LLM-Based Extraction of Causal Relationships among Organizational Factors for Extending Human Reliability Analysis

Tingting Cheng[a], David Doberstein[a,b], Ali Mosleh[a]
[a] Garrick Institute for the Risk Sciences, UCLA, Los Angeles, United States
[b] Department of Computer Science, UCLA, Los Angeles, United States
tingtingc@ucla.edu, ddoberstein@ucla.edu, mosleh@ucla.edu

Abstract: Organizational factors (OFs) play a critical role in shaping human performance and contributing to human error in many safety-critical fields. Human Reliability Analysis (HRA) is a discipline aiming to identify, model, and assess the contribution of human errors to system risk. However, many HRA methods represent OFs either as lists of influencing factors without explicit causal pathways or through causal structures that rely heavily on expert judgment. Although the effects of OFs on human performance have been extensively investigated across safety-critical domains, such as healthcare, energy, aviation and aerospace, and maritime and offshore operations, the relevant evidence is dispersed across accident investigations, regulatory reports, and research publications. This evidence is predominantly unstructured, often expresses causality implicitly and contextually, and may describe similar relationships using different terminology, making systematic identification and synthesis burdensome. To address this gap, our broader research aims to develop a generic, data-driven causal model of organizational influences on human performance. As part of the broader research effort, this paper presents a large language model (LLM)-based pipeline for extracting textual evidence of causal relationships among OFs. A corpus of 266 publications collected in a previous study was processed using PDF-based text extraction and a structured seven-step procedure specifying the extraction task, output schema, formatting rules, causal terminology, sample outputs, definitions of 19 OFs, and 38 target causal relationships. The OF definitions were developed through expert-led literature review and knowledge-based synthesis, while the target relationships were identified a priori based on expert knowledge.
Three LLMs were evaluated using a common extraction task. GPT-5-mini was selected for full-corpus processing because it achieved extraction quality comparable to GPT-5 and Grok-4 at lower computational cost. The pipeline identified 1,538 causal instances, of which 1,482 (96.4%) matched the predefined target relationships. Evidence was found for all 38 relationships with frequencies varying substantially, such as seven occurred more than 50 times, 24 occurred 20–50 times, and seven occurred fewer than 20 times. Information exchange and coordination affecting Decision framing and execution and Employee training and development affecting Knowledge and abilities were the most frequently identified relationships, with 116 and 106 instances, respectively. The extracted causal evidence and frequency-based importance rankings can be used to support construction of a Bayesian belief network causal structure. Through further prompt engineering and literature supplementation, the pipeline was also extended to extract evidence describing how different factor states contribute to human-performance degradation, supporting probabilistic model development presented in a companion study. Limitations included occasional weak or incoherent supporting excerpts and fabrication that need further human review. Overall, the results demonstrate that structured LLM extraction can transform dispersed qualitative evidence into a traceable dataset for reviewing OF causal models and supporting Bayesian network development.

1. Introduction

Human performance is a critical determinant of safety in nuclear power, aviation, chemical processing, maritime, transportation, and other safety-critical systems. Major accidents rarely arise from a single equipment failure; rather, they emerge from interactions among technical conditions, human actions, organizational behaviors, communication processes, training, procedures, and management controls [1]. Human reliability analysis (HRA) provides methods for identifying human failure events (HFEs),

characterizing the conditions that shape performance, and estimating human error probabilities (HEPs) for integrating into the probabilistic risk assessment (PRA). However, organizational influences are difficult to represent because they are interdependent, difficult to measure, and often supported by heterogeneous qualitative evidence [2].

Existing studies often represent OFs as scattered or context-specific factors affecting human performance across miscellaneous safety-critical fields, with limited representation of the causal pathways through which these factors interact and affect performance [2]. Recent studies exploring causality among OFs commonly rely on expert judgment and domain knowledge [3]. Expert knowledge is indispensable, but an exclusively judgment-based structure can be difficult to reproduce, audit, and update. Two analysts may select different links, interpret the same literature differently, or assign different weights to evidence from different domains. These challenges become more important when OF models are connected to performance-influencing factors (PIFs), cognitive failure mechanisms, and HEP quantification [4]. A causal link included upstream in an HRA model can affect downstream inferences even when its empirical basis is weak or opaque [5].

