Joint CSCE Construction Specialty & CRC Conference 2025
Conférence conjointe spécialisée en construction de la SCGC et CRC-2025

Montreal, Quebec
July 28-31, 2025 / *28-31 juillet 2025*

ENHANCING JOB HAZARD ANALYSIS KNOWLEDGE RETRIEVAL THROUGH KNOWLEDGE GRAPHS AND LARGE LANGUAGE MODELS

A.D. Abellanosa[1]*****, E. Pereira[2], L. Lefsrud[3] and Y. Mohamed[1]

[1] Construction and Engineering Management, University of Alberta, Canada
[2] Department of Civil Engineering, University of Calgary, Canada
[3] Lynch School of Engineering Safety and Risk Management, University of Alberta, Canada
*abellano@ualberta.ca

ABSTRACT: Job Hazard Analysis (JHA) is a crucial process for identifying and mitigating risks in construction workplaces. Traditional JHA methods rely heavily on manual expertise, making them time-consuming, knowledge-intensive, and prone to inconsistencies. This research proposes an AI-driven framework that integrates Large Language Models (LLMs) with Neo4j-based Knowledge Graphs (KGs) to enhance JHA workflows by automating hazard identification and mitigation planning. The framework extracts safety-related knowledge from Occupational Safety and Health Administration (OSHA) standards, structuring it into an intelligent KG for efficient hazard retrieval and analysis. By leveraging LLMs for entity recognition and relationship extraction, the system enables automated hazard identification, risk assessment, and regulatory compliance verification. A case study on OSHA lead exposure monitoring compliance illustrates how this approach structures safety regulations and generates actionable hazard insights. Future research will focus on improving the precision of LLM-driven hazard identification, optimizing scalability for large datasets, and conducting user validation studies to refine real-world applicability. The proposed solution bridges traditional knowledge management systems with LLM-driven automation, offering a scalable, cost-effective, and adaptable tool for improving workplace safety in the construction industry.

1. INTRODUCTION

Job Hazard Analysis (JHA) is an essential means in construction safety, with the purpose of identifying potential risks associated with specific tasks and formulating mitigation strategies. Traditional JHA methods often rely on unstructured textual reports, expert judgments, and manual knowledge retrieval (Lu et al. 2015, Pandithawatta et al. 2023, Hong and Cho 2024). These approaches, while foundational, introduce inefficiencies in hazard recognition and decision-making. The fragmented nature of safety-related knowledge across various sources complicates its effective use in proactive risk mitigation (Gao et al. 2022, Wu et al. 2023). Additionally, Chen et al. (2024) highlighted the dynamic nature of construction sites requires ongoing supervision by onsite engineers to ensure that operations comply with safety regulations.

Given the increasing complexity of construction projects, an efficient and structured approach to safety knowledge management is paramount. Knowledge Management (KM) has been an essential tool in construction safety management by facilitating the systematic organization, storage, and retrieval of safety knowledge. Effective KM practices integrate diverse safety information sources, ensuring that relevant safety protocols and best practices are readily available for hazard identification and mitigation (Yuan et al. 2019, Bachar et al. 2025). However, challenges persist, particularly in smaller organizations that lack robust

KM infrastructures, leading to inefficiencies in knowledge dissemination and application. Addressing these limitations through advanced technological interventions, such as knowledge graphs (KGs) and large language models (LLMs), can significantly enhance the accessibility and usability of safety knowledge, fostering a proactive safety culture (Junwu et al. 2024, Bachar et al. 2025).

To further improve construction safety management, the integration of KGs into safety frameworks presents a promising solution. KGs provide a structured representation of safety knowledge by encoding relationships between various hazards, control measures, and safety standards, facilitating a more systematic and context-aware hazard analysis (Xu et al. 2023, Chen et al. 2024, Tao et al. 2025). Moreover, the application of LLMs in automated KG construction can streamline the extraction and organization of safety-related information from unstructured sources, reducing manual efforts and increasing accuracy (Jalali et al. 2024, Thiruganasambandamoorthy et al. 2024). By leveraging LLM-driven KG frameworks, construction safety management can transition from a predominantly manual and experience-driven approach to a more data-driven, scalable, and efficient system, ultimately enhancing safety outcomes and regulatory compliance.

To address these challenges, this study proposes a novel LLM-driven KG construction framework for JHA. The approach encompasses three main components: (1) the systematic extraction of entities and relationships through LLMs to organize hazard-related knowledge; (2) the representation of JHA data in a graph-based format via Neo4j, enhancing relevant knowledge retrieval and reasoning; and (3) the evaluation of the KG's effectiveness as a traversing guide for LLMs, aiming to improve JHA responses and the suggestion of hazard controls in accordance with pertaining safety standards.

This paper focuses on three primary objectives: (1) developing an automated entity and relationship extraction pipeline using LLMs for unstructured documents related to construction safety; (2) constructing a knowledge repository based on a KG to systematically organize JHA-related knowledge; and (3) assessing the feasibility and potential impact of the KG on enhancing hazard identification, information retrieval, and structured decision-making within JHA workflows by demonstrating a simple case study.

The structure of this paper is as follows: Section 2 provides a review of literature relevant to JHA, construction safety knowledge management, KGs, and knowledge extraction techniques using LLMs. Section 3 outlines the proposed methodology, detailing the processes for knowledge extraction from raw documents and the construction of KGs. Section 4 offers a case study as an initial implementation to demonstrate the viability of the proposed approach. The results and discussions are presented in Section 5, with conclusions and directions for future research outlined in Section 6.

