Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria

doi: 10.3850/ESREL2026061419_esrel26-p26576-cd

Knowledge and LLM-Based Approach to Develop Data-Driven Causal Model for Extending Human Reliability Analysis

David Doberstein
Department of Computer Sciences, UCLA, USA. E-mail: ddoberstein@ucla.edu

Tingting Cheng
The B. John Garrick Institute for the Risk Sciences, UCLA, USA. E-mail: tingtingc@ucla.edu

Human performance critically determines safety in high risk domains where organizational failures underlie major accidents. While numerous organizational factors (OFs) influencing human reliability have been identified, their causal interrelations remain poorly understood. A key challenge is that Human Reliability Analysis (HRA) data exists primarily as unstructured text (accident reports, interviews, and regulatory documents) which conventional data mining techniques cannot effectively analyze for implicit, context dependent causal relationships.
This study employs a Large Language Model (LLM) based approach to extract causal relationships among OFs from 266 validated academic and technical documents. Following comparative evaluation of three LLMs (GPT-5, GPT-5-mini, Grok-4), GPT-5-mini was selected based on cost-effectiveness and identification performance. Through structured prompt engineering defining 19 organizational factors and 38 target causal relationships, the system extracted 1,538 causal relationship instances with supporting textual evidence and justifications. Results provide empirical frequency distributions quantifying the prevalence of specific OF interactions. This framework demonstrates a scalable, reproducible method for transforming unstructured linguistic data into analyzable causal knowledge, enhancing the data-driven foundation of HRA. Future work will implement causality-specific prompting and LLM-as-a-judge validation to support development of probabilistic Bayesian network models for quantitative risk assessment.

Keywords: Human Reliability Analysis, Organizational Factors, Large Language Models, Causal Relationships, Information Extraction, Prompt Engineering, Natural Language Processing, Bayesian Networks, High-Risk Industries, Safety Analysis.

1. Introduction

Human reliability analysis (HRA) is dependent on the extent of understanding how organizational factors (OFs) influence high risk systems. Bayesian networks models are a new promising approach to model these influences, but require data on how causal relationships between OFs impact human reliability which is poorly characterized and difficult to extract systematically from text.

This research develops a Large Language Model (LLM) based approach to extract examples of causal relationships between OFs from previously compiled academic literature. This corpus of 266 research papers on OFs, human reliability analysis, and high risk industries was used to extract and justify causalities from 38 target causal relationships constructed from 19 defined OFs. Prompt engineering was employed to guide the LLM to find, justify, and cite examples of causal relationships. Following comparative evaluation of three models (GPT-5, GPT-5-mini, and Grok-4), GPT-5-mini was selected based on extraction performance and cost effectiveness. The process identified 1,538 causal relationship instances with supporting textual evidence, providing frequency data to support future Bayesian network development for HRA.

2. Background

2.1. *Human Reliability Analysis*

Human Reliability Analysis (HRA) was defined as the "systematic study used to identify, quantify, and reduce the probability of a human error when the operator performs a specified task under certain conditions" Ciani et al. (2022). HRA is critical in high risk industries such as nuclear, chemical processing, aviation, and transportation, where failures can lead to catastrophic consequences including loss of life and environmental disasters. Major examples include Three Mile Island, Chernobyl, and Bhopal which demonstrated how human, organizational, and technological factors interact to cause system failures Meshkati (1991). Various methods have been proposed to incorporate organizational factors into HRA, including taxonomic frameworks and probabilistic models Mohaghegh and Mosleh (2009).

2.2. *Organizational Factors in HRA*

Organizational factors (OFs) are organizational structures, processes, and behaviors that influence individual actions at work and affect both the likelihood and severity of accidents Tabibzadeh et al. (2024). However, developing probabilistic models such as Bayesian networks for HRA requires understanding causal relationships among OFs defined as how one organizational condition influences another, ultimately affecting human reliability. Currently, existing representations of causal relationships between OFs rely heavily on expert judgment rather than systematic empirical evidence from literature, limiting transparency and reproducibility Rosness (2009); Vinnem et al. (2009); Mohaghegh and Mosleh (2009). To address this gap, this research adopts a data driven approach to extract 38 expert defined target causal relationships (List 2) among the 19 OFs (List 1) from a comprehensive corpus of related academic literature.