A substantial body of evidence that could strengthen OF causal models already exists. Academic publications, accident investigations, operating-experience reports, regulatory documents, and organizational studies describe, for example, how leadership influences group climate, how policy affects training and resource allocation, how information exchange affects decision framing, and how training develops knowledge and abilities. However, this evidence is predominantly unstructured. Causal meaning may be stated explicitly through expressions such as “leads to,” “results in,” or “contributes to,” but it may also be implicit, distributed across multiple sentences, conditional on context, or expressed through domain-specific terminology. Similar causal mechanisms may be described using different words across industries, while the same terms may represent different constructs in different disciplinary contexts.

Manual literature review provides high-quality interpretation but is resource-intensive and difficult to scale to hundreds of documents. A reviewer must identify relevant passages, determine whether they express causality rather than association, map varying terminology to common OF definitions, assign causal direction, and record why the evidence supports a particular relationship. Keyword search and rule-based text mining can improve retrieval efficiency, but they perform poorly when causality is implicit or when the terminology used in the source differs from that of the target ontology. Co-occurrence of two factor names is also insufficient to establish a causal relationship.

Large language models (LLMs) offer a potential means of converting this dispersed evidence into structured, reviewable records. Their utility, however, depends on disciplined task definition and explicit evidence-selection requirements [6]. Unconstrained prompts can produce plausible but unsupported relationships, normalize terminology inconsistently, or generate explanations that exceed the source text [6]. Recent work on LLM-supported HRA similarly emphasizes that model outputs must remain grounded, traceable, and subject to human review [7]. The present work addresses a narrower but foundational task: extracting textual evidence for a predefined OF causal structure rather than allowing the LLM to freely construct a causal model. Thus, the LLM serves as an evidence-extraction mechanism operating within expert-defined conceptual boundaries, rather than as an independent generator of the causal model.

As part of a broader research effort to develop a generic, knowledge- and data-driven causal model of organizational influences on human performance [8-11], this paper makes three contributions. First, it presents a reproducible LLM-based pipeline that extracts direct textual evidence for 38 predefined causal relationships among 19 OFs from a corpus of 266 publications across multiple safety-critical domains. Second, it evaluates model selection and prompt design for this constrained causal-extraction task. Third, it discusses how the resulting traceable linguistic dataset can support the review, validation, and refinement of OF models for HRA and subsequent Bayesian belief network (BBN) development, while clarifying why extraction frequency alone is not a quantitative measure of causal strength.

The remainder of this paper is organized as follows. Section 2 introduces the OF causal-modeling framework and defines the scope of the extraction task. Section 3 describes the research method and processing workflow. Section 4 presents the results, Section 5 discusses their implications and limitations, and Section 6 concludes the study.

2. OF Causal Modeling Framework and Extraction Scope

2.1. Role within the broader OF-modeling framework

The present study is one component of a broader framework for systematically incorporating OFs into HRA through causal and probabilistic modeling [10,11]. In the broader research, OFs are represented using complementary organizational perspectives, including organizational characteristics and organizational hierarchical levels. Organizational characteristics distinguish structural, process, and behavioral dimensions, while organizational levels distinguish the Strategic Apex, Middle Line Management, and Operating Core. These perspectives support attribution of OFs to organizational functions and provide principles for defining plausible causal directions.

The broader framework combines two sources of knowledge. First, organizational theory, human-factors knowledge, and expert literature review are used to define OF constructs and develop a candidate causal structure. Second, linguistic evidence extracted from published sources is used to examine whether and how those causal relationships are supported in documented studies. The resulting evidence is intended to improve the traceability and reviewability of the causal structure rather than replace expert judgment.

