Knowledge-Based Emergency Decision Support with Large Language Models and Structured Graph

Fangzhou Lin[1], Zhengyi Chen[2], Mingkai Li[3], Boyu Wang[1], Xiao Zhang[1] and Jack C.P. Cheng[1*]

[1]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[2]School of Civil Engineering, Southeast University, Nanjing, 211189, China
[3]Department of the Built Environment, National University of Singapore, Singapore
flinan@connect.ust.hk, zchenfq@connect.ust.hk, mingkai@nus.edu.sg, bwangbb@connect.ust.hk, xiao.zhang@connect.ust.hk, cejcheng@ust.hk,

Abstract –

Sudden incidents can lead to severe consequences, and traditional emergency-response systems often suffer from incomplete information and delayed decision-making. To address this challenge, we propose an intelligent emergency response system that integrates structured knowledge modeling with large language models (LLMs). The system constructs a knowledge base comprising experiential emergency guidelines and building information and employs a vector database together with a knowledge graph to enable precise storage and retrieval. The experiential-knowledge branch dynamically integrates multi-source emergency plans to produce strategies tailored to specific incidents. The building-information branch leverages a Building Information Modeling (BIM)-derived knowledge graph, coupled with a named entity recognition (NER) module, to support route planning and equipment localization. Experimental results indicate that the proposed system effectively mitigates LLM hallucinations and bridges locale-specific knowledge gaps, thereby enhancing the intelligence and reliability of emergency-response operations.

Keywords –

Retrieval-Augmented Generation, Named Entity Recognition, Emergency Response, Knowledge Graph

1 Introduction

Sudden incidents such as fires, cardiac arrest, and hazardous chemical releases pose serious threats to life and property, especially in large or enclosed buildings where evacuation and rescue are difficult [1]. Although modern buildings are equipped with AEDs, fire extinguishers, and related devices that can improve early emergency response [2], their effectiveness depends on users' awareness of device locations and proper procedures [3].

Emergency response depends not only on equipment availability but also on timely information delivery. Traditional floor-plan maps, static signage, manuals, and public-address systems are often inadequate in complex, dynamic, or power-loss situations and cannot provide real-time, incident-specific guidance [4]. Intelligent emergency response systems can integrate building data, incident conditions, and available resources to provide customized evacuation routes, device locations, and response guidance, while also supporting routine training [5].

Recent studies mainly involve BIM, IoT, and artificial intelligence. BIM has been used to construct building information models that manage spatial layouts, evacuation routes, and the locations of emergency equipment—for example, evacuation-route optimization by Ma et al. [6] and BIM-based evacuation simulation by Rüppel et al. [7]. but often lacks effective integration between emergency equipment and spatial information. IoT supports environmental monitoring and dynamic evacuation guidance [8,9], yet most approaches remain rule-based and offer limited personalization. Artificial intelligence in this context has mainly targeted path optimization and evacuation behaviour modelling, with limited, flexible natural-language interaction. For instance, Zhao et al. [10] modelled evacuation behaviour using random forests to support planning, and Wächter et al. [11] applied deep reinforcement learning to generate escape-route strategies. Yet these methods still centre on route optimization rather than improving access to critical emergency information. Under acute stress, human cognition deteriorates, and current systems' weak interactivity hampers rapid access to needed guidance.

In recent years, LLMs have demonstrated considerable potential in complex semantic understanding, spatial reasoning, navigation decision-making, and tool-augmented agent systems. Accordingly, an increasing number of studies have begun to explore their use in environment understanding, task planning, and multi-source information invocation [12] . However, most existing work has focused on navigation-oriented question answering or tool use in general scenarios, with limited attention to building emergency response, a domain characterized by stringent requirements for timeliness and reliability. Meanwhile, although techniques such as BIM-based evacuation analysis [13], and knowledge graph construction have each been explored in prior studies [14], they are developed in isolation: BIM systems primarily address spatial geometry and path analysis, and knowledge graphs focus on structured relationship modeling. As a result, there is still a lack of an integrated framework capable of unifying emergency knowledge generation, building-space reasoning, and equipment localization and retrieval.