2. LITERATURE REVIEW

2.1 JHA and Its Challenges in the Construction Industry

JHA is a fundamental safety tool in the construction industry, crucial for identifying and mitigating potential hazards associated with job tasks by examining the interactions between workers, tasks, tools, and the environment. This process not only highlights potential hazards but also facilitates the implementation of preventative measures to mitigate these risks effectively, making it a vital component of safety management practices (Chi et al. 2014, Lu et al. 2015). Given the dynamic and often unpredictable nature of construction sites, JHA is essential for adapting safety measures to varying conditions, ensuring that strategies are appropriately tailored to each unique scenario (Ouyang and Luo 2022). Additionally, JHA contributes significantly to reducing the high incidence of accidents typical in the construction sector by focusing on proactive hazard identification and risk assessment (Jeon and Cai 2021, Cho et al. 2025). While traditional JHA methods are manual and time-consuming, the integration of automated processes can enhance the efficiency of conducting safety checks, allowing for quicker responses to emerging hazards (Lu et al. 2015, Jeon and Cai 2021). Overall, JHA is part of a comprehensive safety management strategy that includes pre-task safety meetings and safety checklists, aiming to predict and mitigate hazards effectively, albeit with some limitations in highly dynamic environments (Jeon and Cai 2021, Ouyang and Luo 2022).

Traditional JHA is heavily reliant on the skill and knowledge of the individuals conducting the analysis, with inconsistent participation from all relevant stakeholders potentially leading to significant gaps in hazard identification (Oginni et al. 2023). Moreover, the integration of advanced tools like construction safety ontologies is crucial, as they systematize safety management knowledge and support automated safety planning, enhancing the efficiency and accuracy of hazard management (Hong et al. 2024).

Employing JHA in the construction industry faces multiple challenges due to the unique and dynamic nature of construction sites. Each project possesses distinct locations, schedules, and working conditions, making it difficult to accurately predict hazards, thus necessitating frequent updates to hazard analyses (Singh et al. 2023). The manual nature of JHA processes, which are often complex and time-consuming, can delay the implementation of necessary safety measures, highlighting the need for automated procedures to expedite hazard identification and mitigation (Lu et al. 2015). Variability in managing hazards is further complicated by different scenarios across projects, where even similar hazards may require distinct mitigation strategies due to varying site conditions, emphasizing the necessity for daily hazard analysis instead of a one-time assessment at the project's commencement (Chi et al. 2014, Singh et al. 2023).

2.2 KGs in Construction Safety Management

KGs have emerged as a transformative tool in construction safety management, offering various applications that enhance safety protocols and decision-making. By delineating relationships between entity words in construction safety standards, KGs improve understanding and practical application, such as demonstrating essential safety protocols like firefighting equipment inspection in heating sheds (Chen et al. 2024). They also play a pivotal role in accident analysis and prevention, helping to identify patterns, trends, and correlations from structured data to provide insights into accident causation and prevention strategies (Xu et al. 2023, Hong et al. 2024). For compliance checking, KGs automate the process by encoding construction safety regulations and codes into rules, ensuring that building designs and construction processes adhere to relevant standards (Tao et al. 2025). Additionally, transforming unstructured accident reports into structured KGs provides a data-driven foundation for improving safety protocols and practices, making the management process more efficient and effective (Hong et al. 2024).

Despite their significant contributions, KGs in construction safety management encounter several limitations that impede their full potential. These graphs require high accuracy due to their use in emergency management, often necessitating input from domain experts and substantial manual labor to maintain precision (Tao et al. 2025). This reliance on manual processes can delay the implementation and updates of KGs, reducing their effectiveness in rapidly changing construction environments (Tao et al. 2025). Moreover, the scope of existing research on KGs in construction is primarily concentrated on specific areas like underground engineering and urban rail transit, which restricts their applicability to a wider range of construction scenarios (Li et al. 2024). The absence of a mature and clear technical pathway for developing KGs further limits their widespread adoption in the construction industry (Li et al. 2024). Additionally, the lack of deterministic standard terminologies within KGs can lead to inconsistencies and misinterpretations, complicating their use (Lan et al. 2024). The complexity of visualizing relationships and entities within a graph can also make it challenging for construction management personnel to quickly understand and apply the information, further complicating the effective application of KGs in safety management (Chen et al. 2024). These factors collectively highlight the pressing need for further innovation and refinement in the use of KGs within construction safety management to overcome these barriers and fully harness their capabilities.

2.3 LLMs for Automated KG Construction

LLMs such as GPT-4, which are trained on extensive text corpora, demonstrate substantial capabilities in automated knowledge extraction. These models gain an understanding of grammar, syntax, context, and semantics through self-supervised and semi-supervised learning techniques, embedding deep lexical, syntactic, and semantic knowledge (Jalali et al. 2024, Thiruganasambandamoorthy et al. 2024, Yang et al. 2025). In the domain of knowledge extraction, LLMs facilitate the population of ontologies with domain-specific knowledge by generating instances for classes, relationships, and properties. This process involves iterative refinement to balance and structure the ontology (Ciatto et al. 2025). Additionally, LLMs enhance information retrieval tools, improving data retrieval capabilities which support the feeding of relevant information into knowledge bases – a process particularly vital in disciplines such as digital humanities (Chartier et al. 2025). However, the application of LLMs faces several challenges and future directions that need addressing. Issues of trust, especially in critical domains such as medicine and law, require LLMs to be trained with high-quality data to ensure accuracy and reliability (Wehnert et al. 2024). Moreover, integrating LLMs with advanced data analysis methods like LangChain can enhance the automation and sophistication of analytical processes (Hou et al. 2024). Nevertheless, hardware limitations and the demand for extensive computational resources remain significant hurdles for their practical deployment (Zhang et al. 2024).

LLMs integrated with KGs significantly enhance automated knowledge extraction. This combination improves the accuracy of LLM outputs by providing structured context to mitigate hallucinations and inaccuracies (Lavrinovics et al. 2025). Additionally, LLMs enhance entity and relationship extraction, proving especially useful in domains like agriculture (Wang and Zhao 2024). The use of LLMs in constructing KGs through iterative active learning processes, such as extracting relationships from scientific texts, expands knowledge bases effectively (Youn et al. 2024). This synergy also bolsters question-answering systems, where KGs provide essential context for more precise responses (Zhu et al. 2025). Ongoing research aims to improve this integration, focusing on enhancing scalability, accuracy, and the interpretability of extracted knowledge (Babaiha et al. 2023, Mishra et al. 2025). Recent advancements, such as Graph retrieval-augmented generation (GraphRAG), further refine this integration by enabling LLMs to generate structured knowledge graphs that facilitate scalable question answering over large private text corpora (Edge et al. 2024). Unlike conventional retrieval-augmented generation (RAG) systems, which struggle with global sensemaking questions, GraphRAG introduces a two-stage process: first, an entity knowledge graph is generated from source documents, and second, community summaries are precomputed for closely related entities Edge et al. 2024).