This paper focuses only on the linguistic data-extraction component. Detailed development of the organizational classification, knowledge-driven causal model, and probabilistic BBN is reported separately [10,11].

2.2. Predefined extraction ontology

The extraction ontology included definitions for 19 OFs and 38 directed causal relationships. Examples of target relationships include: 1) Leadership → Group Climate; 2) Policy design → Employee Training and Development; 3) Information Exchange and Coordination → Decision Framing and execution; and 4) Employee Training and development → Knowledge/Abilities.

The OF definitions and target causal directions were established before full-corpus extraction. They were developed through expert-led literature review, organizational and human-factors knowledge, and synthesis of the broader OF-modeling framework. One example is as follows.

The OF *Leadership* refers to *the leaders' culture, style, engagement, and skills through which leaders influence employees' motivation, safety attitudes, and team performance. It encompasses how leaders set direction, communicate expectations, allocate resources, and model desired behaviors, shaping how safety and responsibility are prioritized across the organization. Effective leadership fosters trust, accountability, and continuous improvement, serving as a central mechanism linking organizational goals with human performance outcomes*. Its synonyms include *leadership, leadership culture, leadership style, leadership engagement, leadership skills, leadership behavior, leadership practice, leadership commitment, organizational leadership, safety leadership, supervisory leadership, supervision, managerial leadership, management, oversight, command, leader influence, leader behavior*.

The ontology bounded the extraction task. The model was not asked to identify any relationship it considered plausible. Instead, it was instructed to return source passages supporting one or more of the 38 predefined relationships. This design served three purposes. First, it reduced unconstrained causal

generation. Second, it provided consistent labels for aggregating evidence across sources. Third, it allowed the amount and distribution of retrieved evidence to be compared across the target structure.

3. Method

3.1. Research corpus

The input corpus consisted of 266 publications addressing organizational factors, human performance, safety management, HRA, and safety-critical operations [9]. The corpus was assembled in a prior study investigating organizational influences across 9 domains, including nuclear power, aviation and aerospace, healthcare, oil and gas, maritime and offshore, transportation etc. The corpus included peer-reviewed publications, technical reports, and so forth. Publications were selected based on their relevance to organizational determinants of human performance, safety behavior, accident causation, or organizational risk management. Each publication was processed as an independent document-level input. The present analysis reports the number of extracted instances across the complete corpus and their distribution among the 38 target relationships.

3.2. LLM processing pipeline

PDFMiner was used to parse each PDF and extract raw text. The extracted text was passed to a LangChain workflow using ChatPromptTemplates. The selected LLM processed each document and returned a structured JSON array. Each JSON object contained three required fields: (1) a causality label corresponding to one target relationship, (2) a direct excerpt from the source, and (3) a justification explaining why the excerpt supported the assigned relationship. Results were written incrementally to a comma-separated-value file after each document so that completed outputs were preserved if processing was interrupted.

Each source document was converted from PDF format into machine-readable text using PDFMiner. The resulting raw text was prepared for LLM processing. In the implement workflow, reference lists tables and figure captions; equations; headers and footers; page numbers; duplicated text; special characters; and encoding artifacts were removed. The quality of PDF-to-text conversion is important because layout loss, scanned pages, broken sentence boundaries, and character-encoding errors can affect the model's ability to identify causal evidence. Each processed document was passed to the LLM workflow as an independent analysis unit. The LLM-based workflow is shown as Figure 1.