List 1: 19 Organizational Factors

- **Organizational goals and strategies (Org. goals & strat.):** Explicit objectives and plans, including prioritization, measurement, and communication mechanisms.
- **Policy design (Policy):** Formalized rules, procedures, and guidance to translate goals into consistent decision rules and behaviors.
- **Organizational structure design (Org. structure):** Arrangement of roles, reporting lines, and workgroup boundaries that determine task distribution.
- **Leadership (Leadership):** Individuals and practices that set direction, allocate attention and resources, and model behaviors.
- **Organizational culture / value system (Org. culture/value):** Shared values and assumptions shaping decisions and trade-offs (e.g., production vs. safety).
- **Managerial control and safety assurance (Mgmt. control & safety):** Management processes that plan, monitor, verify, and correct operations to align with safety and performance objectives.
- **Employee training and development (Employee training):** Structured activities to build and refresh employees' knowledge, skills, and abilities for reliable task performance.
- **Workplace design and management (Workplace design):** Physical and environmental characteristics of the work setting and practices to maintain and adapt it.
- **Decision framing and execution (Decision framing):** Processes and structures determining how choices are presented, evaluated, assigned, and acted upon.
- **Information exchange & coordination (Info exchange):** Pathways, protocols, tools, and norms enabling timely, accurate sharing of task-relevant information.
- **Group climate (Group climate):** Shared perceptions of workload, fairness, support, and the work environment.
- **Stress (Stress):** Tension or pressure due to perceived urgency, severity of outcomes, or responsibility associated with decisions.
- **Task load (Task load):** Demands imposed by tasks in terms of cognitive complexity, execution difficulty, volume, and information requirements.
- **Bias (Bias):** Systematic tendencies in information selection, interpretation, or judgment that deviate from objective appraisal.
- **Human–System Interaction (HSI):** Quality and characteristics of interfaces through which humans interact with technical systems.
- **Resources (Resources):** Availability and adequacy of tools, equipment, and supports necessary to complete tasks without delay.
- **Team effectiveness (Team eff.):** Degree to which a team coordinates, shares information, and harmonizes contributions to accomplish shared goals.
- **Procedures (Procedures):** Explicit, stepwise instructions intended to guide personnel through tasks accurately and accessibly.
- **Knowledge, abilities (Knowledge/abilities):** Workers' content knowledge, experiential background, applied skills, and physical readiness.

List 2: 38 Causal Relationships

- **Org. goals & strat. → Policy**
- **Org. goals & strat. → Org. structure**
- **Org. goals & strat. → Mgmt. control & safety**
- **Org. structure → Leadership**
- **Org. structure → Decision framing**
- **Org. structure → Info exchange**
- **Org. structure → Team eff.**
- **Policy → Leadership**
- **Policy → Employee training**
- **Policy → Decision framing**
- **Policy → Info exchange**
- **Policy → Mgmt. control & safety**
- **Policy → Procedures**
- **Workplace design → Group climate**
- **Workplace design → Info exchange**
- **Leadership → Employee training**
- **Leadership → Group climate**
- **Leadership → Mgmt. control & safety**
- **Org. culture/value → Employee training**
- **Org. culture/value → Leadership**
- **Decision framing → Info exchange**
- **Info exchange → Employee training**
- **Info exchange → Team eff.**
- **Info exchange → Mgmt. control & safety**
- **Decision framing → Leadership**
- **Decision framing → Group climate**
- **Mgmt. control & safety → Group climate**
- **Mgmt. control & safety → Knowledge/abilities**
- **Mgmt. control & safety → Team eff.**
- **Mgmt. control & safety → Task load**
- **Mgmt. control & safety → HSI**
- **Mgmt. control & safety → Resources**
- **Group climate → Bias**
- **Group climate → Stress**
- **Group climate → Team eff.**
- **Employee training → Knowledge/abilities**
- **Employee training → Team eff.**
- **Employee training → Group climate**

3. Methodology

3.1. *Data Corpus and Target Definitions*

The data source consists of 266 research papers on OFs, human reliability analysis, and high risk industries. This curated corpus of documents was previously compiled for related HRA research. 19 OFs (List 1) and 38 target causal relationships among these factors (List 2) were defined prior to this research based on established HRA frameworks and organizational theory. These definitions provided structure to extracting causal evidence from the literature.