3. PROPOSED METHODOLOGY

Our methodology leverages the computational power of LLMs to create a knowledge repository by constructing KGs stored in Neo4j, enabling automated and intelligent query responses for JHA applications in construction. This process leverages the capabilities of LLMs for automated knowledge extraction through its entity recognition capabilities and Neo4j for sophisticated knowledge-based structuring and visualization. In this initial study, *GPT-4o-mini* is used for the automated extraction of entities and relationships from unstructured texts. The same model is used for the generation of the full JHA report. The methodology for this knowledge retrieval is guided by the GraphRAG strategy (Edge et al. 2024).

3.1 Data Preprocessing and Knowledge Repository Building

The creation of a KG is critical in structuring safety-related knowledge into a graph-anchored vector store to enhance both storage and retrieval efficiency. The preprocessing involves several steps designed to facilitate the KG construction:

1. *Chunking Texts*. Safety documents were segmented into smaller chunks to enhance the accuracy of embeddings and improve retrieval performance of the LLM later in the querying step. Our initial setup included parameters such as a chunk size of 250 and an overlap of 24. The corpus, which can include an array of document lengths, undergoes recursive text splitter from LangChain library[1], forming a searchable index.
2. *Embedding Creation.* Conversion of each chunked texts into embedding vectors is executed using OpenAIEmbeddings library[2], allowing for semantic similarity-based searches that are critical for precise information retrieval.
3. *Graph Storage*. These segmented text data were stored as nodes and relationships in Neo4j, facilitating the organization of data in a graph format that supports efficient navigation and retrieval.

[1] https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/
[2] https://api.python.langchain.com/en/latest/embeddings/langchain_openai.embeddings.base.OpenAIEmbeddings.html

For each textual triplet, a string representation is crafted as {(s)-[r]->(t)}, formatted as source node (s), relationship (r), and target node (t).

4. *Vector Indexing*. To enhance retrieval precision, a vector index was created within Neo4j. This vector representation is utilized to search the corpus index, identifying specific text chunks that display the highest semantic similarities, as determined by cosine similarity. This index is crucial for enabling faster and more precise matching of user queries with the relevant hazard information stored in the system.

3.2 Entity and Relationship Extraction Using LLMs

After the initial data preprocessing and KG formation, the LLM is employed a second time to contextualize the specific construction work. This phase involves:

1. *Contextualization of Construction Work.* The LLM examines user inputs related to specific construction tasks or projects and uses its entity recognition capabilities to identify JHA-related entities specific to the provided work context.
2. *Dynamic Querying and KG Traversal.* With the contextualized entities, the system utilizes Cypher to query Neo4j, identifying relevant relationships and traversing the KG. This step is critical as it allows the system to generate detailed JHA outputs by connecting relevant data points within the KG based on the current project's specific needs.
3. *Generation of JHA Outputs.* As a result of the dynamic querying, the system produces structured outputs that include identified hazards, suggested mitigation strategies, and other relevant safety information tailored to the specific construction context. These outputs are derived from the traversal of the KG, making them both comprehensive and specific to the task at hand.

4. CASE STUDY: OSHA LEAD EXPOSURE MONITORING COMPLIANCE

This case study effectively showcases the proposed LLM-driven KG approach in structuring regulatory information specifically drawn from Occupational Safety and Health Administration (OSHA) 29 CFR 1926 - Safety and Health Regulations for Construction, which pertains to lead exposure monitoring compliance. The extracted dataset vividly details employer responsibilities for monitoring airborne lead concentrations, conducting thorough exposure assessments, and the implementation of necessary controls as per the safety standards.

It's important to note that the full document of OSHA 29 CFR 1926 spans 1122 pages, encompassing a wide range of safety protocols across various construction activities. However, for the focused purpose of this demonstration, we have narrowed our analysis to Subpart E: Personal Protective and Life Saving Equipment, specifically Section 1926.103 – Respiratory Protection. This section provides the specificity and granularity needed to effectively evaluate the LLM-generated JHA responses. Narrowing down to this specific section allows for a more detailed and manageable analysis, ensuring that the LLM's output can be meticulously evaluated against precise regulatory requirements. This targeted approach not only enhances the relevance of our case study but also emphasizes the capability of our methodology to adapt and focus on specific regulatory aspects, which is crucial for practical applications where compliance with specific standards is mandatory. Additionally, this granularity is essential in demonstrating the effectiveness of the LLM in generating JHA responses that are not only compliant with the regulations but also actionable and specific to the unique needs of construction site safety management.

The database described contains a set of 165 nodes and 284 relationships, which are used to structure a comprehensive knowledge graph for regulatory and safety management in a specific domain. The nodes represent a variety of entities including different types of activities, chemicals, clothing, compliance programs, concepts, conditions, documents, effects, emissions, equipment, and various organizational and operational elements like groups, locations, materials, and regulations. Moreover, the relationships between these nodes are equally varied, capturing interactions such as compliance achievements, feasibility considerations, exposure levels, and compliance requirements. These relationships include actions like ACHIEVES_COMPLIANCE, COMPLIES_WITH, EXPOSED_TO, and REGULATED_BY, among others, which link various entities in meaningful ways to reflect regulations, operational requirements, and safety measures. Property keys within the database further enrich the nodes and relationships, adding specific attributes and values that enhance the granularity and utility of the information stored within the KG. An example visualization from Neo4j as provided in Figure 1, illustrating how documents are interconnected with other entities within the knowledge graph.

