RAG-Enhanced Safety Information Retrieval for Construction: Integration of Large Language Models with Domain-Specific Information

Xianxiang Zhao[1], Advik Mehta[2], Falak Sethi[2], Brian Gue[3], Qipei Mei[4], Lingzi Wu[1]

[1]Department of Construction Management, University of Washington, United States
[2]Department of Computer Science, University of Alberta, Canada
[3]PCL Industrial Management Inc, Canada
[4]Department of Civil and Environmental Engineering, University of Alberta, Canada

seanzhao@uw.edu, advik@ualberta.ca, fsethi@ualberta.ca, bmgue@pcl.com, qipei.mei@ualberta.ca, lingwu@uw.edu

Abstract –
In the construction industry, critical safety information is often scattered across numerous documents, standards, and regulations, making it challenging for practitioners to access and comprehend safety knowledge in their daily operations efficiently. To address this challenge, we propose an intelligent and reliable question-answering system for information retrieval and response generation on the construction health, safety, and environment documents via retrieval-augmented generation. Specifically, our system combines a fine-tuned LLaMA-3-8B base model with a vector database constructed using embedding models, enabling accurate information retrieval and enhancing the generated responses' reliability. Initial validation using cosine similarity analysis demonstrates promising results, with our system achieving a cosine similarity score of 0.936, outperforming the LLAMA3-8B base model's score of 0.884 in processing construction safety documentation. The preliminary findings show that: 1) our RAG-enhanced system provides safety information access, and 2) our specialized preprocessing techniques effectively synthesize and retrieve safety information, reducing fragmentation and access time.

Keywords –
Construction Safety; Information Retrieval; Large Language Model (LLM); Retrieval-Augmented Generation (RAG); Question Answering System; Vector Database

1 Introduction

Construction remains one of the most hazardous industries globally, with persistent workplace accidents and injuries despite technological advancements [1]. A significant challenge in ensuring construction safety lies in the scattered nature of safety information across numerous documents, standards, and regulations, making it difficult for practitioners to efficiently access and utilize comprehensive safety information in their daily operations.

Current approaches to construction safety documentation typically rely on traditional documentation systems and manual searches, which have proven inefficient and time-consuming for construction practitioners [2]. These conventional methods often require workers to navigate through extensive documentation to find specific safety information, a process that can be particularly challenging given the dynamic and transient nature of construction operations [3]. The construction industry would benefit from an intelligent and reliable system that can quickly retrieve relevant safety information from various sources and present it in a user-friendly format through natural language interaction.

Traditional construction safety information management relies heavily on manual document searching and interpretation, which is time-consuming and prone to human error. Current approaches lack the capability to efficiently integrate multiple safety information sources while providing natural language interaction. To bridge this gap, this study leveraged large language model (LLM), a type of artificial intelligence system trained on vast amounts of text data, and retrieval-augmented generation (RAG) technology to enable precise information retrieval from various safety document sources. By integrating LLM with RAG, our system provides automated, accurate responses to safety-related queries using construction safety documentation as its knowledge base.

The key contributions of this research included: (1)

proposal and development of a novel RAG-based system that enhanced construction safety information accessibility through automated retrieval and natural language generation, and (2) empirical validation of the system's accuracy and effectiveness in a controlled construction safety management environment.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature, Section 3 presents the prototype development, Section 4 details system validation and evaluation, with conclusions and future directions discussed in Sections 5 and 6.

2 Literature Review

This section reviews the development and limitations of safety information retrieval systems in construction, focusing on the progression from traditional keyword-based approaches to advanced language models. We examine how different technological solutions have attempted to address the challenges of efficient information retrieval, particularly highlighting the persistent limitations in query processing and response generation that impact the practical utility of these systems in construction safety applications.

