EG-ICE 2025
GLASGOW

Enhancing Engineering Risk Analysis with Knowledge Graph-Driven Retrieval-Augmented Generation

LINGHAN OUYANG[1] HAIJIANG LI[1] JIUCAI LIU[1]

[1]BIM for Smart Engineering Centre, School of Engineering, Cardiff University, United Kingdom

ABSTRACT:
The increasing complexity and volume of unstructured risk-related data in engineering projects pose significant challenges for timely and accurate risk analysis. While Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) with external knowledge, traditional RAG systems struggle with context fragmentation and cross-chunk reasoning. This paper proposes a Knowledge Graph-enhanced RAG (KG-RAG) framework that integrates structured semantic relationships into the RAG pipeline to improve information retrieval and response generation. By extracting entities and their interrelations from textual risk assessment reports, the system builds a graph-based knowledge base that enables hierarchical summarization and precise risk identification. It supports both global risk summarization and causal chain tracing through a dual-mode retrieval strategy. A case study on the Jiaozhou Bay Second Subsea Tunnel project illustrates the efficacy of KG-RAG in analysing complex engineering risks, outperforming naïve RAG methods in accuracy, traceability, and decision support. The results suggest KG-RAG offers a scalable and intelligent solution for automating engineering risk assessment.

KEYWORDS:
Knowledge graph, Retrieval-Augmented Generation (RAG), Large language model(LLM), Infrastructure risk assessment.

1. INTRODUCTION

The architecture, engineering, and construction (AEC) industry is widely recognized as one of the most hazardous sectors, characterized by frequent engineering failures and risks to personnel safety. Large-scale construction projects often involve significant risk exposure, resulting in considerable losses in both human life and property. Although risk control measures are essential across all AEC projects, the industry continues to face technical challenges due to the complexity of implementation and the adverse social impacts of post-risk events.

Effective pre-risk analysis involves understanding causal chains between risk sources and events, tracing the contextual evidence behind those relationships, and presenting the findings in a structured and interpretable format. As the volume and complexity of project data increase, professionals face mounting difficulty in detecting emerging risks or responding to known hazards in a timely and traceable manner.

Current risk management practices in AEC involve the handling of extensive pre-risk textual reports as well as multimodal records—such as images and videos—collected during the construction process. Among the various stages of risk mitigation, pre-risk measures—particularly the detailed analysis of textual documentation—are generally more effective in reducing potential losses than in-process or post-risk interventions. As a result, the automated processing of textual risk data has emerged as a key area of research and development.

At present, textual data processing in the AEC domain primarily relies on two main categories of methods: rule-based approaches and neural network models. Rule-based methods derive insights from text using manually constructed grammatical rules and domain-specific knowledge bases. While these methods offer strong interpretability, they struggle to handle complex semantic scenarios due to their limited generative capacity. Neural network-based approaches, in contrast, encode words into vector representations and utilize deep learning architectures to model contextual semantics. These models can capture intricate textual patterns but remain constrained by architectural limitations and parameter size, particularly when processing long or complex texts.

Recently, large language models (LLMs) have emerged as state-of-the-art in natural language processing (NLP). Trained on massive corpora using self-supervised learning, LLMs consist of billions of parameters and demonstrate superior generative capabilities and long-text comprehension compared to previous NLP approaches. Accordingly, they hold significant promise for enhancing the automation of text analysis in risk management. Despite their strengths, general-purpose LLMs face notable limitations in domain-specific applications. These include a lack of contextualized domain knowledge(Wang et al., 2025), the inability to incorporate post-training updates, and the tendency to produce hallucinated or inaccurate outputs due to their probabilistic generation mechanisms. As a result, their effectiveness in specialized professional scenarios remains limited.

To overcome these shortcomings, researchers have explored strategies for adapting LLMs to domain-specific tasks, most notably retrieval-augmented generation (RAG) (Lewis et al., 2020) and fine-tuning. RAG improves factual accuracy and domain alignment by integrating an external, retrievable vector database containing specialized knowledge, thereby enabling evidence-based and traceable responses. However, RAG's effectiveness is highly dependent on the quality and coverage of the underlying knowledge base. It also suffers from performance degradation in handling out-of-domain queries, and the added retrieval step introduces latency, reducing generation speed. Fine-tuning, by contrast, updates a subset of model parameters using domain-specific training data. This approach enhances domain adaptation without compromising inference speed. Nevertheless, it incurs high training costs, requires large amounts of annotated data, and does not guarantee output reliability or traceability. Given the specific requirements of project risk analysis—particularly the need for verifiable and transparent outputs—fine-tuning is often impractical. In this context, RAG emerges as a more suitable solution due to its ability to incorporate domain knowledge while ensuring the traceability of generated content. RAG has been applied to many domains in AEC industry like job safety reports (Bernardi et al., 2024), gas risk assessment in coal (Sun et al., 2025)