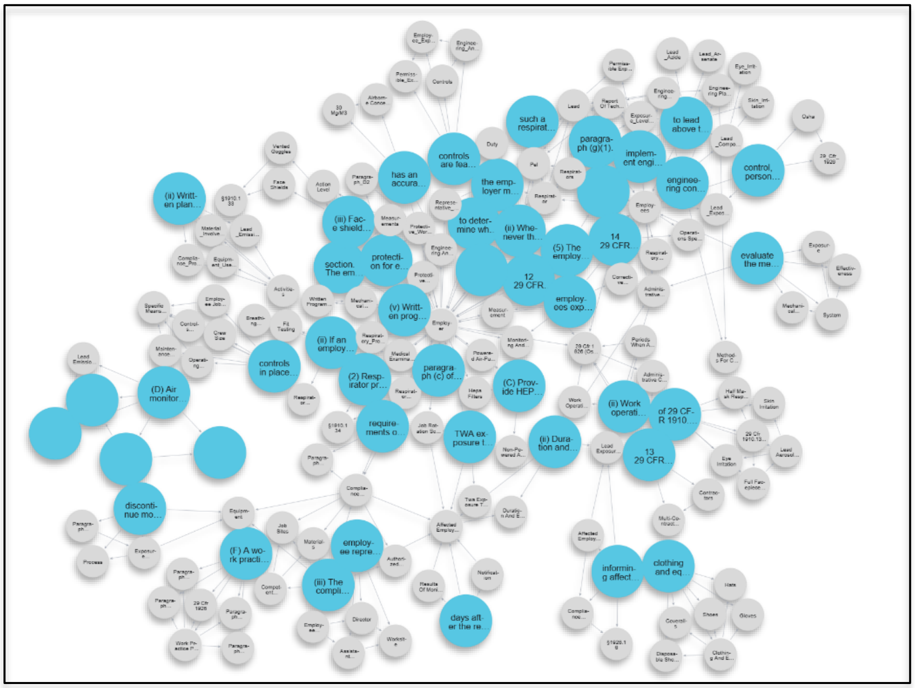


Figure 1: LLM-Generated Entity-Documentation Relationships (Entities-Grey, Document Text-Blue)

The structured KG as illustrated in Figure 2 is the visualization an LLM-extracted KG from Neo4j, illustrating the relationships between respirators, OSHA regulatory requirements (as stipulated in OSHA 29 CFR 1910.134), and workplace safety protocols. Nodes represent key entities such as respirator types, employers, compliance programs, and exposure assessments, while directed edges define their dependencies and interactions. The graph highlights how respirators protect against lead aerosols, their selection by employers, inspection requirements, and the regulatory framework governing their use. However, LLMs often identify identical entities as separate, as seen with the "Respirator" and "Respirators" nodes depicted in Figure 2.

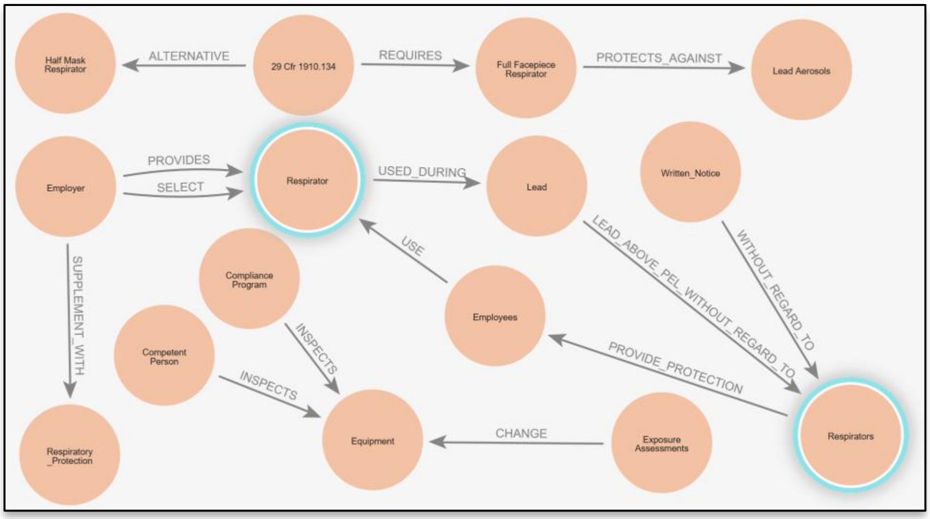


Figure 2: Sample of LLM-Generated KG for Respiratory Protection Compliance (Neo4j Output)

The LLM extracted several key entities from a construction project scenario involving welding operations in an underground tunnel as listed in Table 1. It identified "construction company" as the involved party, accurately pinpointing the "underground tunnel" as the primary location of operations, which is crucial for safety considerations. Hazards such as "hazardous welding fumes," "gases," "oxygen displacement," and "toxic gas accumulation" were comprehensively recognized, highlighting critical areas for safety measure development. Essential equipment like "welding equipment" and "respiratory protection" were also correctly noted, emphasizing the operational needs and safety requirements. Tasks including "installation of high-pressure pipelines" and "welding operations" were linked directly to the hazards and necessary safety equipment. Furthermore, the mitigation strategy of "wearing appropriate respiratory protection" was identified, addressing the risks associated with the hazardous working conditions. These extractions demonstrate the LLM's capability to effectively recognize and organize relevant information for conducting a thorough JHA.