Many natural language processing (NLP) applications have been developed to retrieve construction safety information, each with limitations. Early research focused primarily on discrete tasks such as incident report classification and safety information extraction from unstructured text [4,5]. These traditional NLP approaches have typically treated natural language understanding (NLU) and natural language generation (NLG) as separate components, resulting in fragmented solutions. In 2023, Ricketts et al. highlighted that fully automating safety processes through NLP could be problematic without a strong underlying system [6], particularly when handling complex safety queries that require both accurate understanding and contextual response generation.

The emergence of LLMs has introduced new possibilities for addressing these limitations. Unlike conventional NLP methods, LLMs leverage pre-training on vast, diverse corpora, enabling generalization across various tasks without task-specific data preparation [7,8]. However, while LLMs have demonstrated superior capabilities in analyzing safety incidents [9], commercially available solutions remain too general-purpose for specialized construction safety applications and raise significant data security concerns.

Current construction safety information retrieval primarily relies on manual searches and keyword-based retrieval, which have shown significant limitations in practice. Studies have demonstrated that these document-centric systems require users to navigate through multiple sources and extensive text documents to find specific safety information, thus failing to meet construction workers' diverse needs [10,11]. For example, when workers need to access relevant safety protocols during on-site operations, they often need to search through various safety manuals, regulatory documents, and incident reports separately, making the process time-consuming and risking incomplete information retrieval.

Recent research has demonstrated how RAG technology can enhance model reliability by incorporating accurate, context-specific information during the generation process in healthcare, education, and materials science domains [12–14]. In healthcare, RAG has improved clinical decision-making, pharmacogenomics interpretation, and disease-specific information delivery through specialized applications like GastroBot [12]. In education, RAG has enhanced intelligent tutoring systems and domain-specific knowledge delivery, particularly in technical fields like geotechnical engineering [13]. In materials science, RAG has supported engineering analysis and design processes by enabling access to comprehensive materials databases and enhancing knowledge graph integration [14]. Despite these promising applications across various fields, no study has implemented RAG to enhance construction safety information retrieval, according to our knowledge.

To summarize, research gaps exist in the following two critical aspects: 1) current approaches lack the ability to integrate accurate information retrieval with natural language interaction seamlessly, and 2) existing systems are unable to effectively synthesize and retrieve safety information from multiple sources, leading to fragmented and time-consuming information access processes.

3 System Development

Our proposed system was built upon the LLaMA-3-8B foundation model and integrated the BGE-large embedding model for semantic search capabilities, as shown in Figure 1.

The system architecture comprises three main components: (1) a vector database powered by the BGE-Large embedding model for semantic search functionality, (2) a fine-tuned LLaMA3-8B language model for natural language understanding and response generation, and (3) a context integration module that combines retrieved information with query processing.

The system processes user queries through these components in a streamlined workflow. When a user submits a query, it simultaneously triggers a semantic search in the vector database to retrieve relevant safety documentation while being processed by the language model. The context integration module then combines these elements to generate accurate, contextually grounded responses based on the safety documentation. This architecture ensures reliable safety information delivery while maintaining natural language interaction capabilities.

The development of our system followed two essential steps to ensure its effectiveness. First, we established a comprehensive data processing pipeline to prepare construction safety documentation for the system, as detailed in subsection 3.1. Second, we conducted careful model selection and fine-tuning to optimize the system's performance in the construction safety domain, as detailed in subsection 3.2.

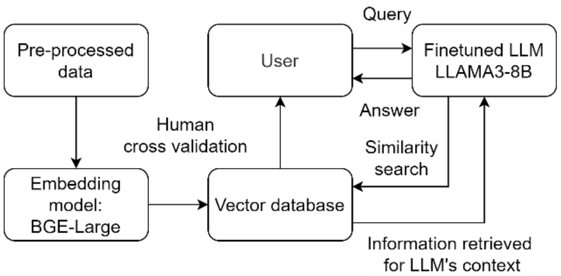

Figure 1 System Design

3.1 Data Processing Pipeline

The data processing pipeline has three key components: data collection, document pre-processing, and information organization and retrieval. First, we established a comprehensive health safety and environment (HSE) information data source as our foundation, as detailed in subsection 3.1.1. Next, a document pre-processing layer converted these diverse document formats into standardized JSON representations while preserving critical safety information, as detailed in subsection 3.1.2. Finally, we implemented an information organization and retrieval system that enabled efficient access to safety information through vector-based search and maintained traceability to original documents, as detailed in subsection 3.1.3.