While effective in many cases, this approach is constrained by chunk-level embedding granularity and the model's context length limitations. This limitation impairs the model's ability to handle complex many-to-many mappings—for example, associating a fractured fault zone with multiple downstream risks (water inrush, tunnel collapse), or identifying how a single event (structural instability) might be caused by several distinct factors (groundwater flow, joint development, overburden pressure).

To overcome this limitation, structured preprocessing of textual inputs is essential. Knowledge graphs (KGs) —constructed by extracting entities and relationships from unstructured text—offer a viable solution(Edge et al., 2024). By representing domain knowledge in a structured, interconnected form, KGs can enhance retrieval precision and contextual richness, thereby improving the performance of subsequent generative tasks within RAG frameworks(Gao et al., 2025, Zhang et al., 2025).

This study proposes a fast and scalable framework for the automated processing of engineering risk in complex construction environments. By integrating LLMs, RAG and KGs, the proposed framework aims to reduce decision-making delays and minimize potential losses associated with risk events.

The key contributions of this study are as follows:

We propose KG-RAG, a domain-specific extension of Retrieval-Augmented Generation that integrates structured knowledge graphs into the language model retrieval pipeline to support contextual reasoning, semantic completeness, and traceable outputs in engineering risk analysis.

We design a dual-pipeline retrieval strategy that aligns with real-world engineering needs, enabling both global risk summarization and topic-specific causal tracing through top-down and bottom-up retrieval modes.

We apply the method to a real-world infrastructure case—the Jiaozhou Bay Second Subsea Tunnel—and demonstrate that KG-RAG outperforms naïve RAG systems in relevance, completeness, coherence, and explainability for complex risk queries.

The rest of this paper is organized as follow: the methodology of hybrid RAG is in Section 2 and the practice of the hybrid RAG on practical engineering cases is introduced in Section 3 in detail. And the conclusions are in Section 4.

2. METHODOLOGY

In this section, the methodology for the automation processing of engineering risk will be presented. To specify the advantage of proposed method. Naïve RAG will be in advance introduced in this section for comparison. Figure 1 shows the outline of the methodology in this paper with comparison. The proposed KG-enhanced RAG and naïve RAG both enhance the performance of LLMs by grounding its response in external knowledge. The main difference lies in that the KG-RAG model used in this paper establishes semantic connections by constructing a graph structure.

2.1 Overview of Naïve RAG

RAG is a widely adopted approach that enhances large language models (LLMs) by supplementing their responses with external textual knowledge. In its standard form, RAG operates by dividing unstructured documents into smaller segments or “chunks,” which are then transformed into vector embeddings and stored in a searchable database. When a user submits a query, the system retrieves the most semantically similar chunks and uses them, along with the query, to prompt an LLM to generate a response. When paired with a rich and well-structured external knowledge base, RAG significantly enhances the factual reliability of LLM-generated responses.

Although this method is effective for general cases, its design is not well suited to the demands of engineering risk analysis. Risk-related information is rarely confined to a single, clearly defined passage; rather, it is often fragmented across sections, with key factors like geological conditions, construction constraints, and hazard triggers appearing in non-contiguous portions of the report. A single concept—such as a fractured fault zone—may be associated with multiple downstream risks like water ingress and tunnel collapse, yet these relationships are dispersed across various document contexts. Naïve RAG, which treats each chunk independently and lacks an understanding of structural or causal links, struggles to consolidate this kind of distributed information.

Moreover, because it retrieves and presents information without modelling how concepts are related, the outputs of naïve RAG can lack the traceability and contextual coherence that engineers require for informed decision-making. In high-stakes domains such as infrastructure planning and pre-risk mitigation, the inability to generate verifiable, well-grounded answers significantly limits the utility of naïve RAG systems.

These challenges underscore the need for a more structured and semantically informed framework—one that can reflect the complexity of real-world engineering risks and support more reliable, context-aware reasoning.