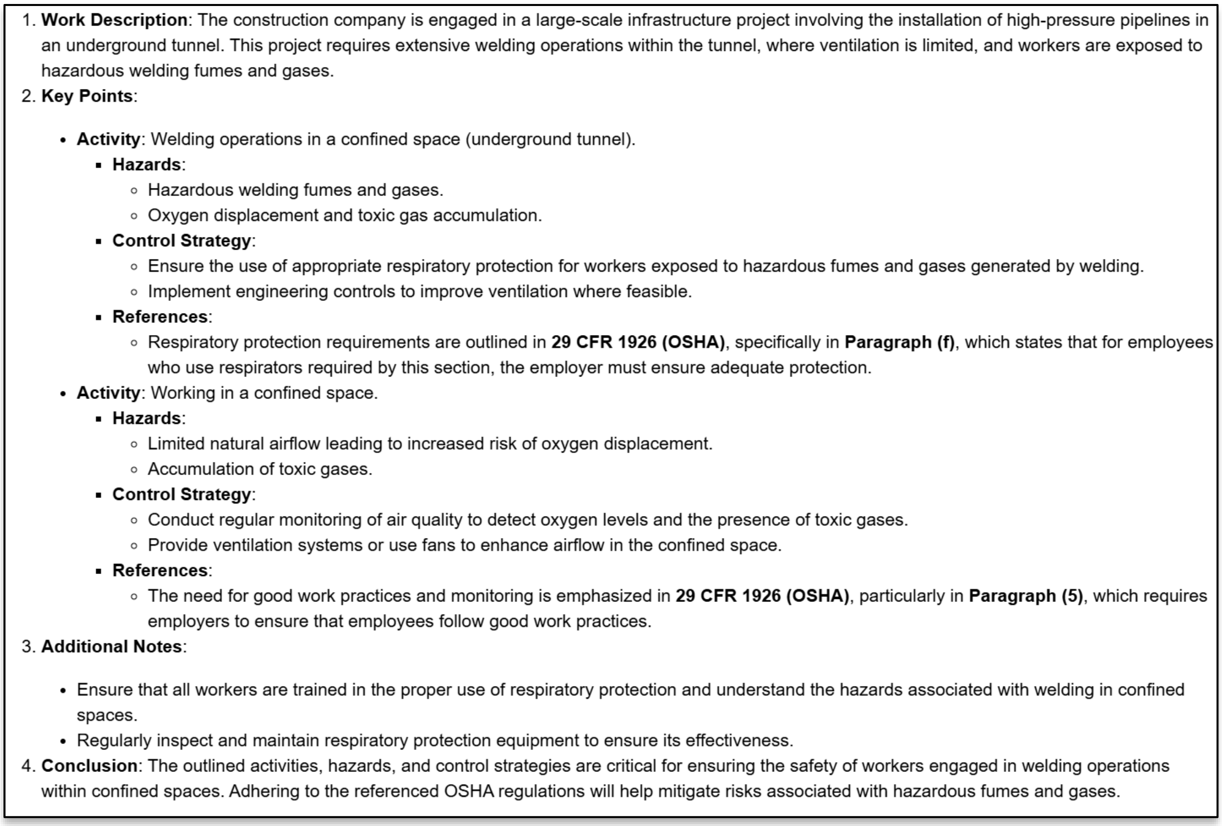
Table 1: Extracted entities from work description

Work description	*"A construction company is engaged in a large-scale infrastructure project involving the installation of high-pressure pipelines in an underground tunnel. The project requires extensive welding operations within the tunnel, where ventilation is limited, and workers are exposed to hazardous welding fumes and gases. The confined space environment further increases the risk of oxygen displacement and toxic gas accumulation. The welding process involves joining sections of steel pipes used for transporting high-pressure fluids. Due to site constraints, much of this work occurs inside the tunnel, where natural airflow is restricted. The workers must wear appropriate respiratory protection due to the presence of hazardous fumes and gases generated by welding."*
Entity	Extracted Entities
Name	"construction company"
Locations	"underground tunnel"
Hazards	"hazardous welding fumes", "gases", "oxygen displacement", "toxic gas accumulation"
Equipment	"welding equipment", "respiratory protection"
Tasks	"installation of high-pressure pipelines", "welding operations"
Mitigations	"wear appropriate respiratory protection"

The output of an LLM-generated JHA as presented in Figure 3, is derived from the traversed KGs through the matching entities in Neo4j. The analysis identifies key hazards associated with welding in confined spaces, such as hazardous fumes, oxygen displacement, and toxic gas accumulation. Control strategies, including respiratory protection, ventilation improvements, and air quality monitoring, are suggested based on regulatory guidelines, particularly OSHA 29 CFR 1926.

The LLM's output on the construction safety scenario was both accurate and comprehensive, identifying key entities for hazards associated with welding in confined spaces such as hazardous fumes, gases, and oxygen displacement. It correctly matched these hazards with appropriate control strategies, including respiratory protection, improved ventilation, and air quality monitoring, aligning with standard safety practices for confined space operations. The model also referenced the OSHA standards, specifically 29 CFR 1926, but it did not cite the correct sections or paragraphs, which would have further grounded the safety measures in legal requirements and enhanced the relevance and credibility of the provided information. Additionally, the presentation was clear and well-structured, with distinct sections for each activity, hazard, and control strategy, demonstrating its capability to deliver targeted safety management information suitable for practical applications.

It is also worth noting that the LLM did not mention any hazard control beyond respiratory protection, which is directly relevant to its existing knowledge base. This limitation highlights the advantage where the LLM did not hallucinate and suggest controls beyond its prior knowledge. It also suggests the potential for further enhancing the LLM's utility by expanding its KG database. Enriching the database with a broader range of hazard controls and safety measures could enable the LLM to generate more comprehensive and varied safety management strategies, thereby improving the depth and applicability of its JHA outputs in real-world scenarios.