3.1.1 Data Collection

We obtained a comprehensive collection of health safety and environmental (HSE) documentation from a leading construction company for our information retrieval system. The dataset comprised 1,299 pages and 327,336 words of safety protocols, guidelines, and procedures, as shown in Table 1. This extensive documentation set served as the source material for our information retrieval system. According to industry experts, our documentation set represents standard industry HSE protocols and procedures.

Table 1 HSE Data Source

Document Type	**Count**
Core HSE Manuals	3
Operational Forms	23
PPE Specifications	11

3.1.2 Document Pre-processing

Our document pre-processing transformed raw safety documents into structured data, as shown in Figure 2. The process consisted of two main steps: content extraction and visual processing.

In the content extraction step, we leveraged the Unstructured library to systematically parse PDF documents [15]. This library enabled precise extraction of three types of content while maintaining the structural relationships between different document elements: tables, images, including diagrams and illustrations, and text.

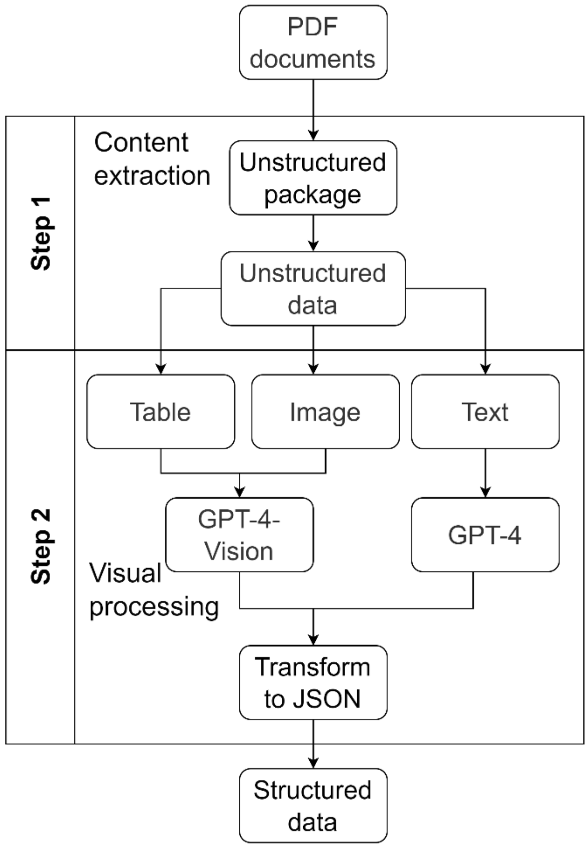

Figure 2 Document Pre-processing Layer

The extracted tables and images presented unique challenges. We addressed these through the implementation of GPT-4-Vision [16], which generated detailed textual representations of the visual content. This approach ensured that critical safety information contained in visual formats was preserved and made accessible for subsequent processing.

All processed content was then consolidated into a standardized JavaScript Object Notation (JSON) structure. This transformation was crucial for construction safety documentation, where the relationships between different content types (e.g., textual instructions, diagrams, and tables) often had significant safety implications. The resulting structured data served as the foundation for generating the fine-tuning dataset and constructing the vector database.

3.1.3 Information Retrieval

The information retrieval process is the key component of our system, implementing similarity search and traceability mechanisms to effectively process and access construction safety documentation.

Our information retrieval system utilizes the BGE-large-en-V1.5 embedding model to convert HSE document into high-dimensional vectors [17,18], enabling similarity search capabilities. We selected this model for its superior performance in capturing technical and domain-specific relationships in construction safety content.