2.2 KG-RAG: A Structured Framework for Engineering Risk Retrieval

To address the limitations of naïve RAG in the context of engineering risk, we propose a domain-aware retrieval framework that integrates knowledge graphs (KGs) with large language models (LLMs). Unlike standard RAG systems, which rely on unstructured text chunks, our method explicitly models the relationships between risk sources, causal factors, and hazard outcomes—enabling more precise retrieval, semantic aggregation, and traceable generation.

Engineering risk reports contain distributed and interdependent references to critical concepts typically. For example, a fault zone may be described in geological sections as fractured rock, in hydrological sections as a groundwater conduit, and in construction planning as a risk collapse in separate parts of the document. These references form a causal chain, yet naïve RAG lacks the ability to consolidate such information. KG-RAG addresses this challenge by transforming raw text into a structured network, where entities are represented as nodes and their causal or associative relationships form the edges of a knowledge graph.

The process begins by dividing source documents into overlapping text segments and using LLMs to extract entities and their relationships in a zero-shot setting. These extracted elements are cleaned and normalized to reduce redundancy and ambiguity, allowing for the consistent construction of the knowledge graph. Once assembled, the graph is partitioned into semantically cohesive communities using the Leiden algorithm. These communities represent clusters of risk-related entities and factors—such as excavation challenges or groundwater-induced failures—that reflect recurring themes in the data and facilitate high-level understanding. with a structured and traceable prompt that combines entity descriptions, retrieved chunk content, and community-level summaries from the knowledge graph.

By embedding risk semantics directly into the retrieval process, KG-RAG moves beyond simple keyword or vector-based matching which is the key method for the Naïve RAG and enables causal reasoning, entity-centric aggregation, and reliable traceability—features that are essential for effective pre-risk planning and decision-making in complex engineering contexts.