1. **Work Description**: The construction company is engaged in a large-scale infrastructure project involving the installation of high-pressure pipelines in an underground tunnel. This project requires extensive welding operations within the tunnel, where ventilation is limited, and workers are exposed to hazardous welding fumes and gases.
2. **Key Points**:
   - **Activity**: Welding operations in a confined space (underground tunnel).
     - **Hazards**:
       - Hazardous welding fumes and gases.
       - Oxygen displacement and toxic gas accumulation.
     - **Control Strategy**:
       - Ensure the use of appropriate respiratory protection for workers exposed to hazardous fumes and gases generated by welding.
       - Implement engineering controls to improve ventilation where feasible.
     - **References**:
       - Respiratory protection requirements are outlined in **29 CFR 1926 (OSHA)**, specifically in **Paragraph (f)**, which states that for employees who use respirators required by this section, the employer must ensure adequate protection.
   - **Activity**: Working in a confined space.
     - **Hazards**:
       - Limited natural airflow leading to increased risk of oxygen displacement.
       - Accumulation of toxic gases.
     - **Control Strategy**:
       - Conduct regular monitoring of air quality to detect oxygen levels and the presence of toxic gases.
       - Provide ventilation systems or use fans to enhance airflow in the confined space.
     - **References**:
       - The need for good work practices and monitoring is emphasized in **29 CFR 1926 (OSHA)**, particularly in **Paragraph (5)**, which requires employers to ensure that employees follow good work practices.
3. **Additional Notes**:
   - Ensure that all workers are trained in the proper use of respiratory protection and understand the hazards associated with welding in confined spaces.
   - Regularly inspect and maintain respiratory protection equipment to ensure its effectiveness.
4. **Conclusion**: The outlined activities, hazards, and control strategies are critical for ensuring the safety of workers engaged in welding operations within confined spaces. Adhering to the referenced OSHA regulations will help mitigate risks associated with hazardous fumes and gases.

Figure 3: LLM-Generated JHA from KG Query

5. CONCLUSION AND FUTURE WORK

This present study introduced a novel framework for leveraging LLMs and Neo4j-based KGs to enhance JHA-related knowledge retrieval. The proposed approach demonstrated a structured hazard-related information extracted from unstructured safety documents, facilitating more automated retrieval of relevant safety knowledge to address construction job hazards.

The implementation of KGs in JHA reveals the potential to improve hazard recognition, assessment, and control by offering structured, queryable hazard insights. The ability to retrieve interconnected hazard knowledge efficiently can aid safety managers in making data-driven decisions, reducing workplace risks, and improving compliance with safety regulations. The structured output demonstrates how LLM-driven knowledge extraction and graph-based organization can enhance hazard identification, regulatory compliance, and decision-making in construction safety workflows.

While the initial implementation shows promise, yet several challenges require attention. Primarily, the accuracy of relationships extracted by the LLMs demands rigorous validation and postprocessing to ensure reliability in practical applications. This validation is crucial not only for verifying the relationships but also for assessing how accurately the model references legal and regulatory standards, such as specific OSHA section and paragraph citations. Misreferencing citations can undermine the trustworthiness and compliance of the safety recommendations. Additionally, scalability concerns must be addressed to manage large-scale safety datasets effectively. As the system extends to cover more diverse sources, the increase in data volume will complicate the validation process, potentially straining computational resources and affecting the timeliness of updates in the KG. Furthermore, there is significant scope for refining the prompting process used in LLM tasks. Experimenting with different prompting strategies could help in better aligning the extracted information with the specific requirements of construction safety management, thus improving the practical utility of the LLM outputs in real-world scenarios.

Future research will focus on improving model precision, optimizing scalability, and conducting user validation studies to refine the framework. Enhancing model precision will involve refining LLM-based entity recognition and relationship extraction. Scalability optimization will assess system performance with larger safety datasets, addressing potential bottlenecks in LLM responses and KG structuring. Moreover, user validation studies will engage industry professionals to assess system usability and effectiveness in JHA workflows. Feedback from safety managers will guide refinements, ensuring the system meets industry needs.

ACKNOWLEDGMENTS

This research was funded by the NSERC Alliance Research Grant ALLRP 556989-20 and supported by the Construction Innovation Centre funding for "Enhancing safety management systems practices on construction projects: A proactive data-driven approach for project safety", held by Dr. Lianne Lefsrud and Dr. Yasser Mohamed. We express our gratitude for their support, which was crucial in facilitating the development and execution of this project.

REFERENCES

Babaiha, N.S., Elsayed, H., Zhang, B., et al. 2023. A natural language processing system for the efficient updating of highly curated pathophysiology mechanism knowledge graphs. *Artificial Intelligence in the Life Sciences*, 4: 100078.

Bachar, R., Urlainis, A., Wang, K.C., et al. 2025. Optimal allocation of safety resources in small and medium construction enterprises. *Safety Science*, 181: 106680.

Chartier, M., Dakkoune, N., Bourgeois, G., et al. 2025. HiBenchLLM: Historical Inquiry Benchmarking for Large Language Models. *Data & Knowledge Engineering*, 156: 102383.

Chen, Y., Lu, G., Wang, K., et al. 2024. Knowledge graph for safety management standards of water conservancy construction engineering. *Automation in Construction*, 168: 105873.

Chi, N.W., Lin, K.Y., Hsieh, S.H. 2014. Using ontology-based text classification to assist Job Hazard Analysis. *Advanced Engineering Informatics*, 28(4): 381–394.

Cho, J., Shin, J., Jang, J., et al. 2025. Automated identification of hazardous zones on construction sites using a 2D digital information model. *Automation in Construction*, 170: 105922.

Ciatto, G., Agiollo, A., Magnini, M., et al. 2025. Large language models as oracles for instantiating ontologies with domain-specific knowledge. *Knowledge-Based Systems*, 310: 112940.

Dong, C., Wang, F., Li, H., et al. 2018. Knowledge dynamics-integrated map as a blueprint for system development: Applications to safety risk management in Wuhan metro project. *Automation in Construction*, 93: 112–122.

Feng, D., Chen, H. 2021. A small samples training framework for deep Learning-based automatic information extraction: Case study of construction accident news reports analysis. *Advanced Engineering Informatics*, 47: 101256.

Gao, S., Ren, G., Li, H. 2022. Knowledge Management in Construction Health and Safety Based on Ontology Modeling. *Applied Sciences*, 12(17): 8574.

Hong, E., Lee, S.Y., Kim, H., et al. 2024. Graph-based intelligent accident hazard ontology using natural language processing for tracking, prediction, and learning. *Automation in Construction*, 168: 105800.

Hong, Y., Cho, J. 2024. Enhancing Individual Worker Risk Awareness: A Location-Based Safety Check System for Real-Time Hazard Warnings in Work-Zones. *Buildings*, 14(1).