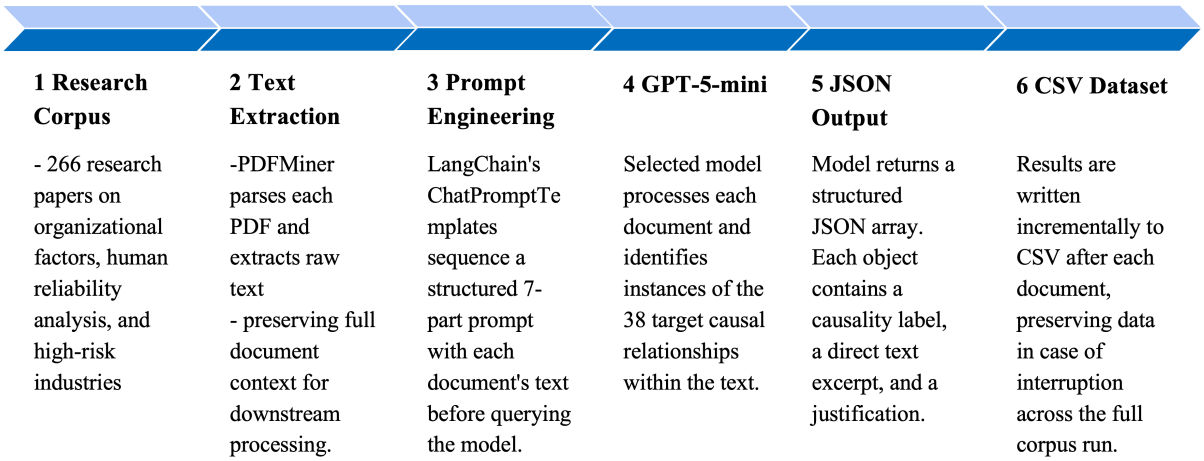

Figure 1. LLM-based causal-evidence extraction workflow

3.3. Prompt engineering step

This subsection describes prompt engineering in detail because, in a constrained causal-evidence extraction task, extraction quality, consistency, and traceability are strongly influenced by how the task

boundaries, causal definitions, factor terminology, target relationships, and output requirements are encoded in the prompt [12]. Therefore, this step was organized into seven components:
1) Task definition: instructing the model to identify textual evidence supporting the predefined relationships rather than summarize the document or generate new causal structures;
2) JSON output schema: standardizing outputs and enabled automated parsing;
3) Formatting and consistency rules: requiring consistent factor labels and record structure;
4) Definition of causal relationships and causal terminology: distinguishing directional causal evidence from simple co-occurrence or association. Common causal terms were provided as guidance, but the model was also permitted to identify implicitly expressed causal mechanisms.
5) Sample output: demonstrating the intended level of specificity;
6) Definitions of the 19 OFs: enabling the model to map synonymous and domain-specific terminology to the common ontology;
7) List of the 38 target relationships: constraining outputs to the predefined causal structure.

3.4. Pilot model comparison and selection

Grok-4, GPT-5, and GPT-5-mini were compared on the same test document and prompt. The pilot comparison considered:

- number of relevant relationships identified;
- completeness and specificity of supporting excerpts;
- depth and coherence of the justification; and
- computational cost.

Grok-4 identified three relationships and generally returned shorter excerpts and tenuous explanations. GPT-5 identified seven relationships with detailed excerpts and high-quality justifications. GPT-5-mini identified eight relationships with similarly detailed excerpts and explanations. Because GPT-5-mini matched or exceeded the observed extraction performance of GPT-5 in the pilot comparison at substantially lower cost, it was selected for full-corpus processing.

This pilot comparison supported an engineering selection decision but was not a formal benchmark. The models were evaluated on a limited common task, and no independently labeled gold standard was used. Consequently, the comparison should not be interpreted as a general ranking of the models.

3.5. Output processing and evaluation

The JSON records returned by the model were parsed and aggregated across the 266 documents. Each causality label was compared against the enumerated list of 38 target relationships. Records were assigned to one of two categories:

- target-matching instances, whose labels corresponded to one of the predefined directed relationships; or
- out-of-scope instances, which combined valid OF names into a relationship not included in the target list.

The number of instances was then calculated for each target relationship. Relationships were grouped into three descriptive frequency categories:

- high frequency: more than 50 instances;
- moderate frequency: 20–50 instances; and
- low frequency: fewer than 20 instances.