To address these limitations, this study proposes a knowledge-driven intelligent emergency response system that combines LLMs with structured emergency knowledge. The main technical contributions of this work go beyond system integration in three aspects. First, we design a dual-branch knowledge framework that couples a retrieval-augmented experiential knowledge base with a BIM-derived building knowledge graph, enabling both incident-handling guidance and spatial decision support within a unified architecture. Second, we propose a graph-construction method that transforms BIM objects and their topological and semantic relations into a queryable emergency knowledge graph, allowing explicit modeling of space connectivity and equipment allocation. Third, we develop an entity-aware retrieval pipeline in which a DeBERTa-based NER model extracts location and equipment entities from natural-language queries and links them to graph reasoning for evacuation-route planning and emergency-equipment localization. Through this combination, the proposed system improves the precision, contextual relevance, and practical usability of emergency-response support.

2 Methodology

This study proposes a knowledge-driven intelligent emergency response system comprising three core modules that support efficient information retrieval and emergency decision making (Figure. 1). Module I (Emergency-RAG) interacts with users to provide real-time emergency information and, by combining a curated knowledge base with a LLM, generates accurate and readable guidance. For queries about evacuation routes or the locations of emergency equipment, Module II employs a DeBERTa-based model to parse user input, accurately identify building spaces or devices, and produce optimal routes or location information. Module III extracts building-space and equipment data from BIM models and constructs the corresponding knowledge graph. Through the coordinated operation of these modules, the system delivers precise and timely information support in emergency scenarios and enhances the overall intelligence of emergency response.

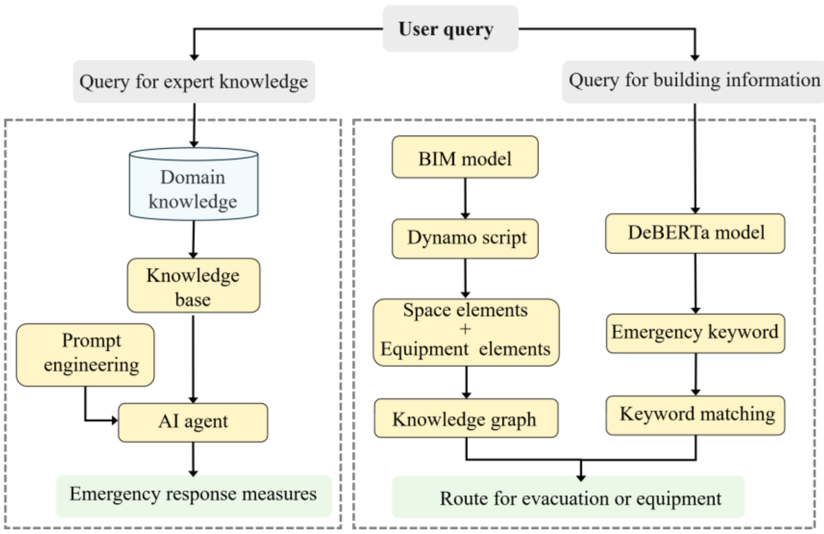

Figure 1. The workflow of emergency response system

2.1 Emergency-RAG for Providing Measures

In the experiential-knowledge branch of the Emergency-RAG platform, the knowledge base is constructed by aggregating high-quality emergency information to provide reliable guidance during critical incidents. The current knowledge base contains about 1,000 structured expert entries, focusing primarily on three core scenarios: fire safety, first aid, and laboratory emergencies. The platform integrates authoritative materials from the Hong Kong Fire Services Department (e.g., fire escape, self- and mutual-aid guidance, CPR, and AED use), as well as resources from multiple international fire departments, the International Committee of the Red Cross (ICRC) first-aid guidelines, the World Health Organization (WHO) emergency response protocols, and university laboratory emergency manuals. All sources are standardized and organized into a structured natural-language knowledge base to improve retrieval efficiency and matching accuracy. The knowledge base is also designed to be extensible, allowing additional event types and customized response strategies to be incorporated for specialized scenarios, thereby ensuring scalability and versatility of the emergency response system.