To effectively manage and trace information and their reference source, we developed a traceability framework with multi-vector retrieval capabilities. We designed and implemented a custom JSON mapping system that preserves structural relationships while facilitating efficient retrieval. Our system pre-processes documents, extracts nodes, and preserves comprehensive metadata to maintain bidirectional traceability between processed content and source documents.

3.2 Model selection and Fine-tuning

Model selection and fine-tuning enhances our system's capability from general knowledge to domain-specific expertise in construction safety information retrieval. Fine-tuning complements our RAG framework by enabling the model to better interpret construction safety terminology, recognize domain-specific relationships, and effectively integrate retrieved content into coherent responses. This synergy between fine-tuning and RAG is essential, as neither component alone would achieve the accuracy and relevance needed for construction safety applications. This process consisted of two main stages: base model selection and model fine-tuning, as detailed in subsections 3.2.1 and 3.2.2.

3.2.1 Base Model Selection

Our base models were selected considering the following reasons: 1) less than 8 billion parameters due to GPU memory limitation of our training hardware (RTX A6000), and 2) open source for data privacy protection and security when processing construction safety documents. Based on the context window and massive multitask language understanding (MMLU) score as listed in Table 2, we selected Llama-3-8B. MMLU score is a comprehensive benchmark that evaluates models across 57 subjects, including science, engineering, and professional knowledge.

Table 2 Model Selection Criteria and Performance

Model	Context Window	MMLU
LLaMA-3-8B	**8K**	**66.6**
Gemma-7B	8K	64.3
GPT-Neo-2.7B	2K	25.0

3.2.2 Model Fine-tuning

Our model's fine-tuning consists of two main steps: (1) Q-A pairs generation from structured dataset using Langchain Tuna, and (2) model fine-tuning through LLAMA Factory. Figure 3 illustrates this fine-tuning pipeline.

First, we utilized Langchain Tuna to generate high-quality training data from our structured construction safety dataset. This tool created 11,837 structured question-answer (Q-A) pairs that comprehensively cover various aspects of construction safety protocols [19]. These Q-A pairs are designed to maintain technical accuracy and contextual relevance for safety documentation. Figure 4 presents an example of these generated Q-A pairs.

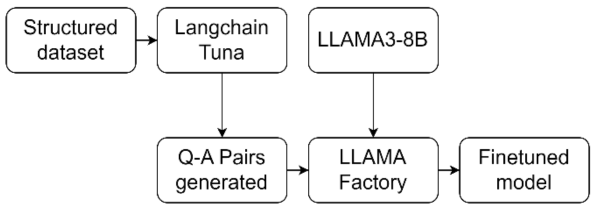

Figure 3 Fine-tuning steps

As our second step, our model's fine-tuning leverages the LLAMA Factory framework [20] to enhance the LLaMA-3-8B model's understanding of construction safety protocols. We chose Low-Rank Adaptation (LoRA) technology [21] for the fine-tuning, as it enables efficient domain adaptation while preserving the model's fundamental capabilities. We used LLAMA Factory using GPUs with CUDA v12 (RTX A6000), with carefully selected hyperparameters shown in Table 3 based on empirical experiments and model characteristics:

Table 3 Fine-tuning Parameters

Parameter	Value
Batch Size	2
Learning Rate	5.00E-05
Training Epochs	3
LoRA+ Ratio	16

