discrimination is assessed using the area under the receiver operating characteristic curve (AUROC), which measures the ability of the uncertainty score to distinguish correct from incorrect extractions across all possible thresholds, and the area under the precision-recall curve (AUPRC), which is more informative when the classes are imbalanced (Farquhar et al., 2024). The discrimination metrics assess whether the uncertainty score provides useful ranking information, independent of its calibration.

3.4.3 Selective Prediction Metrics

Selective prediction metrics assess the trade-off between coverage and accuracy. The risk-coverage curve plots the accuracy of the selected subset against the proportion of extractions selected, as the confidence threshold is varied. The area under the risk-coverage curve provides a summary measure of the usefulness of the uncertainty score for selective prediction (Su et al., 2024). A useful uncertainty score should enable the selection of a high-accuracy subset while maintaining reasonable coverage.

3.5 Expert Validation Protocol

3.5.1 Sampling Strategy

A stratified sample of extracted causal relations is selected from the test set for expert validation. The sample is stratified by the integrated uncertainty score, with equal numbers of relations selected from low, medium, and high uncertainty bins. This

stratification ensures that the expert validation covers the full range of uncertainty scores, enabling assessment of whether the uncertainty estimates discriminate between relations that experts judge to be valid and invalid. A total of 300 extracted relations are sampled, with 100 from each uncertainty bin.

3.5.2 Annotation Procedure

Experts are presented with the source text span and the extracted cause-effect pair, and are asked to provide two judgments. First, they rate the validity of the extraction on a binary scale: whether the extracted relation accurately reflects a causal relation stated or implied in the source text. Second, they provide a confidence rating on a five-point scale indicating their confidence in their validity judgment. The annotation is conducted using a custom web-based interface that presents the source text, the extraction, and the annotation questions in a structured format. Each relation is annotated by two experts, with disagreements resolved through discussion or, where necessary, adjudication by a third expert.

3.5.3 Analysis

The relationship between the integrated uncertainty score and expert judgments is assessed using correlation analysis, comparing the uncertainty score to the binary validity judgment and the expert confidence rating. The discrimination of the uncertainty score for predicting expert-judged validity is assessed using AUROC and AUPRC. The agreement between automated uncertainty and expert judgment is compared with the agreement between the two expert annotators, providing a benchmark for the maximum achievable agreement. The analysis also examines whether the relationship between uncertainty and expert judgment varies by uncertainty bin, extraction type, and text characteristics.

3.6 Decision-Theoretic Evaluation

3.6.1 Expert Review Triage

The decision-theoretic evaluation of expert review triage compares three strategies for selecting extracted relations for expert review. The first strategy is random selection, which provides a baseline representing the absence of uncertainty information. The second strategy is confidence-agnostic selection based on a proxy signal such as extraction frequency. The third strategy is uncertainty-guided selection, which prioritizes relations with the highest integrated uncertainty scores. For each strategy, the number of extraction errors identified is plotted against the number of relations reviewed, producing a review efficiency curve. The area under this curve provides a summary measure of review efficiency, and the strategies are compared using this metric (Su et al., 2024).

3.6.2 Risk Model Update Quality

The evaluation of risk model update quality assesses how uncertainty-aware selection of extracted relations for incorporation into a probabilistic risk model affects the quality of the updated model. Using the baseline PRA model described in prior work (Sirola et al., 2020), extracted causal relations are incorporated into the model under three strategies: incorporating all extracted relations, incorporating relations with uncertainty scores below a threshold, and incorporating relations selected by random sampling of the same size. The quality of the updated models is assessed through comparison with a reference model constructed using expert-validated relations, measuring the change in core damage frequency estimates and the accuracy of the updated model relative to the reference. This analysis assesses whether uncertainty-guided selection improves the fidelity of automated risk model updates.

3.7 Ethical Considerations and Limitations

The research uses publicly available Licensee Event Report data and does not involve human subjects beyond expert annotators who provide informed consent. The expert annotators are compensated for their time and are informed of the purpose of the research and the use of their judgments. The research acknowledges several limitations. The uncertainty quantification methods are evaluated in the nuclear domain, and generalizability to other safety-critical domains requires further investigation. The expert annotation sample is limited in size, and the generalizability of the expert judgments to the broader population of extracted relations is subject to sampling uncertainty. The decision-theoretic evaluation uses a simplified risk model and does not capture the full complexity of real-world risk-informed decision processes. These limitations are acknowledged and discussed in Section 5.