The workflow of retrieval and response generation is shown in Figure 2. To enable semantic retrieval, we employ the M3E model for text embeddings [15], converting knowledge items into high-dimensional vectors stored in a FAISS vector database [16]. At query time, user inputs are embedded with M3E and retrieved via FAISS using cosine similarity, which rapidly surfaces the most relevant entries. FAISS provides efficient indexing and nearest-neighbor search, supporting fast and accurate guidance while reducing omissions and noise typical of keyword matching.

The retrieved entries are then passed to the ChatGLM large language model [17] to optimize the final response. ChatGLM performs semantic consolidation and context understanding to produce concise, executable guidance. Prompt-engineering refinements further ensure clear, scenario-appropriate instructions. The processed answer is returned to the user, providing high-quality incident-handling guidance under time-critical conditions.

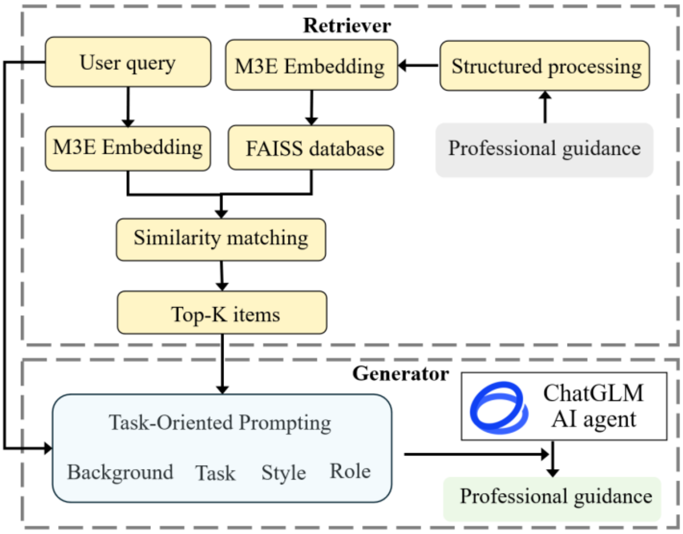

Figure 2. Retrieval and generation of expert knowledge

Task-Oriented Prompting [18] serves as the foundation of our prompt-engineering design and establishes a systematic framework across three dimensions: role setting, task constraints, and output-format control. For role setting, the model is specified as an emergency-response expert with access to a comprehensive emergency knowledge base and decision-support capabilities, which ensures that its outputs meet professional standards and provide sound strategies under complex conditions. For task constraints, the model is required to retrieve, given the user's query, the two most relevant items of emergency knowledge from the knowledge base and, after synthesis, to produce an optimal recommendation. This mechanism effectively reduces hallucinations during generation and improves the reliability of the information. For output-format control, the system mandates concise answers, priority ordering when multiple options are feasible, and step-by-step instructions for complex procedures, thereby enhancing readability and practical utility.

2.2 Establishment of Emergency Knowledge Graph

In the BIM-based emergency knowledge graph, building objects and their relationships are represented as triples (entity–relation–entity), describing the topological and semantic associations among spaces, building components, and emergency equipment. Two core relationships are defined: links and has. The links relationship represents spatial accessibility. When two spaces are connected by a door, a link is created (e.g., Living_room_109 links Corridor_1). Stairs and elevators are modeled as aggregated nodes and connected to the spaces on the floors they serve to maintain cross-floor evacuation continuity. An Open Space node represents outdoor areas, and doors leading outside are connected to this node to ensure evacuation routes reach a safe outdoor area. The has relationship describes the association between spaces and emergency equipment. Emergency devices are modeled as independent nodes and linked to their corresponding spaces (e.g., Lobby_1 has Fire_Extinguisher), enabling equipment querying and localization. During construction, spatial elements are extracted from the BIM model and mapped to graph nodes, while spatial connectivity and equipment configuration are represented as links and has edges and stored in Neo4j. The method relies on building objects and their relationships rather than a specific BIM platform. Although a Revit model is used in this study, the method is also applicable to IFC-based models. By parsing entities such as IfcSpace and IfcDoor, the same graph structure can be generated using the same mapping logic.