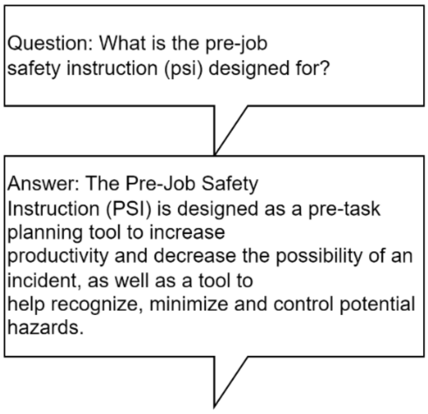

Figure 4 An example of Q-A Pairs generated by Langchain Tuna

While our system architecture incorporates both RAG and fine-tuning, it is important to clarify their distinct yet complementary roles in enhancing construction safety information retrieval. RAG primarily addresses the challenge of information access by retrieving relevant context from the vector database, while fine-tuning addresses the challenge of domain understanding by adapting the model to construction-specific terminology and concepts. Fine-tuning is essential in our approach because construction safety documentation contains specialized terminology, regulatory references, and technical concepts that general-purpose models may not adequately understand. Without domain adaptation through fine-tuning, even a model with perfect retrieval capabilities would struggle to correctly interpret construction-specific queries or synthesize retrieved information into coherent, domain-appropriate responses.

4 Validation and Evaluation

We compared our model with the base LLaMA-3-8B model to evaluate its performance. Our validation and evaluation followed three steps: test case selection, quantitative analysis, and qualitative analysis, as detailed in subsections 4.1, 4.2, and 4.3.

4.1 Test Case Selection

For test case selection in our validation and evaluation, we identified a representative scenario focusing on crisis response protocols from a safety operational manual. We used one specific test question ("What are the steps for Crisis Response?") as our validation case. This case was specifically selected to evaluate our system's ability to handle complex procedural information—a seven-step protocol with sequential dependencies and decision points. Such protocols represent a particularly challenging test case for safety information systems as they require both accurate retrieval of each procedural step and correct preservation of the sequential relationships between steps. This type of complex procedural information is abundant in construction safety documentation and essential for proper safety management. By successfully processing this challenging protocol, our system demonstrates its reliability for handling the diverse information needs in construction safety applications. This case was chosen to evaluate our model's comprehensive understanding capabilities in two key aspects: (1) the ability to process and interpret complex textual information in safety procedures, and (2) the capability to accurately comprehend and convert visual flowchart representations into structured data. The source information was originally presented as a visual flowchart (Figure 5), which was pre-processed using GPT-4V for text conversion and subsequently stored in our vector database.

We designed the test case query as "What are the steps for Crisis Response?" This query was selected to assess the model's capabilities in: comprehending and articulating complete protocol sequences, handling interconnected steps in safety procedures, and maintaining accuracy in complex information retrieval scenarios.

The selection of this test case is based on the real-world case, as protocol-based information represents a substantial portion of construction safety documentation. The ability to accurately retrieve complete step sequences is crucial for maintaining safety compliance in real-world applications.

We present the comparative responses from the LLaMA-3-8B model and our model in Figure 6 and

Figure 7, respectively, demonstrating their performance differences in handling this complex query.

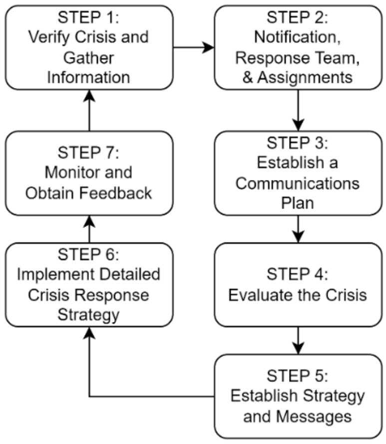

Figure 5 An example adapted from the graph information in the PDF document

4.2 Quantitative Analysis

To evaluate our model's performance quantitatively, we employed two complementary approaches: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) for assessing text generation quality [22], and cosine similarity for measuring semantic relevance between responses and reference documents [23].

The ROUGE metrics we adopted provide a comprehensive evaluation framework that includes precision (accuracy of generated content), recall (coverage of reference content), and F1 score (harmonic mean of precision and recall). This multidimensional analysis through ROUGE metrics offers a complete view of the system's performance beyond simple text overlap. Our evaluation examined three ROUGE variants: ROUGE-1 calculates unigram overlap to assess content coverage, ROUGE-2 measures bigram overlap to evaluate fluency, and ROUGE-L employs the longest common subsequence algorithm to evaluate structural similarity. Additionally, cosine similarity provides a direct measure of semantic alignment between the model's outputs and the source safety documentation.