Hou, H., Shen, L., Jia, J., et al. 2024. An integrated framework for flood disaster information extraction and analysis leveraging social media data: A case study of the Shouguang flood in China. *Science of The Total Environment*, 949: 174948.

Jalali, M., Luo, Y., Caulfield, L., et al. 2024. Large language models in electronic laboratory notebooks: Transforming materials science research workflows. *Materials Today Communications*, 40: 109801.

Jeon, J.H., Cai, H. 2021. Classification of construction hazard-related perceptions using: Wearable electroencephalogram and virtual reality. *Automation in Construction*, 132: 103975.

Junwu, W., Yipeng, L., Jingtao, F. 2024. Integrating Bayesian networks and ontology to improve safety knowledge management in construction behavior: A conceptual framework. *Ain Shams Engineering Journal*, 15(9): 102906.

Lan, M., Gardoni, P., Weng, W., et al. 2024. Modeling the evolution of industrial accidents triggered by natural disasters using dynamic graphs: A case study of typhoon-induced domino accidents in storage tank areas. *Reliability Engineering & System Safety*, 241: 109656.

Lavrinovics, E., Biswas, R., Bjerva, J., et al. 2025. Knowledge Graphs, Large Language Models, and Hallucinations: An NLP Perspective. *Journal of Web Semantics*, 85: 100844.

Li, Q., Yang, Y., Yao, G., et al. 2024. Classification and application of deep learning in construction engineering and management – A systematic literature review and future innovations. *Case Studies in Construction Materials*, 21: e04051.

Lu, Y., Li, Q., Zhou, Z., et al. 2015. Ontology-based knowledge modeling for automated construction safety checking. *Safety Science*, 79: 11–18.

Mishra, C., Sarma, H., M., S. 2025. PageLLM: Incremental approach for updating a Security Knowledge Graph by using Page ranking and Large language model. *Information Processing & Management*, 62(3): 104045.

Nabawy, M., Ofori, G., Morcos, M., et al. 2021. Risk identification framework in construction of Egyptian mega housing projects. *Ain Shams Engineering Journal*, 12(2): 2047–2056.

Oginni, D., Camelia, F., Chatzimichailidou, M., et al. 2023. Applying System-Theoretic Process Analysis (STPA)-based methodology supported by Systems Engineering models to a UK rail project. *Safety Science*, 167: 106275.

Ouyang, Y., Luo, X. 2022. Differences between inexperienced and experienced safety supervisors in identifying construction hazards: Seeking insights for training the inexperienced. *Advanced Engineering Informatics*, 52: 101602.

Pandithawatta, S., Ahn, S., Rameezdeen, R., et al. 2023. Development of a Knowledge Graph for Automatic Job Hazard Analysis: The Schema. *Sensors*, 23(8): 3893.

Singh, S.P., Mansuri, L.E., Patel, D.A., et al. 2023. Harnessing BIM with risk assessment for generating automated safety schedule and developing application for safety training. *Safety Science*, 164: 106179.

Tao, Z., Liu, X., Li, Y., et al. 2025. Intelligent emergency assisted decision-making method based on standard digitalization: Hazardous chemical accidents in industrial parks. *Journal of Safety Science and Resilience*, 6(1): 79–92.

Thiruganasambandamoorthy, V., Probst, M.A., Poterucha, T.J., et al. 2024. Role of Artificial Intelligence in Improving Syncope Management. *Canadian Journal of Cardiology*, 40(10): 1852–1864.

Tserng, H.P., Yin, S.Y.L., Dzeng, R.J., et al. 2009. A study of ontology-based risk management framework of construction projects through project life cycle. *Automation in Construction*, 18(7): 994–1008.

Wang, H., Zhao, R. 2024. Knowledge graph of agricultural engineering technology based on large language model. *Displays*, 85: 102820.

Wehnert, S., Chedella, P., Asche, J., et al. 2024. A dynamic approach for visualizing and exploring concept hierarchies from textbooks. *Frontiers in Artificial Intelligence*, 7: 1285026.

Wu, W., Wen, C., Yuan, Q., et al. 2023. Construction and application of knowledge graph for construction accidents based on deep learning. *Engineering, Construction and Architectural Management*, ahead-of-print(ahead-of-print).

Xu, H., Wei, Y., Cai, Y., et al. 2023. Knowledge graph and CBR-based approach for automated analysis of bridge operational accidents: Case representation and retrieval. *PLOS ONE*, 18(11): e0294130.

Yang, W., Chen, Y., Xu, J., et al. 2025. Automatically learning linguistic structures for entity relation extraction. *Information Processing & Management*, 62(1): 103904.

Youn, J., Li, F., Simmons, G., et al. 2024. FoodAtlas: Automated knowledge extraction of food and chemicals from literature. *Computers in Biology and Medicine*, 181: 109072.

Yuan, J., Li, X., Xiahou, X., et al. 2019. Accident prevention through design (PtD): Integration of building information modeling and PtD knowledge base. *Automation in Construction*, 102: 86–104.