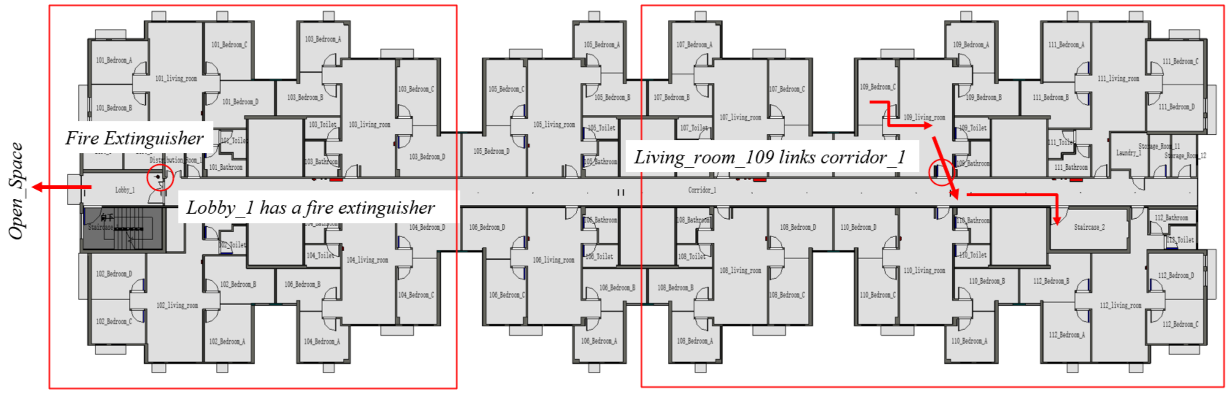

Figure 3. Definition of knowledge graph triplets

2.3 Information Extraction and Retrieval

As illustrated in Figure 4, the building-information retrieval pipeline provides both evacuation guidance and emergency-equipment search. When a user submits a query, the system first invokes a NER model to parse the text, extract key entities and their types, and thus identify the retrieval target. For example, given "Room 1 is on fire, how can I obtain a fire blanket?" or "I am in the control room, how should I escape?", the NER model accurately extracts entities such as "fire blanket" or "control room" and determines whether the query concerns equipment localization or evacuation guidance. The recognized entities are then converted into standardized query for searching the knowledge graph.

Next, based on the entities parsed by the DeBERTa-based NER model [19], the system composes graph queries for the Neo4j database and performs fuzzy search to match the corresponding building entities. The knowledge graph stores spatial connectivity and equipment affiliation, ensuring that queries are precisely mapped to the building environment. Retrieval is conducted using the Cypher query language, and the fuzzy-matching mechanism returns the most relevant results even when user phrasing differs from graph entries. The system then locates the source and target nodes in the graph and determines the search scope according to the query type. For evacuation queries, it enumerates all reachable paths under the spatial-connectivity relations, computes path lengths, and recommends the shortest path. For equipment-location queries, it first retrieves device nodes with the space and then derives the route using spatial connectivity.

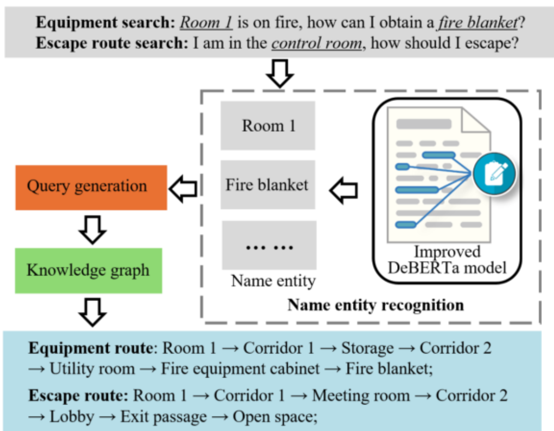

Figure 4. Retrieval process of building information

The process of the NER model identifying key entities is illustrated in Figure 5. The method begins with a natural language query sentence: "Room 1 is on fire, how can I obtain a fire blanket?" The sentence is first converted into a token sequence that the model can process. The sequence starts with a special classification token [CLS], followed by the decomposition of the sentence into a series of subword units (tokens). It ends with a separator token [SEP], forming the complete input sequence.Each token is fed into an embedding layer, where it is transformed into a high-dimensional dense vector representation. These embedding vectors incorporate lexical semantics, positional information, and potential sentence segment information, constituting the initial input for the model.