3.1.1. *LLM Selection and Evaluation*

Three LLMs, OpenAI's GPT-5 (the most current OpenAI model), OpenAI's GPT-5-mini (a lighter version of GPT-5), and xAI's Grok-4 (the most current xAI model), were chosen based on convenience and evaluated on their ability to identify multiple causal relationships from corpus documents. Each model was tested with the same prompt containing the 38 target causal relationships, definitions of all 19 OFs, and simple examples of a causal relationship. A sample PDF containing multiple causal relationships was processed by each model. Outputs were evaluated based on the scope and completeness of evidence, the quality and reasoning of justifications, and the quantity and quality of extracted examples.

3.1.2. *Prompt Engineering*

A single all encompassing prompt was designed to extract all 38 causal relationships from each document in the corpus. The prompt was structured in seven parts: (1) task definition informing the model of its purpose to extract explicit textual evidence of causal relationships, (2) output specifications requiring a JSON array with objects containing "causality", "excerpt", and "justification" keys per textual evidence, (3) nine formatting rules to ensure consistency and eliminate common errors, (4) a definition of causal relationships along with a list of common words which may indicate a relevant relationship, (5) a sample output to demonstrate expected format, (6) a json array containing OF names and definitions as seen in List 1, and (7) a list of causal relationships as seen in List 2.

3.1.3. *Implementation and Data Processing*

A Python script was developed to process the document corpus, utilizing LangChain to interface with GPT-5-mini and output results to CSV format, writing incrementally to preserve data in case of interruption. PDFMiner extracted text from the research papers which LangChain's ChatPromptTemplates sequenced with the initial prompt to provide necessary context. Built in JSON and CSV libraries converted model output to a csv format which could be easily read, distributed, and stored.

4. Results

4.1. *Model Comparison*

The three candidate models showed distinct differences in their ability to extract causal relationships under the same conditions: same sample document, identical prompt, and same script. Grok-4 identified evidence for three target causal relationships, while GPT-5 and GPT-5-mini identified seven and eight relationships respectively. In addition to identifying more target causal relationships, both GPT-5 and GPT-5-mini consistently produced longer excerpts and more detailed justifications. Given the difficulty of extracting sufficient causal evidence from academic text, it was advantageous for the model to produce more output for evidence collection. Between the two GPT-5 variants, the qualitative output was comparable in terms of reasoning clarity and relevance. Given the comparable extraction quality and the lower computational cost of GPT-5-mini, it was selected for processing the full corpus.

4.2. *Extraction Results*

Across the full corpus of 266 papers, the process identified 1,538 instances of causal relationships, of which 1,482 corresponded to the 38 targeted causal relationships among 19 defined OFs. A broadly distributed pattern across the 38 defined relationships was observed: 24 relationships occurred between 20–50 times, representing the majority of evidence, while 7 were relatively rare (< 20 occurrences) and another 7 were highly frequent (> 50 occurrences). The most frequent relationships were Info exchange & coordination $\rightarrow$ Decision framing, execution (116 occurrences) and Employee training and development $\rightarrow$ Knowledge, abilities (106 occurrences), while some relationships, such as Decision framing and execution $\rightarrow$ Leadership, were rare (1 occurrence). These instances provide a sufficiently large evidence base to support frequency-based analysis and the future development of Bayesian networks.

5. Discussion and Future Work

Three main challenges were observed in the extraction process. First, the model occasionally identified hallucinated causal relationships composed from two of the 19 OFs in List 1 but not present in the targeted 38 causal relationships in List 2. Second, inconsistencies in spelling and abbreviation usage were observed such as converting “&” to “and” or “info” to “information”. Third, the model occasionally identified irrelevant or incoherent pieces of evidence or made unclear arguments in the justification. Despite these issues, the extracted dataset remained mainly consistent with target causal relationships and predefined OFs, providing a usable basis for frequency based analysis and Bayesian network development.