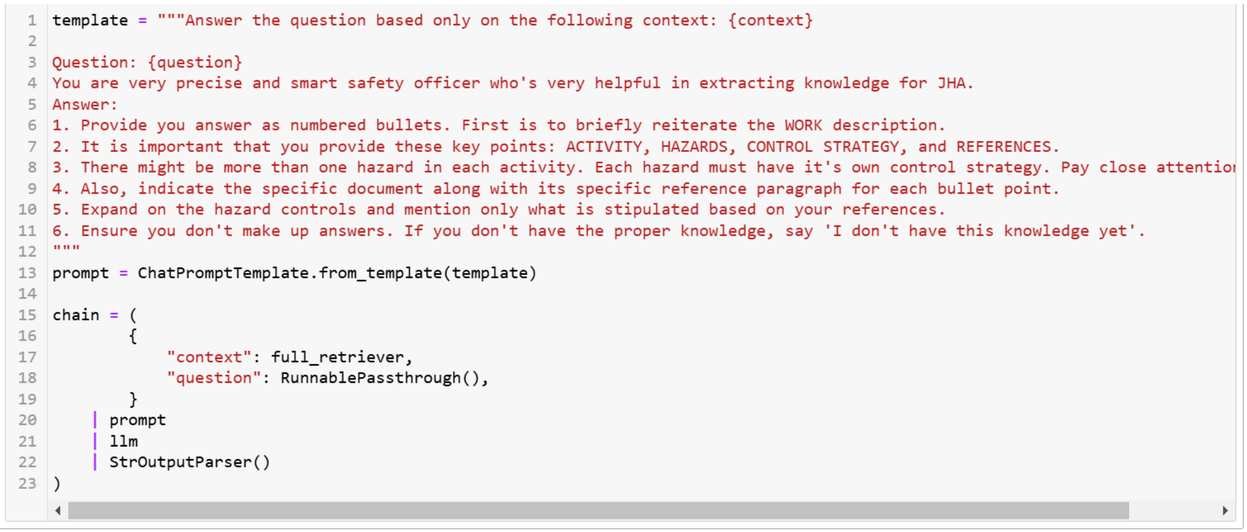
Zhang, Y., Yang, Z., Yang, Y., et al. 2024. Location-enhanced syntactic knowledge for biomedical relation extraction. *Journal of Biomedical Informatics*, 156: 104676.

Zhou, Z., Goh, Y.M., Li, Q. 2015. Overview and analysis of safety management studies in the construction industry. *Safety Science*, 72: 337–350.

Zhu, R., Liu, B., Tian, Q., et al. 2025. Knowledge graph based question-answering model with subgraph retrieval optimization. *Computers & Operations Research*, 177: 106995

APPENDIX

1. LLM prompting strategy for employed to retrieve context (matching entities and metadata from KGs queried from Neo4j).

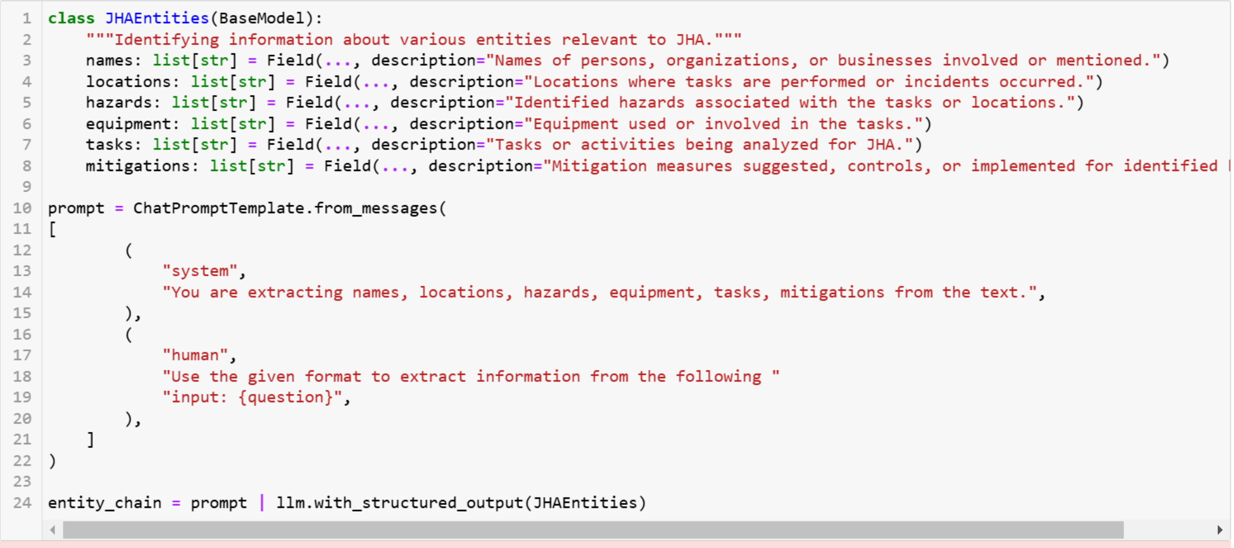


```
template = """Answer the question based only on the following context: {context}

Question: {question}
You are very precise and smart safety officer who's very helpful in extracting knowledge for JHA.
Answer:
1. Provide you answer as numbered bullets. First is to briefly reiterate the WORK description.
2. It is important that you provide these key points: ACTIVITY, HAZARDS, CONTROL STRATEGY, and REFERENCES.
3. There might be more than one hazard in each activity. Each hazard must have it's own control strategy. Pay close attentio
4. Also, indicate the specific document along with its specific reference paragraph for each bullet point.
5. Expand on the hazard controls and mention only what is stipulated based on your references.
6. Ensure you don't make up answers. If you don't have the proper knowledge, say 'I don't have this knowledge yet'.
"""
prompt = ChatPromptTemplate.from_template(template)

chain = (
        {
            "context": full_retriever,
            "question": RunnablePassthrough(),
        }
    | prompt
    | llm
    | StrOutputParser()
)
```

2. LLM prompting for entity retrieval from the user inputs of work description {question}. Each entity is provided with descriptions or definitions to contextualize such extracted entities.



```
class JHAEntities(BaseModel):
    """Identifying information about various entities relevant to JHA."""
    names: list[str] = Field(..., description="Names of persons, organizations, or businesses involved or mentioned.")
    locations: list[str] = Field(..., description="Locations where tasks are performed or incidents occurred.")
    hazards: list[str] = Field(..., description="Identified hazards associated with the tasks or locations.")
    equipment: list[str] = Field(..., description="Equipment used or involved in the tasks.")
    tasks: list[str] = Field(..., description="Tasks or activities being analyzed for JHA.")
    mitigations: list[str] = Field(..., description="Mitigation measures suggested, controls, or implemented for identified

prompt = ChatPromptTemplate.from_messages(
[
        (
            "system",
            "You are extracting names, locations, hazards, equipment, tasks, mitigations from the text.",
        ),
        (
            "human",
            "Use the given format to extract information from the following "
            "input: {question}",
        ),
    ]
)

entity_chain = prompt | llm.with_structured_output(JHAEntities)
```