The embedding vectors are then fed into a multi-layer DeBERTa encoder for deep representation learning. Each encoder layer employs a disentangled attention mechanism that separates token representations into content and positional vectors, enabling attention computation based on content–content and position–position interactions. This mechanism improves the model's ability to capture semantic relationships and syntactic dependencies. The attention outputs are further processed through a feedforward neural network, producing contextualized representations for each token.

The final output of the model is the contextual representation corresponding to each input token. The representation of the [CLS] token at the beginning of the sequence aggregates the global semantic information of the entire sentence and is typically used for sentence-level classification tasks. The representations of other tokens in the sequence encode their specific meanings within the given context, making them suitable for token-level tasks such as named entity recognition or answer span extraction in question-answering systems. In this application example, the representations learned by the model can support downstream natural language understanding tasks, such as identifying user intent from the query sentence, extracting key entities (e.g., "Room 1" and "fire blanket"), or providing deep semantic foundations for generating safety guidance responses.

The training dataset for the NER model is constructed from equipment manuals, emergency management regulations, and design documents, supplemented with LLM-generated synthetic data to improve data coverage. Approximately 70% of the dataset consists of synthetic data, resulting in a corpus of about 10,000 words and 5,000 sentences. The dataset is annotated following the BIO tagging scheme, yielding around 4,000 labeled entities, including equipment entities and location entities. The DeBERT model achieves a precision of 84.2% on this dataset, indicating that most entities can be correctly identified in practical scenarios.

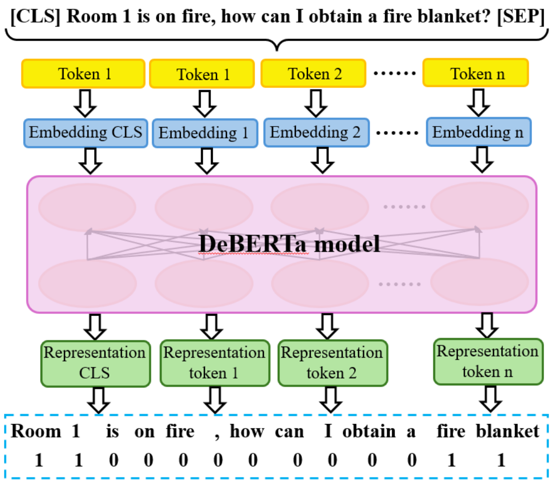

Figure 5. Named entity recognition process and representation based on DeBERTa model

3 Experiment

As illustrated in Figure 6, when a user queries building information, the system generates the result through the following steps. First, an improved DeBERTa-based NER model processes the input to extract key entities, including locations (e.g., “001 bedroom,” “control room”) and devices (e.g., “AED,” “manual alarm button”). Next, the system composes queries tailored to the knowledge graph and retrieves relevant entities via a fuzzy-matching mechanism. It then identifies the source and target nodes in the graph, enumerates all feasible routes under the spatial-connectivity relations, computes their lengths, and recommends the shortest path. The user can follow this recommended path to move through the intermediate spaces and, in the final space, locate the required device.

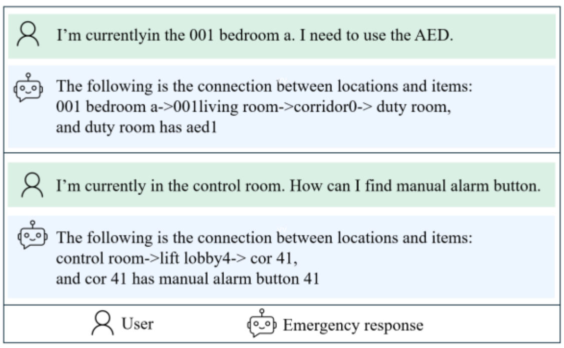

Figure 6. Results of route query

In the expert-knowledge retrieval experiment, we locally deployed Emergency-RAG and configured ChatGLM, DeepSeek, and ChatGPT on the open-source Dify LLM platform for comparative evaluation. To ensure comparability, Emergency-RAG used the proposed prompt-engineering workflow, whereas the baseline models followed an equivalent workflow without background information.