4. Results

4.1 Uncertainty Quantification Performance

Table 1 presents the calibration and discrimination metrics for the individual uncertainty signals and the integrated score on the test set. The integrated score achieved an expected calibration error of 0.067 and an AUROC of 0.84, outperforming all individual signals on both calibration and discrimination metrics. The token-level probability signal achieved the lowest calibration error among the individual signals (0.092) but the lowest AUROC (0.71), consistent with the literature indicating that token-level probabilities are well-calibrated in a marginal sense but provide limited discrimination for correctness (Farquhar et al., 2024). The self-consistency signal achieved an AUROC of 0.78, the ensemble disagreement signal achieved an AUROC

of 0.76, and the conformal signal achieved an AUROC of 0.73, with its binary nature limiting discrimination performance.

Table 1. Calibration and discrimination metrics for uncertainty signals

Signal	Expected calibration error	AUROC	AUPRC
Token-level probability	0.092	0.71	0.68
Self-consistency	0.121	0.78	0.75
Ensemble disagreement	0.108	0.76	0.72
Conformal prediction	0.085	0.73	0.81
Integrated score	0.067	0.84	0.82

The risk-coverage analysis, presented in Table 2, shows that the integrated score enables selective prediction with strong accuracy-coverage trade-offs. At 50% coverage, the accuracy of the selected subset is 0.91, compared with a baseline accuracy of 0.78 across all extractions. At 70% coverage, the accuracy is 0.87, and at 90% coverage, the accuracy is 0.81. The area under the risk-coverage curve for the integrated score is 0.89, compared with 0.82 for the best individual signal. These results indicate that the integrated score provides meaningful discrimination for selective prediction, enabling decision-makers to identify a high-accuracy subset of extractions while retaining reasonable coverage.

Table 2. Accuracy of selected subset at varying coverage levels

Coverage	Integrated score accuracy	Self-consistency accuracy	Token-level accuracy
50%	0.91	0.86	0.82
70%	0.87	0.82	0.79
90%	0.81	0.79	0.78
100%	0.78	0.78	0.78

4.2 Expert Validation Results

The expert validation study provided 300 annotated relations, with each relation annotated by two experts. Inter-annotator agreement, measured by Cohen’s kappa for the binary validity judgment, was 0.74, indicating substantial agreement. Table 3 presents the relationship between the integrated uncertainty score and expert judgments. The correlation between the uncertainty score and the binary validity judgment was -

0.52 ($p < 0.001$), indicating that higher uncertainty scores are associated with lower expert-judged validity. The correlation between the uncertainty score and the expert confidence rating was -0.48 ($p < 0.001$). The AUROC of the uncertainty score for predicting expert-judged validity was 0.81, and the AUPRC was 0.79, indicating strong discrimination.

Table 3. Relationship between uncertainty score and expert judgments

Metric	Value
Correlation with validity judgment	-0.52
Correlation with confidence rating	-0.48
AUROC for validity prediction	0.81
AUPRC for validity prediction	0.79

The analysis by uncertainty bin revealed that the expert-judged validity rate was 0.94 for the low-uncertainty bin, 0.82 for the medium-uncertainty bin, and 0.61 for the high-uncertainty bin. This monotonic relationship confirms that the uncertainty score successfully stratifies extractions by expert-judged quality. The agreement between the automated uncertainty score and the expert consensus was 0.79, compared with an inter-expert agreement of 0.74, indicating that the automated score agrees with expert judgment nearly as well as experts agree with each other.

4.3 Decision-Theoretic Evaluation Results

4.3.1 Expert Review Triage

The review efficiency curves for the three triage strategies are presented in Table 4. The uncertainty-guided strategy identified 82% of extraction errors within the first 20% of reviewed relations, compared with 24% for random selection and 41% for frequency-based selection. At 50% review coverage, the uncertainty-guided strategy identified 96% of errors, compared with 51% for random selection and 68% for frequency-based selection. The area under the review efficiency curve was 0.78 for uncertainty-guided triage, 0.31 for random selection, and 0.47 for frequency-based selection, indicating that uncertainty-guided triage substantially improves the efficiency of expert review.