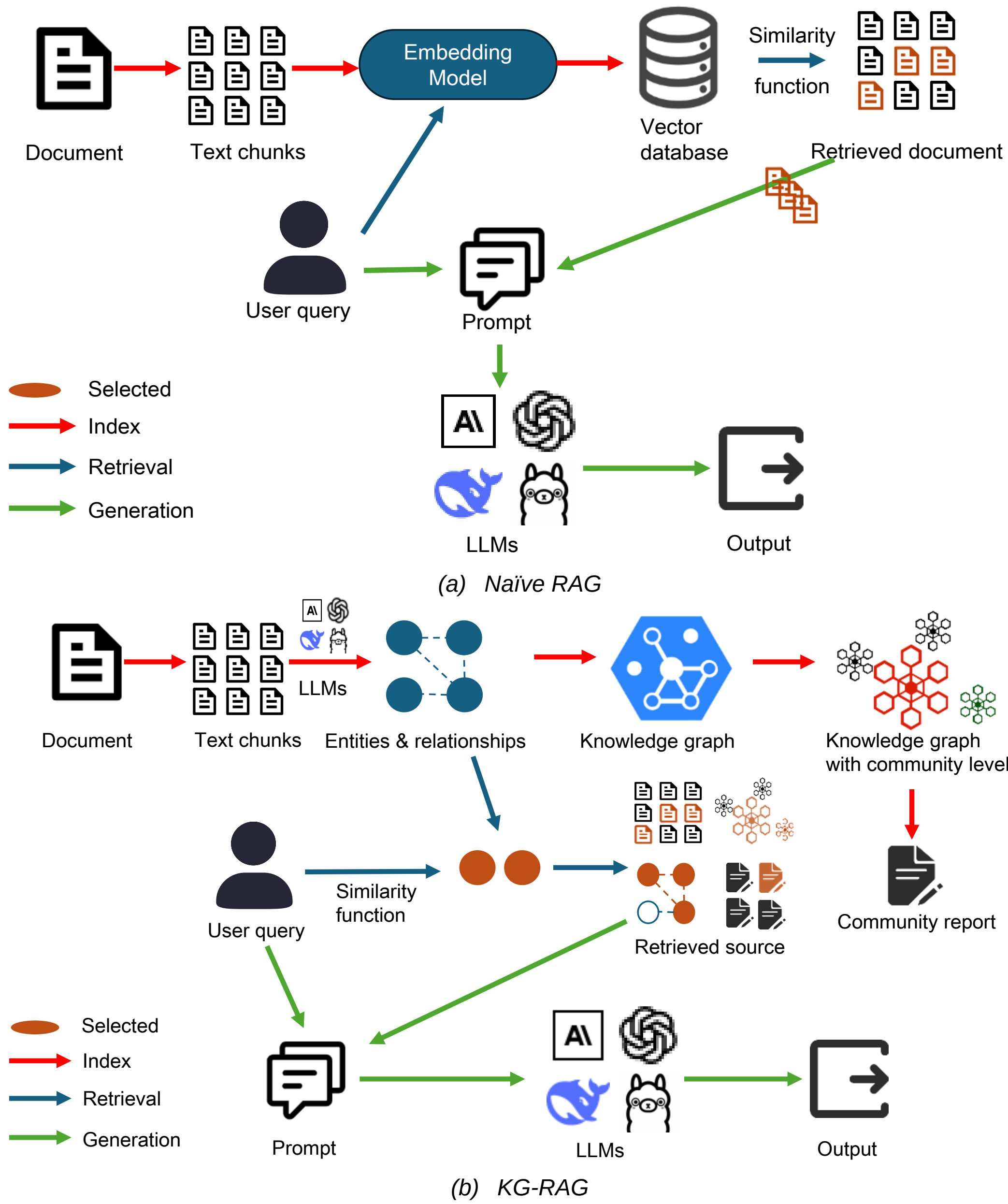

Figure 1: Flowchart of methodology of (a) Naïve RAG and (b) KG-RAG

2.3 LLM-Based Evaluation of Risk Assessment Outputs

While the underlying causal chains in engineering risk are grounded in objective physical phenomena, the assessment process itself remains context-dependent and subjective in expression. Unlike NLP tasks where benchmark datasets and reference outputs exist, Reports generated by different professionals may vary in structure and emphasis—even when addressing the same scenario. As such, there is no universally accepted ground truth against which system outputs can be compared in a definitive way.

In this context, conventional automatic evaluation metrics—such as BLEU or BERTScore, which rely on surface-level similarity—are insufficient for assessing the quality of risk-focused language generation. These metrics cannot capture critical attributes such as factual grounding, traceability, or the logical organization of risk reasoning.

To address this challenge, we adopt a large language model (LLM) as a neutral evaluator, which is increasingly used for assessing open-generation tasks. By designing structured prompts and scoring rubrics, the LLM is instructed to judge system outputs with certain dimensions which is shown in Table 1.

Table 1 Evaluation dimension for output of RAG system

Evaluation dimensions	content
Relevance	Does the answer directly and accurately address the query?
Completeness	Does the answer cover all major points related to the query?
Traceability	Does the answer cite or clearly reference supporting information?
Coherence	Is the answer logically organized and clearly written?
Accuracy	Are the technical details plausible and factually grounded?

The prompt provided to the LLM includes the original user query and three system-generated responses—one from the proposed framework and two from naïve RAG systems using different language models as comparisons. For each dimension, the LLM is instructed to assign a score on a 10-point scale and briefly justify its judgment. The output from LLM judge will be regarded as the reference for output's quality.

3. APPLICATION CASE STUDY

To validate the performance of the proposed KG-RAG method, A real project risk assessment report is set as example, building a graph engineering risk database for the future query.

3.1 Case introduction

The Jiaozhou Bay Second Subsea tunnel is in Qingdao, Shandong Province, China, which is the largest subsea road tunnel in construction in the world. The total length of the project is 17.48 kilometres, with the main tunnel extending 14.37 kilometres in total, including a 9.95-kilometer section beneath the sea. Based on the current cross-bay transport scheme, the tunnel will increase the road capacity, significantly reduce the time required for cross-bay transport during peak hours and facilitate the life of citizens and urban development.

The engineering geology and hydrogeological conditions along the Jiaozhou Bay Second Subsea Tunnel are extremely complex. The subsea section intersects several fault zones, which are composed of multiple minor fault zones. Additionally, several small-scale fractured zones are developed on both the Huangdao and Qingdao sides of the land. A combined construction method is adopted along with the tunnel alignment, using drill-and-blast on the western section and tunneling on the eastern section. Due to the overlying seawater and the complex geological environment, geological investigation is limited. Combined with constraints in technical capabilities and other factors, the construction phase of the subsea tunnel is subject to significant uncertainty and safety risks.

Figure 2: A comprehensive overview of the geographical location of The Jiaozhou Bay Second Subsea tunnel and cross-bay transportation in Qingdao

Considering the large scale of the tunnel project, the limited number of construction work fronts, the high difficulty of construction organization, and the elevated risks of construction and environmental impact in certain sections, there is a degree of uncertainty in the project timeline. Therefore, risk assessment is essential during the construction of the Jiaozhou Bay Second Subsea Tunnel to effectively control and mitigate potential risk incidents.