When prompted with “What should I do if I experience poisoning in the laboratory?” (Figure 7), the four models exhibited different levels of coverage of key response items. ChatGPT omitted personnel evacuation. DeepSeek omitted both evacuation and seeking medical care. The original ChatGLM missed the largest number of items, including reporting to management, locating an antidote kit, and other first-aid measures. After integrating the proposed improvements, Emergency-RAG covered all relevant response items, providing a more comprehensive and complete emergency-response guideline.

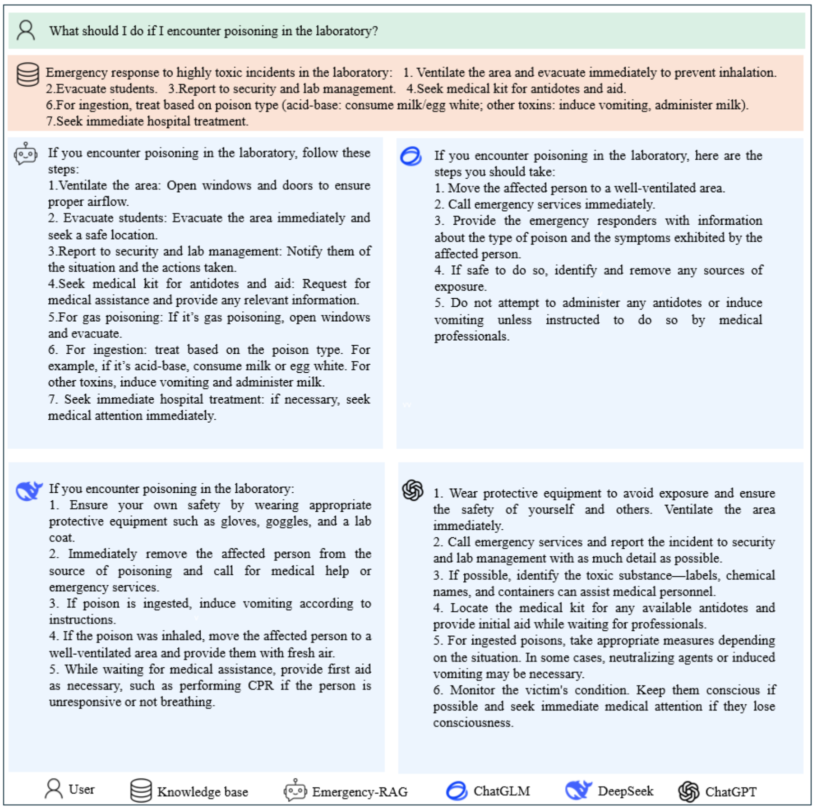

Figure 7. Comparison of query results

BERTScore [20] was employed to evaluate the semantic similarity between the knowledge base and the responses generated by different agents. Unlike traditional lexical-based metrics, BERTScore leverages contextual embeddings to capture deep semantic alignment, making it particularly suitable for assessing the quality of open-ended responses in emergency scenarios. Across multiple queries, the evaluation results are summarized in Table 1. With knowledge-base grounding, Emergency-RAG consistently outperforms the other agents on all metrics, achieving 72.69% precision, 83.03% recall, and a 77.48% F1 score, substantially surpassing competing systems. Moreover, under the constraints of professional guidelines, Emergency-RAG effectively mitigates hallucinations.

Notably, Emergency-RAG can deliver customized responses. For example, it can integrate local knowledge (such as regional emergency numbers) and tailor outputs to user needs. This capability is valuable in diverse domains, including factories, hospitals, schools, transportation hubs, and metro systems. In the case of a gas leak, for instance, prompt ventilation may be appropriate in residential settings when the leak is minor, whereas industrial environments with larger releases require targeted exhaust ventilation rather than general ventilation to prevent flammable concentrations from reaching the explosive limit. Emergency-RAG can construct and apply scenario-specific knowledge bases to provide more targeted guidance.