These thresholds were used to summarize the distribution of retrieved evidence and should not be interpreted as thresholds for causal importance or statistical significance.

4. Results

4.1. Corpus-level extraction

The full-corpus pipeline generated 1,538 candidate causal instances across the 266 publications. Of these, 1,482 records matched one of the 38 predefined target relationships, yielding a target-match rate of 96.4%. The remaining 56 instances, or 3.6%, represented relationships that were outside the predefined target structure. At least one candidate evidence record was retrieved for every target relationship.

Table 1: Summary of extraction results

Measure	Result
Documents processed	266
Candidate instances	1,538
Instances matching target links	1,482 (96.4%)
Out-of-scope relationship labels	56 (3.6%)
Target relationships with at least one instance	38 of 38

The 96.4% result indicates that the structured prompt and predefined ontology strongly constrained the generated relationship labels. However, it should not be interpreted as 96.4% extraction accuracy because the excerpts and causal interpretations were not all independently validated.

4.2. Distribution across target relationships

The extracted evidence was unevenly distributed across the 38 target relationships. Seven relationships occurred more than 50 times. Twenty-four relationships occurred between 20 and 50 times, and seven relationships occurred fewer than 20 times.

Table 2. Frequency categories among the 38 target relationships

Frequency category	Definition	Number of relationships
High	More than 50 instances	7
Moderate	20–50 instances	24
Low	Fewer than 20 instances	7

The distribution indicates that certain organizational mechanisms are repeatedly described in the corpus, whereas others are less frequently represented. A low extraction frequency indicates relatively limited textual support, whereas a high frequency reflects broader and repeated support in the literature. Overall, the distribution was consistent with expert expectations regarding the relative prominence of the causal relationships. Figure 2 shows the number of extracted causals instances for each of the 38 target relationships.

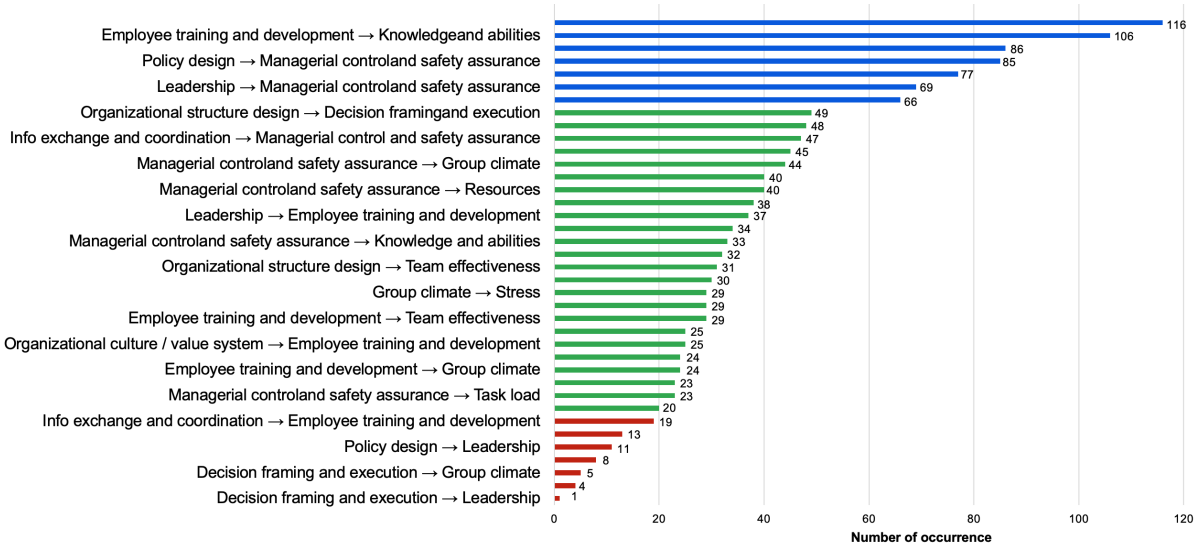

Figure 2. Number of extracted causal instances for each of the 38 target causal relationships