Based on relevant standard and practical experience, risk assessment of major projects is generally conducted at two levels: risk sources and risk events. For risk sources, investigation and recordation as well as classification and summary work need to be carried out. As for risk events, they are inferred based on general engineering experience and are classified into two types: those based on risk sources such as the rainfall-induced failure and those inherent such as potential delay caused by malfunction of machines. Professionals will conduct risk classification based on the possibility of risk events occurring and the potential losses of personnel and property after the occurrence.

The core of risk assessment is a risk source-risk event mapping into many-to-one and one-to-many scenarios. Hence, the structural graph format is a good format for engineering project risk assessments.

3.2 Graph Construction

The knowledge corpus comes from risk assessment report from Tongji University and Qingdao Guoxin Jiaozhou Bay Second Submarine Tunnel Co., Ltd. (Appointing party of this tunnel). Tongji University is the appointed party for the risk assessment task. The professional carefully studied the document and data provided by the appointing party (including various aspects such as hydrology and geology, design scheme, social research reports, feasibility analysis reports and other preliminary materials). They strictly conducted risk analysis in accordance with relevant industry standards and compiled this *Risk Assessment Report for Jiaozhou Bay Second Subsea Tunnel*. The report has passed the project review and is now being put into practice.

The risk assessment report of this project is voluminous and complex in content; hence, this study only selects the risk assessment report of the land area section on the Huangdao side as the knowledge corpus for establishing the knowledge graph. The selected two chapters of the original report are divided into 19 documents and the figures and tables are transferred to the text with GPT-4o model from OpenAI. The documents have 20631 words in total.

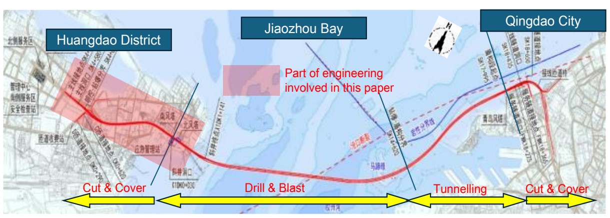

Figure 3: *The geographical location of the land-based part of the project and its relative position within the overall project.*

The documents are cut into 750-word-long chunks with 100 words overlapping to avoid some continuous semantic meaning is disturbed. 51 chunks are created, and some chunks are randomly selected for manual entity extraction. The input and output of manual entity extraction were used as part of the prompt for LLMs (gpt-4o-mini in this paper if no special mention) in the subsequent automated process, building a paradigm to guide LLMs in the entity extraction operation.

With the entities and relationships extracted from the chunks, the graph with 1535 nodes and 4273 relationships is built which is shown in figure. The automatic graph construction using LLms cost about 800 thousand tokens in total. Leiden algorithm is introduced in building community from the graph. The result shows that this graph contains 17 level-0 communities, 78 level-1 communities and 23 level-2 communities (level-0 community is the most general community and level-2 community is the most specific community). A community should have the strong inner connection among the nodes and is relatively weak connected to other community, hence, a community could be regards as a group of entities with the shared topic. Based each community, the LLMs helps generation a community report which summarize the main point of the community. By means of graphs and communities, the generation of cross-chunk, shared-topic and retrievable text has been achieved. In the graph construction phase, all the data including chunks, entities, relationships, communities and reports are created and ready for the future retrieval.

Figure 4: The knowledge graph corpus and community of risk assessment of project

3.3 Response to the query

To demonstrate the adaptability of the KG-RAG framework, queries were categorized into two types based on common risk analysis practices: global summarization and topic-specific queries. Each type was handled using a tailored retrieval and generation pipeline.

For global summarization tasks, such as identifying the most critical risk sources in a project, the challenge lies in aggregating thematically related information that is distributed across multiple sections of a large report. Naïve RAG often fails to consolidate this dispersed content due to its chunk-based retrieval strategy. In contrast, KG-RAG leverages pre-identified semantic communities within the knowledge graph to support high-level thematic summarization. Community reports, generated during the graph construction phase, serve as compact representations of these themes. By ranking these reports based on their semantic similarity to the query, KG-RAG selects the most relevant subsets for response generation. This top-down pipeline enables a high-level overview with contextual coherence and grounded references.