The extraction process could be further improved by simplifying the tasks assigned to the model by breaking the extraction into separate prompts for each target causal relationship. This may reduce hallucinations and improve the clarity of justifications. Exploring alternative LLMs that are newer or better suited for this task may enhance performance as different models vary in processing context and reasoning critically. Additionally, implementing a retrieval augmented generation (RAG) framework could improve efficiency and scalability. Such a workflow would include semantic text chunking, vector embedding, vector storage for similarity based retrieval, more prompt engineering, and LLM-as-a-judge validation, enabling researchers to query the corpus multiple times without reprocessing all documents. This approach aligns with recent advances in knowledge and LLM based causal extraction, providing a more structured and accessible dataset for frequency analysis, creating claims, and supporting the development of probabilistic models all for furthering human reliability assessment.

Table 1. Frequency of extracted targeted causal relationships across 266 papers from List 2.

Cause Factor	Effect Factor	Freq.
Org. goals & strat.	Policy	25
Org. goals & strat.	Org. structure	13
Org. goals & strat.	Mgmt. control & safety	34
Org. structure	Leadership	20
Org. structure	Decision framing	49
Org. structure	Info exchange	38
Org. structure	Team eff.	31
Policy	Leadership	11
Policy	Employee training	45
Policy	Decision framing	48
Policy	Info exchange	40
Policy	Mgmt. control & safety	85
Policy	Procedures	77
Workplace design	Group climate	32
Workplace design	Info exchange	30
Leadership	Employee training	37
Leadership	Group climate	86
Leadership	Mgmt. control & safety	69
Org. culture/value	Employee training	25
Org. culture/value	Leadership	23
Info exchange	Decision framing	116
Info exchange	Employee training	19
Info exchange	Team eff.	66
Info exchange	Mgmt. control & safety	47
Decision framing	Leadership	1
Decision framing	Group climate	5
Mgmt. control & safety	Group climate	44
Mgmt. control & safety	Knowledge/abilities	33
Mgmt. control & safety	Team eff.	24
Mgmt. control & safety	Task load	23
Mgmt. control & safety	HSI	4
Mgmt. control & safety	Resources	40
Group climate	Bias	8
Group climate	Stress	29
Group climate	Team eff.	29
Employee training	Knowledge/abilities	106
Employee training	Team eff.	29
Employee training	Group climate	24
Total instances		**1482**

6. Conclusion

This study presents a LLM based approach to extracting targeted causal relationships among defined OFs from a document corpus of 266 research papers, resulting in 1,538 identified instances. Identical evaluation of three LLMs showed that GPT-5-mini provided the most cost effective solution to achieve desired output quality. The dataset offers a structured evidence base for frequency analysis and the eventual development of Bayesian network models for human reliability assessment in high risk industries. Despite the output exhibiting challenges such as occasional hallucinations and small inconsistencies regarding spelling and abbreviations, the methodology produced structured and interpretable results. Future work could implement a retrieval augmented generation (RAG) framework to enhance efficiency, enabling repeated queries of the corpus without reprocessing all documents, and supporting development of claims and creation of probabilistic models for quantitative risk assessment. Overall, this approach provides a data driven foundation for advancing the integration of OFs into HRA.

References

Ciani, L., G. Guidi, and G. Patrizi (2022). Human reliability in railway engineering: Literature review and bibliometric analysis of the last two decades. *Safety Science 151*.

Meshkati, N. (1991). Human factors in large-scale technological systems' accidents: Three mile island, bhopal, chernobyl. *Industrial Crisis Quarterly 5*(2), 133–154.

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

Rosness, R. (2009). Risk influence analysis a methodology for identification and assessment of risk reduction strategies. *Reliability Engineering System Safety 223*(1), 87–97.

Tabibzadeh, M., D. Zaratsyan, H. Jafary, T. Cheng, M. Ramos, and A. Mosleh (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 68*(1), 802–804.

Vinnem, J. E., J. Seljelid, S. Haugen, S. Sklet, and T. Aven (2009). Generalized methodology for operational risk analysis of offshore installations. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 223*, 87–97.