Table 1. Comparison of query responses from different agents

LLMs	Metric	Case 1	Case 2	Case 3	Case 4	Average
Emergency-RAG	Precision	77.2	69.73	71.4	78.3	74.1
	Recall	84.4	84.58	81.9	83.8	83.7
	F1-score	80.7	76.4	76.3	80.9	78.57
ChatGLM	Precision	61.6	52.74	60.3	63.4	59.5
	Recall	70.5	63.35	63.9	67.9	66.4
	F1-score	65.7	57.6	62.0	65.6	62.7
DeepSeek	Precision	61.9	53.15	58.4	65.6	59.7
	Recall	71.8	62.32	63.8	68.2	66.5
	F1-score	66.5	57.4	61.0	66.9	62.9
ChatGPT	Precision	60.4	50.48	64.1	61.5	59.1
	Recall	71.6	57.4	70.5	72.2	67.9
	F1-score	65.5	53.7	67.2	66.4	63.2

4 Conclusions and Discussion

Sudden incidents pose serious threats to life and property and often lead to severe consequences. To address this challenge, we propose an intelligent emergency-response system that integrates structured knowledge modeling with LLMs. The system builds a knowledge base containing experiential guidance and building information to mitigate LLM hallucinations and gaps in local knowledge. Experiential guidance is stored in a vector database and retrieved via similarity search to ground LLM outputs, thereby providing precise emergency instructions. Building information is organized as a knowledge graph derived from BIM models, which captures relationships between spaces and equipment to support route planning and device localization. Experiments demonstrate that the system effectively integrates heterogeneous data, enhances emergency-response capability and overall intelligence, and rapidly generates strategies tailored to specific scenarios. It is particularly suitable for environments with frequent incidents and high safety requirements. By delivering accurate route planning and equipment localization, the system streamlines response workflows and, through routine preparedness drills, strengthens public awareness and self-rescue skills, helping to overcome the information gaps and decision delays common in traditional approaches.

Despite the aforementioned advantages, the proposed system still has several limitations. First, the system currently relies on manual input of location information, which may be inaccurate and susceptible to human factors, especially in complex or time-critical situations. Second, the performance of the NER model is not yet ideal, and recognition errors may occur in complex sentences or multi-entity scenarios. Future work will focus on improving the model architecture to enhance entity recognition performance. In addition, most experiential knowledge is derived from general online sources, lacking guidance tailored to specific regions, building types, or specialized environments. Due to the limited scale of the current dataset, the system evaluation is mainly based on a small number of case studies, which constrains the comprehensiveness of the results. Future research will conduct more extensive evaluations by incorporating a wider range of test scenarios and retrieval mechanisms, as well as exploring different retrieval strategies to further improve system performance. Furthermore, IoT technologies will be integrated to enable automatic localization, real-time path planning, and equipment navigation, while the integration with maps and visualization tools will provide more intuitive guidance. Finally, expert interviews and field investigations will be conducted to collect localized and domain-specific knowledge, thereby improving the accuracy of the knowledge base and enhancing the overall intelligence of the system.

5 Acknowledgment

This research is supported by the Guangdong Provincial Key Laboratory of Modern Civil Engineering Technology (2021B1212040003). The authors also would like to thank the partial support provided by the Digital Twin-based ESG Platform for Property and Facility Management Industries (PRP/009/23LI).