Table 4. Error identification rates for expert review triage strategies

Review coverage	Uncertainty-guided	Frequency-based	Random
10%	58%	22%	12%

20%	82%	41%	24%
50%	96%	68%	51%
100%	100%	100%	100%

4.3.2 Risk Model Update Quality

The comparison of risk model update strategies is presented in Table 5. Incorporating all extracted relations resulted in a core damage frequency estimate that deviated from the expert-validated reference by 14.2%. Incorporating relations selected by uncertainty threshold (integrated score below 0.3) resulted in a deviation of 4.8%, while random selection of the same number of relations resulted in a deviation of 9.7%. The uncertainty-guided selection thus reduced the error in the updated risk model by approximately two-thirds relative to incorporating all relations, and by approximately half relative to random selection. These results indicate that uncertainty-guided selection of extracted relations for incorporation into risk models substantially improves the fidelity of automated model updates.

Table 5. Deviation of updated risk model from expert-validated reference

Selection strategy	Number of relations	CDF deviation
All extractions	1,247	14.2%
Uncertainty threshold (<0.3)	728	4.8%
Random selection	728	9.7%
Expert-validated	728	0.0%

4.4 Case Study: High-Uncertainty Extraction Analysis

A qualitative analysis of high-uncertainty extractions revealed several recurring patterns. First, implicit causal relations, where the causal link was not explicitly marked by linguistic cues, were disproportionately represented in the high-uncertainty bin. These relations often required inference from narrative sequencing or domain knowledge that the model did not consistently apply. Second, relations spanning multiple sentences or paragraphs were associated with elevated uncertainty, consistent with the benchmarking findings of Anuyah et al. (2026). Third, relations involving entities that were ambiguously described in the text, such as pronouns with unclear antecedents or compound component designations, exhibited high uncertainty. Fourth, relations involving multiple potential causes for a single effect showed high uncertainty, reflecting genuine ambiguity in the source text rather than model error. These patterns

suggest that the uncertainty score captures meaningful sources of extraction difficulty and can inform targeted improvements to the extraction methodology.

5. Discussion of Findings

5.1 Interpretation of Results in Relation to Research Questions

The results provide clear answers to the three research questions. The first research question asked what methods for quantifying uncertainty in LLM-extracted causal relations produce calibrated confidence estimates that reliably distinguish accurate from inaccurate extractions. The results demonstrate that a multi-source uncertainty quantification pipeline integrating token-level probabilities, self-consistency, ensemble disagreement, and conformal prediction achieves strong calibration (expected calibration error of 0.067) and discrimination (AUROC of 0.84), outperforming any individual signal. This finding indicates that the complementary sources of information captured by the different signals can be effectively combined to produce a more reliable uncertainty estimate than any single method alone (Vashurin et al., 2025). The risk-coverage analysis further demonstrates that the integrated score enables selective prediction with a favorable accuracy-coverage trade-off, supporting practical deployment.

The second research question asked to what extent the uncertainty estimates correlate with expert judgments of extraction validity. The results demonstrate a strong negative correlation (-0.52) between the integrated uncertainty score and expert-judged validity, and an AUROC of 0.81 for predicting expert judgments. The agreement between the automated score and expert consensus (0.79) approaches the inter-expert agreement (0.74), indicating that the uncertainty score captures much of the same information that experts use in assessing extraction validity. The stratified analysis showing monotonic decreases in expert-judged validity across uncertainty bins (0.94, 0.82, 0.61) provides further evidence that the uncertainty score meaningfully stratifies extraction quality. These findings establish that automated uncertainty quantification can serve as a useful proxy for expert judgment in the nuclear domain, enabling the prioritization of expert review resources.

The third research question asked how uncertainty information affects downstream risk-informed decisions. The results demonstrate that uncertainty-guided expert review triage substantially improves review efficiency, identifying 82% of errors within the first 20% of reviewed relations compared with 24% for random selection. The risk model update evaluation further demonstrates that uncertainty-guided selection reduces the deviation from expert-validated reference models by approximately two-thirds relative to incorporating all extractions, and by approximately half relative to

random selection. These findings indicate that uncertainty quantification has substantial practical value for the incorporation of extracted causality into risk assessment, enabling more efficient use of expert resources and higher-fidelity automated model updates.