For topic-specific queries—such as tracing risks related to a specific fault zone—the system adopts a bottom-up pipeline. Here, the query is matched directly to relevant entities in the knowledge graph. Once key entities are identified, their linked nodes, textual references, and associated communities are retrieved. This allows for a focused extraction of evidence and semantic reasoning based on causal and associative relationships. The LLM then synthesizes this information into a targeted response, enriched with cross-referenced evidence and contextual understanding.

By separating query handling into these two distinct modes, KG-RAG ensures both broad thematic coverage and fine-grained precision in risk information retrieval. This approach aligns with real-world engineering workflows, where both overview-level risk mapping and focused diagnostic inquiries are routinely required.

3.4 Comparison with the naïve RAG

To present the differences between proposed KG-RAG and other models more intuitively, Naive RAG was also constructed using Anything LLMs from Mintplex Labs Inc. for comparison. The same chunk size and embedding model were adopted, and gpt-4o-mini and gpt-4o models were used as LLMs in Naïve RAG.

The example questions selected in this section are designed based on the classification of "global summarization and topic-specific queries", and comparisons of the outputs under different methods are conducted respectively. The example problem for global summarization is "What are the top three risk sources that require the most attention in this project?". This problem needs to extract and summarize risks from various aspects such as those in the site, design, and construction process, and evaluate their importance based on factors such as severity and occurrence frequency. Undoubtedly, it belongs to the global summarization problem. The example problem corresponding to topic-specific queries is "What risks might be triggered by the fault zone and what are the other inducing factors for each of these risks?". This problem clearly indicates the risk source "Fault zone" and links to a one-to-many risk mapping starting from "Fault zone" for querying. This task is clearly within the scope of topic-specific queries.

The outputs were evaluated using the GPT-o3 prompt framework introduced in Table 2. Considering the potential for hallucinations when large language models (LLMs) process long textual inputs, the retrieved content used in the generation is not included in the evaluation prompt. Instead, this component will be reviewed manually to ensure factual consistency. GPT-o3 as the most advanced LLM from OpenAI is introduced here as the judge of the output.

Table 2 Prompt for GPT-o3 to judge the quality of output from three RAG system

Component	Content
Task Description	You are an expert evaluator in engineering risk communication. Your task is to assess the quality of three AI-generated answers to a technical risk-related query. For each score, please provide a brief justification.
Evaluation Criteria	1. Relevance – Does the answer directly and accurately address the query? 2. Completeness – Does the answer include all major risk factors or explanations relevant to the query? 3. Traceability – Are claims supported by clear references to the original data or document sources? 4. Coherence – Is the response logically organized and easy to follow? 5. Accuracy – Are the technical details plausible and factually grounded in real engineering knowledge?
Scoring Instructions	For each criterion, assign a score from 1 to 10 (10 = highest). Provide a one-sentence explanation for each score.
Input Structure	The prompt includes: • The original query • Answer A (KG-RAG output) • Answer B (Naïve RAG with GPT-4o) • Answer C (Naïve RAG with GPT-4o-mini)
Output Format	Provide results as: Relevance: A – X/10, B – Y/10, C – Z/10 → [Explanation] Completeness: ... ... Overall Judgment: Indicate which answer is best overall and explain why.

In the query "What are the top three risk sources that require the most attention in this project?" Naïve RAG with gpt-4o only generates some general answers and with gpt-4o-mini regards some assessment index in the construction schedule risk analyses as the assessment result for the whole project mistakenly. But the proposed KG-RAG generation well-organized answer with summary, detailed risk information and the knowledge source, getting the highest score in all dimensions from the GPT-o3 judge. For this question, 16 most relevant community reports are used for the initial generation of the answers, covering a wide range of the potential risk source from aquifer to intrusive formations of veins and rocks. Each community report will involve more than 10 entities and relationships in general which are widely distributed in every part of the original documents. The structure of output includes sub-inducing risk source, making the answer more logistic and convincing. Table 3 shows the structure and retrieved data of the output from proposed KG-RAG. But for the Naïve RAG, the retrieved chunks are quite limited (8 for gpt-4o-mini and 4 for gpt-4o) and most of them are context of each other. Because the similarity score between descriptive and declarative chunk and general query are too low, Naïve RAG performed badly in retrieving the useful chunk, causing the information limited in the output. All output only provides some number of retrieve sources, therefore the LLMs could not give out a reasonable score for traceability without the retrieved data included. Through manual review, the data involved in the proposed method has a much broader distribution in terms of semantics and the original documents.