Following standard practice in NLP evaluation [24], we pre-processed all texts for consistent tokenization and normalization. As shown in Table 4, the experimental results demonstrate substantial improvements across all metrics, indicating that our model better captures both content accuracy and structural coherence in construction safety responses.

As shown in Table 4, our model demonstrates substantial improvements across all evaluation metrics. It is important to note that each ROUGE score reported represents the F1 measure, which balances precision (the percentage of generated content that is accurate) and recall (the percentage of reference content successfully captured). The significant improvements in ROUGE scores (44.9%-60.4%) indicate that our model not only produces more accurate content but also achieves superior coverage of the essential safety information. These comprehensive metrics collectively validate our system's effectiveness for construction safety information retrieval and response generation tasks.

The cosine similarity improvement from 0.884 to 0.936, while appearing modest in absolute terms, is particularly meaningful considering the high baseline. As visualized in Figure 6 and Figure 7, this enhancement manifests in the model's responses being more focused and domain aware. The base model's responses (Figure 6) tend to be more generic, while our model (Figure 7) provides more precise, construction safety-specific responses with clear source references.

Table 4 Model Selection Criteria and Performance

Metric	Our Model	Base Model	Improvement
ROUGE-1	70.8%	10.4%	60.4%
ROUGE-2	47.6%	2.7%	44.9%
ROUGE-L	67.7%	7.3%	60.4%
Cosine Similarity	93.6%	88.4%	5.2%

4.3 Qualitative Analysis

From a qualitative perspective, the comparison between model responses (Figure 6 and

Figure 7) clearly demonstrates the superiority of our model. When processing the crisis response flowchart, our model accurately captured and reproduced the exact procedural steps and decision points from the source documentation. As shown in Figure 7, our model provides direct references to the original safety documentation, with each reference corresponding to specific sections in the source HSE manuals. These references enable users to verify information and access additional context from the original documents, thereby enhancing the traceability and reliability of the provided information.

For instance, when asked about emergency response procedures, our model correctly identified the sequence "Initial Assessment → Hazard Evaluation → Response Selection" without introducing additional, unsourced steps or procedures. In contrast, the base model tended to supplement the flowchart information with general safety protocols do not present in the original documentation, demonstrating hallucination tendencies. The base model's responses often included generic safety knowledge and theoretical frameworks that, while potentially valid, were not part of the source material. Our model's ability to strictly adhere to the source content while maintaining the logical flow of the cyclic graph structure showcases its enhanced reliability for safety-critical information retrieval tasks.

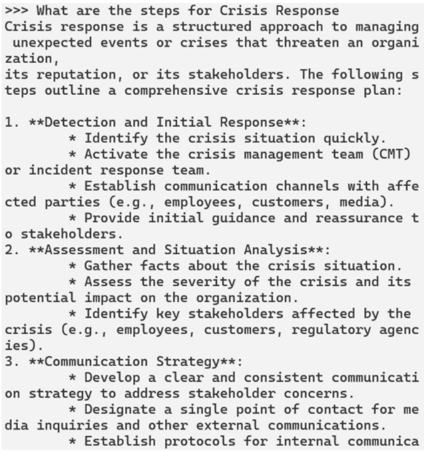

Figure 6 Answer from llama3-8B base model

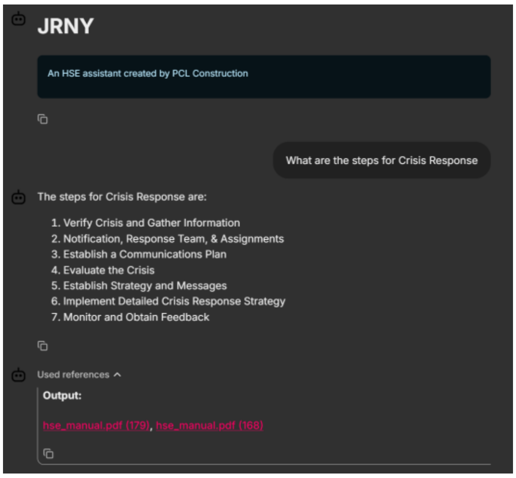

Figure 7 Answer with source references

5 Discussion and Conclusion

This study addressed the critical challenge of efficiently retrieving and processing construction safety information, traditionally hindered by complex documentation structures and technical terminology barriers. We proposed an RAG architecture that combines fine-tuned LLaMA-3-8B with specialized document preprocessing capabilities. Our system demonstrated superior performance in handling construction safety queries through a natural language interface while maintaining high accuracy in information retrieval.