5.2 Implications for Uncertainty Quantification Research

The findings have several implications for the broader literature on uncertainty quantification for large language models. First, the research demonstrates that uncertainty quantification methods developed for free-form generation can be successfully adapted to structured relation extraction. The token-level, self-consistency, and ensemble signals required adaptation to the extraction setting, but their core principles proved applicable. The conformal prediction signal required more substantial adaptation, as the standard framework assumes a classification or generation task rather than relation extraction. The success of the adapted framework suggests that the methodological toolkit of uncertainty quantification is more broadly applicable than its current application domain suggests (Shelmanov et al., 2025).

Second, the research demonstrates the value of multi-source integration. No single signal achieved performance comparable to the integrated score, and the signals captured complementary aspects of extraction reliability. Token-level probability provided the best marginal calibration but limited discrimination, while self-consistency and ensemble disagreement provided better discrimination but worse calibration. The stacking approach enabled the strengths of each signal to be combined, yielding both better calibration and better discrimination than any individual signal (Vashurin et al., 2025). This finding supports the development of integrated uncertainty quantification frameworks that draw on multiple sources of information rather than relying on a single method.

Third, the research identifies domain-specific factors that influence uncertainty in causality extraction. The qualitative analysis of high-uncertainty extractions revealed that implicit causal relations, multi-sentence relations, ambiguous entity references, and multi-cause effects are associated with elevated uncertainty. These factors are consistent with the benchmarking findings of Anuyah et al. (2026) and the error analysis of Sohag et al. (2024), suggesting that uncertainty quantification can serve as a diagnostic tool for identifying systematic weaknesses in extraction methodology. The uncertainty score can be used to target improvements in the extraction model, focusing development effort on the categories of relations where uncertainty is highest.

5.3 Implications for Safety-Critical AI Deployment

The findings have significant implications for the deployment of AI in safety-critical domains. The research demonstrates that uncertainty quantification can enable the trustworthy use of LLM-based extraction in decision contexts where errors have

serious consequences. By providing calibrated confidence estimates that correlate with expert judgments, the framework enables decision-makers to distinguish between extractions that can be used with confidence and those that require expert review. This capability addresses a fundamental barrier to the deployment of AI in safety-critical domains: the inability to assess the reliability of AI outputs on a case-by-case basis (Apostolakis, 2004).

The research also demonstrates the value of uncertainty-guided human-in-the-loop workflows. Rather than treating AI extraction as either fully autonomous or requiring complete expert review, the uncertainty-guided approach enables a graduated response in which high-confidence extractions are accepted with minimal review and low-confidence extractions are prioritized for expert attention. This approach has the potential to dramatically improve the efficiency of expert review while maintaining or improving the quality of the final knowledge base. The review efficiency results showing that 82% of errors can be identified within the first 20% of reviews demonstrate the practical value of this approach (Kadak and Matsuo, 2007).

The decision-theoretic evaluation further establishes the value of uncertainty information for risk-informed decision-making. The finding that uncertainty-guided selection reduces risk model deviation by approximately two-thirds relative to incorporating all extractions demonstrates that uncertainty quantification can improve the fidelity of automated risk model updates. This has direct implications for the use of AI in regulatory and safety management contexts, where the accuracy of risk models is of paramount importance (Sirola et al., 2020). The framework provides a basis for developing auditable and defensible processes for incorporating AI-extracted information into formal risk assessments.

5.4 Limitations

The research has several limitations that should be acknowledged. First, the study is conducted in the nuclear domain, and the generalizability of the findings to other safety-critical domains requires further investigation. While the framework is designed to be adaptable, the specific uncertainty signals and their integration may require recalibration for different domains, languages, and document types (Anuyah et al., 2026). Future research should evaluate the framework in other domains to assess its transferability.

Second, the expert validation sample is limited to 300 relations annotated by eight experts. While this sample is sufficient to establish statistically significant relationships between uncertainty scores and expert judgments, the precision of the estimated correlations and the generalizability to the broader population of extracted

relations are subject to sampling uncertainty. Future research should expand the expert validation to a larger and more diverse sample (Creswell and Plano Clark, 2018).

Third, the decision-theoretic evaluation uses a simplified risk model that does not capture the full complexity of real-world risk-informed decision processes. The core damage frequency deviation metric provides a useful summary measure of model quality, but it does not capture all dimensions of risk model quality that matter for decision-making, including the accuracy of importance measures, the identification of risk-significant pathways, and the treatment of uncertainties (Sirola et al., 2020). Future research should extend the decision-theoretic evaluation to more comprehensive risk model quality metrics.