References

[1] M.O. Sanni-Anibire, A.S. Mahmoud, M.A. Hassanain, B.A. Salami, A risk assessment approach for enhancing construction safety performance, Safety

Science, 121, pp. 15-29. (2020) http://doi.org/10.1016/j.ssci.2019.08.044
[2] F. Lin, B. Wang, Z. Chen, X. Zhang, C. Song, L. Yang, J.C.P. Cheng, Efficient visual inspection of fire safety equipment in buildings, Automation in Construction, 171, 105970. (2025) http://doi.org/10.1016/j.autcon.2025.105970
[3] B. Zhong, X. Pan, P.E.D. Love, J. Sun, C. Tao, Hazard analysis: A deep learning and text mining framework for accident prevention, Advanced Engineering Informatics, 46, 101152. (2020) http://doi.org/10.1016/j.aei.2020.101152
[4] N. Bharosa, J. Lee, M. Janssen, Challenges and obstacles in sharing and coordinating information during multi-agency disaster response: Propositions from field exercises, Information Systems Frontiers, 12(1), pp. 49-65. (2009) http://doi.org/10.1007/s10796-009-9174-z
[5] S. Matarneh, F. Elghaish, F.P. Rahimian, N. Dawood, D. Edwards, Automated and interconnected facility management system: An open IFC cloud-based BIM solution, Automation in Construction, 143, 104569. (2022) http://doi.org/10.1016/j.autcon.2022.104569
[6] G. Ma, Z. Wu, BIM-based building fire emergency management: Combining building users' behavior decisions, Automation in Construction, 109, 102975. (2020) http://doi.org/10.1016/j.autcon.2019.102975
[7] U. Rüppel, K. Schatz, Designing a BIM-based serious game for fire safety evacuation simulations, Advanced Engineering Informatics, 25(4), pp. 600-611. (2011) http://doi.org/10.1016/j.aei.2011.08.001
[8] K. Saini, S. Kalra, S.K. Sood, Disaster emergency response framework for smart buildings, Future Generation Computer Systems, 131, pp. 106-120. (2022) http://doi.org/10.1016/j.future.2022.01.015
[9] H. Jiang, Mobile Fire Evacuation System for Large Public Buildings Based on Artificial Intelligence and IoT, IEEE Access, 7, pp. 64101-64109. (2019) http://doi.org/10.1109/access.2019.2915241
[10] X. Zhao, R. Lovreglio, D. Nilsson, Modelling and interpreting pre-evacuation decision-making using machine learning, Automation in Construction, 113, 103140. (2020) http://doi.org/10.1016/j.autcon.2020.103140
[11] T. Wächter, J. Rexilius, M. König, Escape Route Strategies in Complex Emergency Situations using Deep Reinforcement Learning, 2023 19th International Conference on Intelligent Environments,(2023), pp. 1-8.http://doi.org/10.1109/ie57519.2023.10179101
[12] Z. Lin, D. Zhang, Y. Shen, X. Li, Reasoning text-to-image retrieval with large language models and digital twin representations, Knowledge-Based Systems, 336. (2026) http://doi.org/10.1016/j.knosys.2026.115313
[13] N. Lotfi, B. Behnam, F. Peyman, A BIM-based framework for evacuation assessment of high-rise buildings under post-earthquake fires, Journal of Building Engineering, 43. (2021) http://doi.org/10.1016/j.jobe.2021.102559
[14] G.N. Lilis, M. Wang, K. Katsigarakis, D. Mavrokapnidis, I. Korolija, R. Dimitrios, BIM-based semantic enrichment and knowledge graph generation via geometric relation checking, Automation in Construction, 173. (2025) http://doi.org/10.1016/j.autcon.2025.106081
[15] Y. Wang, Q. Sun, S. He, M3e: Moka massive mixed embedding model,(2023).http://doi.org/https://github.com/wangyuxinwhy/uniem
[16] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H.J.a.p.a. Jégou, The faiss library, arXiv preprint. (2024) http://doi.org/10.48550/arXiv.2401.08281
[17] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H.J.a.p.a. Zhao, Chatglm: A family of large language models from glm-130b to glm-4 all tools, arXiv preprint. (2024) http://doi.org/10.48550/arXiv.2406.12793
[18] E. Basic, A.J.a.p.a. Giaretta, Large Language Models and Code Security: A Systematic Literature Review, arXiv preprint. (2024) http://doi.org/10.48550/arXiv.2412.15004
[19] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint. (2020) http://doi.org/10.48550/arXiv.2006.03654
[20] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y.J.a.p.a. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint. (2019) http://doi.org/http://arxiv.org/abs/1904.09675