Table 3: A summary of the KG-RAG's output for the global summarization query

No	Risk source	Sub-risk source	Retrieved data
1	Groundwater Presence	high groundwater levels	Reports (47)
		potential for water inrush and localized flooding	Reports (115, 98, 117)
		groundwater dynamics	Reports (32, 34, 64, 101, +more)
2	Geological Conditions	Geological uncertainty	Reports (63, 105, 78)
		Fractured zones	Reports (88, 113, 44)
		Intrusive rock dikes	Reports (115,98)
3	Hazardous Chemical Processes	the Petrochemical Zone	Reports (19, 48, 49, 53, +more)
		Liquefied Petroleum Gas and Hydrogen	Reports (21)

Table 4: a comparison with the output of proposed KG-RAG and naïve RAG

Query	"What are the top three risk sources that require the most attention in this project?"
KG-RAG (gpt-4o-mini)	…… 1.Groundwater Presence Groundwater presence is a significant risk factor that can lead to instability during excavation activities. It has been noted that areas with high groundwater levels, particularly in sections such as NK5+495 - NK5+759, contribute to a collapse probability level of 3, indicating a high risk that must be managed effectively [Data: Reports (47)]. The potential for water inrush and localized flooding necessitates the implementation of effective drainage and monitoring systems to mitigate these risks [Data: Reports (115, 98, 117)]. Furthermore, groundwater dynamics can complicate construction efforts, leading to mud and water inflow, which increases the likelihood of structural failure [Data: Reports (32, 34, 64, 101, 81, +more)]. ……(2. Geological Conditions, 3. Hazardous Chemical Processes and conclusion are omitted here due to the limitation of space.)
Naïve RAG (gpt-4o-mini)	The top three risk sources that need the most attention in this project are: 1. Site geological conditions risk. 2. Impact risks from adjacent structures. 3. Construction organization and technical plan risks.
Naïve RAG (gpt-4o)	The top three risk sources that need the most attention in this project are: 1. Construction Organization (Risk Probability Level: 5, Risk Loss Level: C, Overall Risk Level: Level IV) 2. Traffic Conditions (Risk Probability Level: 4, Risk Loss Level: D, Overall Risk Level: Level IV) 3. Meteorological Conditions (Risk Probability Level: 4, Risk Loss Level: D, Overall Risk Level: Level IV)

Table 5: a comparison for global summarization query by LLMs judge

Evaluation dimensions	KG-RAG (gpt-4o-mini)	Naïve RAG (gpt-4o-mini)	Naïve RAG (gpt-4o)
Relevance	9	7	6
Completeness	8	4	4
Traceability	5	3	3
Coherence	9	6	5
Accuracy	7	6	5

For all the three outputs, scores were given by GPT-o3 judge in Table 5, demonstrating that output from proposed method get a higher score in all dimensions. Specifically, significant improvements were observed in completeness, relevance, and coherence.

In the query "What risks might be triggered by faults zone and what are the other inducing factors for each of these risks?", As shown in Table 6, the KG-RAG output presents a more structured explanation of causal risk pathways, supplemented with knowledge captured in the graph. Even if the chunks retrieved in Naïve RAG is limited comparing with the proposed KG-RAG, the most important source of this question is retrieved across all three methods. This ensures that the output of Naive RAG remains grounded and comprehensive and is also in line with the prior understanding of topic-specific queries. Chunk(34) plays an important role in answering this query and it is a conclusion of different risks that may be triggered by various kinds of unstable rock masses. Therefore, it contains the relevant content of faults zone and has a very high degree of similarity. Proposed KG-RAG retrieved this chunk as well but through entities from the chunk instead of the similarity. The entities descriptions also contribute to the generation with some factors which are not mentioned in Chunk(34) like fractured rock mass and groundwater flow. Descriptions from similar entities provide the necessary data for the generation. Due to improved retrieval, GPT-judge feedback shows the Naïve RAG's output is highly qualified in topic-specific query. The difference in coherence is likely from the preference of LLM judge for extra reason statements.