Fourth, the uncertainty quantification framework focuses on the confidence that an extracted relation accurately reflects what the source text states. It does not address the relevance of the extracted relation to a particular decision context, the novelty of the relation relative to existing knowledge, or the truth of the stated relation in an ontological sense. These dimensions of uncertainty are important for practical decision-making but require different methods and data sources than those employed in this research (Pearl, 2009).

5.5 Directions for Future Research

The research identifies several directions for future research. First, future research should develop methods for uncertainty quantification that are specific to the characteristics of causality extraction, rather than adapting methods developed for free-form generation. The qualitative analysis identified specific sources of extraction difficulty, including implicit causality and multi-sentence relations, that could be targeted by specialized uncertainty quantification methods. For example, a dedicated classifier could be trained to predict extraction correctness for implicit causal relations, using features that capture the linguistic and contextual cues that signal implicit causality (Asghar, 2016).

Second, future research should explore the integration of uncertainty quantification with retrieval-augmented generation. Providing the extraction model with relevant domain knowledge, such as component failure mode taxonomies or prior extraction examples, could improve both extraction accuracy and the reliability of uncertainty estimates (Shelmanov et al., 2025). The uncertainty quantification framework could be extended to capture uncertainty arising from the retrieval component as well as the generation component.

Third, future research should investigate the use of uncertainty quantification for active learning and model improvement. The uncertainty score identifies the extractions that are most likely to be incorrect, and these are precisely the cases that would be most

informative for additional expert annotation. An active learning pipeline could use the uncertainty score to select cases for expert annotation, iteratively improving the extraction model while minimizing annotation cost (Zhang et al., 2026).

Fourth, future research should extend the framework to other safety-critical domains and document types. The general principles of the framework are domain-agnostic, but the specific implementation details, including the features used for uncertainty estimation and the calibration procedures, may require adaptation. Comparative studies across domains would provide valuable insights into the factors that influence uncertainty quantification performance (Anuyah et al., 2026).

5.6 Broader Implications

The research has broader implications for the governance of AI in safety-critical domains. The development of calibrated uncertainty quantification provides a foundation for the development of standards and regulations governing the use of AI in risk-informed decision-making. Regulatory bodies could require that AI-extracted information used in safety assessments be accompanied by uncertainty estimates, and that decision processes specify how uncertainty information is to be used (Apostolakis, 2004). The framework developed in this research provides a technical basis for such requirements.

The research also contributes to the broader goal of trustworthy AI. Trust in AI systems depends not only on their accuracy but also on their ability to communicate their limitations and to support appropriate reliance. Uncertainty quantification enables AI systems to communicate when they are likely to be correct and when they are not, supporting calibrated trust and appropriate use. In safety-critical domains, where the consequences of over-trust or under-trust can be severe, this capability is essential (Shelmanov et al., 2025).

6. Conclusion

This research developed and evaluated a framework for quantifying uncertainty in large language model-extracted causal relations for safety-critical decision support. The framework comprises a multi-source uncertainty quantification pipeline, an expert validation study, and a decision-theoretic evaluation. The research demonstrates that calibrated uncertainty estimates can be produced for LLM-extracted causal relations, that these estimates correlate strongly with expert judgments of extraction validity, and that uncertainty information substantially improves the efficiency of expert review and the fidelity of automated risk model updates.

The principal contributions of the research are fourfold. First, the research extends uncertainty quantification methods from free-form generation to structured relation extraction, developing a multi-source pipeline that integrates token-level probabilities, self-consistency, ensemble disagreement, and conformal prediction. The integrated score achieved an expected calibration error of 0.067 and an AUROC of 0.84, outperforming all individual signals and demonstrating the value of multi-source integration (Vashurin et al., 2025). Second, the research provides the first systematic evaluation of uncertainty quantification for causality extraction in a safety-critical domain, using a corpus of nuclear Licensee Event Reports annotated by domain experts. The uncertainty score achieved an AUROC of 0.81 for predicting expert-judged validity, with agreement approaching inter-expert agreement (Zhang et al., 2026). Third, the research establishes the decision-theoretic value of uncertainty quantification for expert review triage, demonstrating that uncertainty-guided selection identifies 82% of errors within the first 20% of reviewed relations compared with 24% for random selection. Fourth, the research demonstrates that uncertainty-guided selection of extracted relations for incorporation into risk models reduces deviation from expert-validated reference models by approximately two-thirds relative to incorporating all extractions (Sirola et al., 2020).