Unlike previous studies that rely on commercial LLMs with data privacy concerns, our approach leverages a locally deployable, smaller model (LLaMA-3-8B) enhanced through domain-specific fine-tuning and RAG architecture. Our approach provides enhanced security for sensitive safety information and greater customizability through construction-specific adaptation.

The reliability of our system was systematically validated through multiple complementary approaches. The quantitative evaluation showed substantial improvements across all metrics, with our model achieving significantly higher ROUGE scores than the base model and improved cosine similarity from 88.4% to 93.6%. These technical improvements manifested in more precise, construction safety-specific responses with better source reference integration, as evidenced in our system's outputs (Figure 7). Particularly noteworthy was our system's ability to correctly process and present the seven-step crisis response protocol, demonstrating its capability to handle complex procedural information with high fidelity. This performance validates both the retrieval capabilities of our RAG architecture and the domain understanding provided by fine-tuning.

This research addresses critical gaps in construction safety information management through two key contributions: (1) the development of a RAG architecture that seamlessly integrates accurate information retrieval with natural language interaction, enabling intuitive access to safety information, and (2) the implementation of specialized document preprocessing techniques that effectively synthesize and retrieve safety information from multiple sources, significantly reducing information fragmentation and access time.

This research has the following limitations. First, while the system excelled at information retrieval, it did not address broader information management challenges such as information creation, organization, and updating. Additionally, its performance varied across different types of construction safety documentation, and its effectiveness in actual construction site conditions needed further validation. Second, the system's capabilities were confined to English-language documentation, restricting its utility in international construction contexts. These limitations suggested opportunities for future research in developing more comprehensive and multilingual implementations.

6 Acknowledgements

This project was supported by the University of Alberta Research Exploration Fund. We would also like to express our gratitude to PCL Industrial Management Inc. for their continued support throughout the projects.

7 References

[1] G.M. Waehrer, X.S. Dong, T. Miller, E. Haile, Y. Men, Costs of occupational injuries in construction

in the United States, Accident Analysis & Prevention 39 (2007) 1258–1266. https://doi.org/10.1016/j.aap.2007.03.012.

[2] J. Teizer, Right-time vs real-time pro-active construction safety and health system architecture, CI 16 (2016) 253–280. https://doi.org/10.1108/CI-10-2015-0049.

[3] L. Wu, S. AbouRizk, Towards Construction's Digital Future: A Roadmap for Enhancing Data Value, in: S. Walbridge, M. Nik-Bakht, K.T.W. Ng, M. Shome, M.S. Alam, A. El Damatty, G. Lovegrove (Eds.), Proceedings of the Canadian Society of Civil Engineering Annual Conference 2021, Springer Nature Singapore, Singapore, 2023: pp. 225–238. https://doi.org/10.1007/978-981-19-1029-6_17.

[4] R. Ganguli, P. Miller, R. Pothina, Effectiveness of Natural Language Processing Based Machine Learning in Analyzing Incident Narratives at a Mine, Minerals 11 (2021) 776. https://doi.org/10.3390/min11070776.

[5] M. Locatelli, E. Seghezzi, L. Pellegrini, L.C. Tagliabue, G.M. Di Giuda, Exploring Natural Language Processing in Construction and Integration with Building Information Modeling: A Scientometric Analysis, Buildings 11 (2021) 583. https://doi.org/10.3390/buildings11120583.