4.3. Illustrative relationship frequencies

The most frequently extracted relationship was: Information Exchange and Coordination→ Decision Framing and Execution, with 116 instances. This relationship captures evidence that the availability, timeliness, completeness, and quality of exchanged information affect how decision makers understand a situation, interpret alternatives, and frame problems. The second most frequently extracted relationship was Employee Training and Development → Knowledge/Abilities, with 106 instances. This result is substantively consistent with the role of training in developing task-related knowledge, skills, competencies, and preparedness.

At the other end of the distribution, Decision Framing and execution → Leadership appeared only once. The low count may reflect an uncommon bottom-up relationship, limited alignment between the terminology in the sources and the ontology, or a corpus that more frequently describes top-down organizational influences.

The observed counts represent the amount of evidence retrievable under the present corpus, ontology, prompt, and model. They do not represent effect sizes, causal coefficients, conditional probabilities, or objective rankings of safety significance.

Table 3. Recommended format for illustrative extracted evidence (example)

Target relationship	Direct source excerpt	LLM justification	Human-review status	Human comments
Employee training and development → Team effectiveness	Improved inter-group communication and inclusive training could help increase mutual understanding of the safety and work needs of both groups and might improve transparency of the work system for technicians.	This sentence explicitly claims that inclusive training improves mutual understanding and system transparency between groups, which are aspects of team effectiveness.	Accept	-
Org structure design → Information exchange	the main factors contributing to the formation of information-seeking ties are both formal organizational structure (as reflected in the	This sentence explicitly states that formal organizational structure contributes to (i.e., causes) the formation of	Accept	-

and coordination	position of the supervisory/administrative division) and informal and tacit dynamics (as reflected in the effects of friendship and EBP scores).	information-seeking ties, which correspond to information exchange and coordination.		
Policy design → Decision framing and execution	Such an expedited process may enable HOF research to better align to the quick technical developments and technology implementation processes that are the subject of the study.	This sentence indicates that adopting an expedited policy/process would enable HOF research to align (i.e., influence how research is planned and executed), showing policy design affects decision framing and execution.	Reject	The excerpt describes the significance of the research but does not support the target relationship

5. Discussion

5.1. Value for organizational-factor modeling

The primary contribution of the extracted dataset is traceability. Instead of documenting a causal link only as an expert judgment, analysts can associate the link with a set of excerpts, source documents, and explicit justifications. This does not eliminate expert judgment; rather, it changes the evidence available to the expert. Analysts can examine whether the evidence supports the proposed direction, determine whether it is domain-relevant, and identify contradictory or conditional findings. The dataset can also support model revision. Links with abundant, high-quality evidence may be retained with greater confidence. Rare links can be prioritized for targeted searches or expert elicitation. Frequently extracted relationships not included in the original structure can be examined as candidate extensions, although they should not be added solely because an LLM generated them. Differences across industries may reveal where a generic OF model requires domain-specific adaptations.

5.2. Relationship to HRA and Bayesian-network development

In an HRA application, OFs are most useful when connected to mechanisms that influence human performance, such as PIFs, cognitive functions, or HFE dependency. A BBN provides one possible representation because it can encode directed dependencies, combine heterogeneous evidence, propagate observations, and represent uncertainty. The extracted records can support the qualitative BBN structure and provide an evidence inventory for conditional probability table (CPT) development.

However, the observed frequency distribution should not be inserted directly into CPTs. Publication frequency is not equivalent to conditional probability. A defensible path from text to BBN parameters would require at least: coding of factor states and contextual conditions; distinction between direct and mediated effects; treatment of contradictory evidence; and expert review. Frequencies may serve as one input to prior elicitation or as a screening indicator for where stronger empirical support exists, but they must be transformed through a transparent evidence-synthesis method.