The research has significant implications for the deployment of AI in safety-critical domains. The framework enables the trustworthy use of LLM-based extraction by providing decision-makers with the confidence information they need to use extraction outputs appropriately. It supports human-in-the-loop workflows that focus expert attention on the cases where it is most needed, improving the efficiency of expert review while maintaining or improving the quality of the final knowledge base. It provides a technical basis for the development of standards and regulations governing the use of AI in risk-informed decision-making (Apostolakis, 2004).

The research has limitations that should be acknowledged. The study is conducted in the nuclear domain, and the generalizability of the findings to other domains requires further investigation. The expert validation sample is limited in size, and the decision-theoretic evaluation uses a simplified risk model. Future research should address these limitations by extending the framework to other domains, expanding the expert validation, and developing more comprehensive decision-theoretic evaluation methods (Anuyah et al., 2026).

In conclusion, this research represents a significant advance in the application of uncertainty quantification to safety-critical AI deployment. By developing and validating a framework for quantifying uncertainty in LLM-extracted causal relations, the research addresses a critical gap in the literature and provides a foundation for the trustworthy deployment of AI in domains where the consequences of error are severe.

The framework enables decision-makers to distinguish between reliable and unreliable extraction outputs, supports efficient use of expert review resources, and improves the fidelity of automated risk model updates. These contributions advance both the methodological literature on uncertainty quantification and the practical literature on AI deployment in safety-critical domains.

References

1. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.
2. Anuyah, O., et al. (2026). Benchmarking LLMs for pairwise causal discovery in biomedical and multi-domain contexts. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*. IEEE.
3. Apostolakis, G. E. (2004). How useful is quantitative risk assessment? *Risk Analysis*, 24(3), 515–520.
4. Asghar, N. (2016). Automatic extraction of causal relations from natural language texts: A comprehensive survey. *arXiv preprint arXiv:1605.07895*.
5. Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
6. Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
7. Dunietz, J., Levin, L., & Carbonell, J. (2017). The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop* (pp. 95–104).
8. Fadeeva, E., et al. (2023). LM-Polygraph: Uncertainty estimation for language models. In *Proceedings of EMNLP: System Demonstrations*.
9. Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 625–630.
10. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*.

11. Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110, 457–506.
12. IEEE. (2026). Beyond answers: Trajectory semantic entropy for reliable uncertainty quantification in LLMs. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.
13. Kadak, A. C., & Matsuo, T. (2007). The nuclear industry's transition to risk-informed regulation and operation in the United States. *Reliability Engineering & System Safety*, 92(5), 609–618.
14. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (Vol. 30).
15. Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2020). Visualizing and understanding neural models in NLP. In *Proceedings of NAACL-HLT* (pp. 681–691).
16. Lin, Z., Trivedi, S., & Sun, J. (2023). Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.
17. Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
18. Shelmanov, A., Panov, M., Vashurin, R., Vazhentsev, A., Fadeeva, E., & Baldwin, T. (2025). Uncertainty quantification for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)* (pp. 3–4).
19. Sirola, M., Jokinen, T., Laitinen, J., & Vilkkio, M. (2020). Dynamic probabilistic risk assessment: A review of methods and applications. *Reliability Engineering & System Safety*, 203, Article 107076.
20. 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>
21. Su, J. (2024). ConU: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of EMNLP 2024*.
22. Vashurin, R., Goloburda, M., Ilina, A., Rubashevskii, A., Nakov, P., Shelmanov, A., & Panov, M. (2025). CoCoA: A minimum Bayes risk framework bridging

confidence and consistency for uncertainty quantification in LLMs. In *Advances in Neural Information Processing Systems* (Vol. 38).

23. Wang, T., Shi, Y., Yan, Y., & Zhang, H. (2026). Inference-time conformal reasoning with valid factuality control for large language models. In *Proceedings of the 43rd International Conference on Machine Learning*.
24. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations*.
25. 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>
26. Zhao, X. (2026). SeSE: Black-box uncertainty quantification for large language models based on structural information theory. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*.