Table 6 A summary of the KG-RAG's output for the topic-specific query

	Risks induced by faults zone	Inducing factors of risk	source
1	Structural Instability	Fractured Rock Mass	Entities (369, 536, 380)
		Joint Development	Chunk (34)
		Overburden Thickness	Chunk (34)
2	Water Inrush	Groundwater Accumulation	Entities (488, 631)
		Rock Veins	Chunk (34)
		Excessive Span	Chunk (34)
3	Tunnel Collapse	Highly Developed Joint Structures	Chunk (34)
		Fractured Fault Zones	Entities (536, 271)
		Groundwater Flow	Entities (488, 631)

Table 7: a comparison for topic-specific query by LLMs judge

Evaluation dimensions	KG-RAG (gpt-4o-mini)	Naïve RAG (gpt-4o-mini)	Naïve RAG (gpt-4o)
Relevance	8	6	7
Completeness	5	4	5
Traceability	3	2	2
Coherence	9	6	7
Accuracy	7	6	6

It is also worth noting that the current KG-RAG framework is built entirely upon unstructured textual data. As engineering risk assessment increasingly incorporates structured and multimodal sources—such as time-series sensor data, monitoring imagery, or geospatial models—the ability to integrate these data types will become essential for achieving more precise and comprehensive evaluations. At present, KG-RAG exhibits limited extensibility in this regard, highlighting an important direction for future development.

4.CONCLUSIONS

This study presents a knowledge graph-enhanced Retrieval-Augmented Generation (KG-RAG) framework for automated engineering risk analysis. By integrating structured knowledge representations into the RAG pipeline, the approach addresses key limitations of naïve RAG models—specifically, inadequate cross-chunk reasoning and limited contextual comprehension in large-scale unstructured data environments. In contrast to conventional NLP-driven summarization methods, KG-RAG is designed as a domain-aware decision-support tool that enhances the identification, retrieval, and interpretation of critical risk factors embedded in unstructured engineering documents.

By extracting structured entities (e.g., risk sources, hazard types, triggering conditions) and their causal relationships from any risk-related document, the system constructs a semantic knowledge graph that mirrors real-world risk registers and hazard propagation chains. This structure enables cross-chunk reasoning, semantic aggregation, and precise traceability — capabilities that are essential for risk mapping, back analysis, and scenario-based evaluation.

The proposed method supports two distinct analytical pipelines: a top-down approach for global risk summarization based on community-detected themes, and a bottom-up approach for topic-specific risk queries centered on individual entities and their interconnections. These dual pathways align closely with the way risk engineers conduct comprehensive or focused assessments.

Through a case study on the Jiaozhou Bay Second Subsea Tunnel, KG-RAG demonstrated superior performance over naïve RAG systems in terms of relevance, completeness, coherence, and explainability. The results underscore the framework's potential to automate and enhance high-stakes engineering risk evaluation by providing transparent, evidence-grounded outputs.

In summary, KG-RAG bridges the gap between advanced language models and the real-world demands of engineering risk management, offering a scalable, interpretable, and technically sound platform for practical deployment in complex infrastructure projects.

ACKNOWLEDGEMENTS

This research was funded by the China Scholarship Council (No. 202406260052) and We would like to express our gratitude to the personnel involved in from Jiaozhou Bay Second Subsea Tunnel project and Tongji University for their contributions to the risk report.

REFERENCES

BERNARDI, M. L., CIMITILE, M. & PECORI, R. Automatic Job Safety Report Generation using RAG-based LLMs. *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024. 1-8.

EDGE, D., TRINH, H., CHENG, N., BRADLEY, J., CHAO, A., MODY, A., TRUITT, S. & LARSON, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint* arXiv:2404.16130.

GAO, Z., CAO, Y., WANG, H., KE, A., FENG, Y., XIE, X. & ZHOU, S. K. 2025. FRAG: A Flexible Modular Framework for Retrieval-Augmented Generation based on Knowledge Graphs. *arXiv preprint* arXiv:2501.09957.

LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., KüTTLER, H., LEWIS, M., YIH, W.-T. & ROCKTäSCHEL, T. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

SUN, Y., HAN, Y. & LIU, X. 2025. Intelligent Gas Risk Assessment and Report Generation for Coal Mines: An Innovative Framework Based on GLM Fine-Tuning. *Electronics*, 14, 379.

WANG, Y., LUO, H. & FANG, W. 2025. An integrated approach for automatic safety inspection in construction: Domain knowledge with multimodal large language model. *Advanced Engineering Informatics*, 65.

ZHANG, J., RUAN, X., SI, H. & WANG, X. 2025. Dynamic hazard analysis on construction sites using knowledge graphs integrated with real-time information. *Automation in Construction*, 170.