5.3. Extraction challenges

Three recurrent challenges were identified. First, the model occasionally produced out-of-scope relationships by combining two valid OFs into a causal link that was absent from the target list. Fifty-six of 1,538 records had this form. This behavior shows that including a target list reduces but does not eliminate generative drift. Second, spelling and abbreviation inconsistencies complicated automated

matching. Examples included replacing an ampersand with "and" or expanding "Info" to "Information." These are semantically minor but operationally important when thousands of records are processed. Third, some records contained excerpts that were irrelevant, insufficiently causal, or incoherent with the justification.

5.3 Extension to state-level evidence

Relationship-level evidence establishes that one OF may influence another, but it is insufficient for BBN quantification. Probabilistic modeling requires evidence concerning how defined states of a parent factor affect defined states of a child factor. For example, the general relationship *Employee Training and Development → Knowledge/Abilities* must be refined into state-level propositions, such as inadequate training increases the likelihood of insufficient task knowledge; incomplete training reduces preparedness for abnormal conditions; or high-quality recurrent training improves knowledge retention and response capability.

Following the relationship-level extraction reported in this paper, the broader research extended the pipeline through additional prompt engineering and literature supplementation to extract state-level evidence describing human-performance degradation. For example, parent-factor state, child-factor state, effect polarity, effect strength are extracted serving to populate CPTs of BBN.

5.4. Comparison with structured LLM support for HRA

Studies suggest that LLM-supported HRA should use controlled sources, typed intermediate products, explicit uncertainty states, and review gates rather than unconstrained free-text generation [13]. The present study is consistent with that principle. The seven-part prompt, closed target ontology, JSON schema, direct excerpts, and incremental dataset create a constrained extraction task with auditable intermediate products. Nevertheless, additional source evaluation and human validation are required because LLM-extracted records represent candidate rather than validated evidence, and their content may vary with the model, prompt, and generation settings.

A useful extension would treat each extracted instance as a typed evidence object containing a stable relationship ID, source identifier, page or section location, exact excerpt, causal direction, evidence type, domain, contextual conditions, extraction model and prompt version, validation status, and reviewer disposition. Such a representation would allow the dataset to be integrated into a governed knowledge base or retrievable HRA workflow while preserving the distinction between LLM-extracted candidate evidence and human-accepted evidence.

5.5. Limitation and future work

This study has several limitations. First, the corpus composition has not yet been characterized in sufficient detail to fully assess coverage and selection bias. Second, the model comparison was based on a limited qualitative pilot rather than a standardized benchmark. The reported 96.4% target-match rate reflects conformity with the predefined relationship labels, not factual correctness or causal validity, because a valid label may still be paired with weak or insufficient evidence. Full-document prompting may also be affected by context-window constraints, PDF parsing errors, and loss of document layout, while the extraction process may favor causal relationships stated explicitly in English-language prose. In addition, multiple publications may repeat the same theory or cite the same primary study, causing raw frequencies to overstate independent evidentiary support. Reproducibility therefore requires documentation of model versions, generation settings, prompt versions, preprocessing code, and the complete source corpus.

Future work will focus on improving extraction accuracy and traceability. Specifically, semantic chunking, embeddings, and similarity-based retrieval can further improve source localization and reduce the need to repeatedly process full documents. Formal validation of these evidences should involve at least two qualified reviewers independently assessing relationship labels, causal direction,

excerpt sufficiency, and justification quality, with inter-rater agreement and final acceptance rates reported.