[6] J. Ricketts, D. Barry, W. Guo, J. Pelham, A Scoping Literature Review of Natural Language Processing Application to Safety Occurrence Reports, Safety 9 (2023) 22. https://doi.org/10.3390/safety9020022.

[7] S.A. Prieto, E.T. Mengiste, B. García de Soto, Investigating the Use of ChatGPT for the Scheduling of Construction Projects, Buildings 13 (2023) 857. https://doi.org/10.3390/buildings13040857.

[8] H. Aladağ, Assessing the Accuracy of ChatGPT Use for Risk Management in Construction Projects, Sustainability 15 (2023) 16071. https://doi.org/10.3390/su152216071.

[9] M. Smetana, L. Salles De Salles, I. Sukharev, L. Khazanovich, Highway Construction Safety Analysis Using Large Language Models, Applied Sciences 14 (2024) 1352. https://doi.org/10.3390/app14041352.

[10] A.M. Basahel, Safety Leadership, Safety Attitudes, Safety Knowledge and Motivation toward Safety-Related Behaviors in Electrical Substation Construction Projects, IJERPH 18 (2021) 4196. https://doi.org/10.3390/ijerph18084196.

[11] C.I. Arum, T.O. Osunsanmi, C.O. Aigbavboa, Appraisal of the Challenges to Ensuring Occupational Health and Safety Compliance within the Nigerian Construction Industry, Mocs (2019) 488–495. https://doi.org/10.29173/mocs130.

[12] Q. Zhou, C. Liu, Y. Duan, K. Sun, Y. Li, H. Kan, Z. Gu, J. Shu, J. Hu, GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation, Front. Med. 11 (2024) 1392555. https://doi.org/10.3389/fmed.2024.1392555.

[13] A. Tophel, L. Chen, U. Hettiyadura, J. Kodikara, Towards an AI Tutor for Undergraduate Geotechnical Engineering: A Comparative Study of Evaluating the Efficiency of Large Language Model Application Programming Interfaces, (2024). https://doi.org/10.21203/rs.3.rs-4658661/v1.

[14] M.J. Buehler, Generative Retrieval-Augmented Ontologic Graph and Multiagent Strategies for Interpretive Large Language Model-Based Materials Design, ACS Eng. Au 4 (2024) 241–277. https://doi.org/10.1021/acsengineeringau.3c00058.

[15] Unstructured-IO/unstructured, (2024). https://github.com/Unstructured-IO/unstructured (accessed December 13, 2024).

[16] U. Lee, Y. Jeong, J. Koh, G. Byun, Y. Lee, H. Lee, S. Eun, J. Moon, C. Lim, H. Kim, I see you: teacher analytics with GPT-4 vision-powered observational assessment, Smart Learn. Environ. 11 (2024) 48. https://doi.org/10.1186/s40561-024-00335-4.

[17] P. Zhang, S. Xiao, Z. Liu, Z. Dou, J.-Y. Nie, Retrieve Anything To Augment Large Language Models, (2023). https://doi.org/10.48550/arXiv.2310.07554.

[18] S. Xiao, Z. Liu, Y. Shao, Z. Cao, RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder, (2022). https://doi.org/10.48550/arXiv.2205.12035.

[19] Introducing Tuna - A Tool for Rapidly Generating Synthetic Fine-Tuning Datasets, LangChain Blog (2023). https://blog.langchain.dev/introducing-tuna-a-tool-for-rapidly-generating-synthetic-fine-tuning-datasets/ (accessed December 21, 2024).

[20] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, Y. Ma, LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models, (2024). https://arxiv.org/abs/2403.13372 (accessed December 13, 2024).

[21] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, (2021). https://doi.org/10.48550/arXiv.2106.09685.

[22] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004: pp. 74–81.

https://aclanthology.org/W04-1013 (accessed December 21, 2024).

[23] F. Rahutomo, T. Kitasuka, M. Aritsugi, Semantic Cosine Similarity, (n.d.).

[24] S. Bird, Natural Language Processing with Python, 2009.