6. Conclusion

This study developed an LLM-based pipeline for extracting causal evidence relevant to organizational-factor modeling in HRA. Using a structured seven-part prompt and a predefined ontology of 19 OFs and 38 directed relationships, GPT-5-mini processed 266 publications and produced 1,538 candidate instances. Of these, 1,482 matched the target structure, and evidence was found for every predefined relationship. The results demonstrate that LLMs can substantially reduce the clerical burden of locating and structuring qualitative causal evidence while preserving direct excerpts and explanations for review. The contribution is an evidence foundation, not an automated causal model or quantified HRA. Raw extraction counts do not represent causal strengths or probabilities, and model outputs require validation, source governance, deduplication, and expert interpretation. With these controls, the framework can support transparent review of OFs, identify evidence gaps, and provide inputs to subsequent BBN development. More broadly, the work illustrates a disciplined role for LLMs in safety analysis: extracting and organizing evidence within a constrained schema while leaving acceptance and quantification to qualified analysts.

References

[1] Peters, S., Morrow, S., Dennis, S., Lane, J., Ma, Z., & Siu, N. (2019). Organizational factors in PRA: Twisting knobs and beyond. In *Proceedings of the 2019 International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA 2019)* (pp. 860–869). American Nuclear Society.

[2] Pence, J., & Mohaghegh, Z. (2020). A discourse on the incorporation of organizational factors into probabilistic risk assessment: Key questions and categorical review. *Risk Analysis, 40*(6), 1183–1211.

[3] Akerstrom, M., Wahlström, J., Lindegård, A., Arvidsson, I., & Fagerlind Ståhl, A.-C. (2025). Organisational-level risk and health-promoting factors within the healthcare sector—A systematic search and review. *Frontiers in Medicine, 11*, Article 1509023.

[4] Mohaghegh, Z., & Mosleh, A. (2009). Incorporating organizational factors into probabilistic risk assessment of complex socio-technical systems: Principles and theoretical foundations. *Safety Science, 47*(8), 1139–1158.

[5] Ekanem, N. J., Mosleh, A., & Shen, S.-H. (2016). Phoenix—A model-based human reliability analysis methodology: Qualitative analysis procedure. *Reliability Engineering & System Safety, 145*, 301–315.

[6] Liu, M. X., Liu, F., Fiannaca, A. J., Koo, T., Dixon, L., Terry, M., & Cai, C. J. (2024). "We need structured output": Towards user-centered constraints on large language model output. In *Extended abstracts of the CHI Conference on Human Factors in Computing Systems* (Article 10, pp. 1–9). Association for Computing Machinery.

[7] Xiao, X., Chen, P., Jia, Q., Tong, J., Liang, J., & Wang, H. (2025). *A dynamic and high-precision method for scenario-based HRA synthetic data collection in multi-agent collaborative environments driven by LLMs* [Preprint]. arXiv.

[8] Cheng, T., Tabibzadeh, M., Jafary, H., Gill, H., Zaratsyan, D., & Mosleh, A. (2025). Identification and conceptual modeling of organizational factors affecting operational safety toward extending human reliability analysis methods. In *Proceedings of the 35th European Safety and Reliability Conference and the 33rd Society for Risk Analysis Europe Conference*. Stavanger, Norway.

[9] Tabibzadeh, M., Zaratsyan, D., Jafary, H., Cheng, T., Ramos, M., & Mosleh, A. (2024). Identification of organizational factors affecting the safety of operations: Foundation for extending human reliability analysis methods. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications.

[10] Cheng, T., Mosleh, A., Tabibzadeh, M., Jafary, H., Doberstein, D., & Gill, H. (2026). *Incorporating organizational factors into human reliability analysis, Part I: Knowledge- and*

linguistic data-driven causal and probabilistic model development [Manuscript submitted for publication].
[11] Cheng, T., Mosleh, A., Tabibzadeh, M., & Jafary, H. (2026). *Incorporating organizational factors into human reliability analysis, Part II: Integration of the organizational factor model into Phoenix HRA and case studies* [Manuscript submitted for publication].
[12] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT* [Preprint]. arXiv.
[13] Hildebrandt, M. (2026). Knowledge requirements, agentic architectures, and prompt evolution for LLM-supported human reliability analysis. In *Proceedings of the 18th International Conference on Probabilistic Safety Assessment and Management (PSAM 18)*. Pittsburgh, PA, United States.