MRAG-Based Generative AI Framework for Infrastructure Maintenance Decision Support

Juyoung Jang[1], Seungsoo Lee[1], Mingeon Cho[1], Gichun Cha[1] Solmoi Park[2] and Seunghee Park[1,2]

[1]Department of Global Smart City, Sungkyunkwan University, South Korea
[2]School of Civil, Architectural Engineering and Landscape Architecture, Sungkyunkwan University, South Korea

juyoung0704@skku.edu, skklss@skku.edu, raonik6713@naver.com, ckckicun@skku.edu, solmoipark@skku.edu, shparkpc@skku.edu

Abstract –

As infrastructure systems such as bridges continue to age, maintenance engineers are increasingly required to make timely and consistent decisions based on large volumes of unstructured maintenance documents. In current practice, decision-making relies heavily on inspectors' experience and manual interpretation of guidelines, inspection reports, and past cases, often resulting in inconsistent judgments and limited reuse of accumulated knowledge. This study proposes a Large Language Model (LLM)-based decision support framework for infrastructure maintenance using a Multi-Head Retrieval Augmented Generation (MRAG) architecture. Maintenance documents are preprocessed into semantically coherent text segments, embedded, and organized into facility specific knowledge repositories. Upon receiving a query, the system identifies the relevant facility context and performs parallel semantic retrieval across multiple domain specific heads, enabling facility-oriented standards to be considered while minimizing interference from unrelated documents. When facility specific evidence is insufficient, supplementary retrieval is selectively conducted using a shared common repository. To mitigate hallucination and enhance reliability, the framework applies relevance-based filtering and confidence aware response control. If reliable evidence cannot be identified, the system explicitly notifies the user of insufficient information instead of generating speculative outputs. When sufficient context is available, the LLM synthesizes retrieved evidence to provide decision support responses with explicit document and page references. Overall, the proposed framework promotes transparent and consistent maintenance decision making and demonstrates strong potential as a scalable generative AI solution for diverse infrastructure maintenance domains.

Keywords –

Large Language Model; Multi-Head RAG; Infrastructure Maintenance; Decision Support System

1 Introduction

As infrastructure systems worldwide continue to age, ensuring structural safety while improving the efficiency of maintenance practices has become an increasingly critical challenge [1]. Major infrastructure facilities, including bridges, experience growing inspection and maintenance demands as their service life progresses, highlighting the need for systematic and consistent decision-making approaches. In current practice, infrastructure maintenance decisions are primarily supported by visual inspections and material testing. When defects or abnormal conditions are identified at the component level, relevant standards and guidelines are reviewed to determine appropriate management and response actions. However, such traditional inspection practices rely heavily on inspectors' experience and subjective judgment [2]. Even under similar conditions, different inspectors may reach different conclusions, making it difficult to ensure consistency and reliability in maintenance-related decisions [3]. This variability can negatively affect long term lifecycle management and maintenance planning [4]. These challenges are further amplified by the increasing participation of newly trained inspection personnel, as differences in experience and expertise often lead to discrepancies in judgment despite the application of standardized criteria. Consequently, maintenance outcomes may depend strongly on individual proficiency, underscoring the need for tools that support more objective and repeatable decision-making. Another critical limitation arises from the document centric nature of infrastructure maintenance. Standards, manuals, inspection reports, and repair case records are typically stored in unstructured formats such as PDF or HWP files [5]. Maintenance engineers must

manually search and interpret these documents, which is time-consuming and inefficient, particularly in field environments where timely decisions are required. As a result, similar maintenance issues may be addressed differently across departments or personnel. This challenge is also evident in Korean maintenance practice, where decision making often depends on individual interpretation of document-based standards and guidelines.

Recent advances in artificial intelligence and natural language processing have increased interest in generative AI techniques for effectively utilizing unstructured documents. Large Language Models (LLMs) demonstrate strong capabilities in generating context aware responses and summarizing complex technical texts [6, 7]. However, such models are also prone to hallucination, generating responses that are insufficiently grounded in factual sources [8]. To address this limitation, Retrieval Augmented Generation (RAG) has been introduced as a framework that integrates external document retrieval with language model-based generation [9]. By combining retrieval and generation within a unified pipeline, updated information can be incorporated without retraining the model [10]. However, infrastructure maintenance documents differ significantly in scope and purpose, including common standards, facility specific guidelines, and historical inspection records. When such diverse and unstructured documents are retrieved from a single unified repository, irrelevant information may interfere with retrieval results, motivating the adoption of a Multi-Head Retrieval Augmented Generation (MRAG) structure that performs parallel, domain separated retrieval.

Building on this concept, this study investigates whether a generative AI-based framework can effectively support document-driven decision-making in infrastructure maintenance. The proposed approach aims to reduce dependence on expert judgment, enhance consistency through automated retrieval and interpretation of maintenance documents, and provide a practical foundation for digital transformation and AI-driven automation in infrastructure maintenance workflows.

2 Literature Review

Infrastructure maintenance has been studied as a complex decision making process aimed at ensuring long term structural safety and serviceability [3]. Previous studies have proposed various methods, including standardized evaluation criteria, quantitative indicators, and rule-based or data-driven systems [3, 4]. While these approaches provide a certain level of consistency, they often rely on predefined rules or limited datasets and struggle to flexibly incorporate diverse field conditions and document-based knowledge. In practice, the interpretation of inspection results frequently depends on the experience of inspectors, leading to inconsistent decisions under similar conditions [2]. To address this issue, decision support systems have been introduced; however, many existing systems focus primarily on numerical analysis and provide limited support for utilizing qualitative information embedded in technical documents [5]. Maintenance-related knowledge is largely accumulated in unstructured documents such as standards, manuals, inspection reports, and repair records. Conventional keyword-based retrieval methods offer limited contextual understanding, making it difficult to effectively utilize such document-centric knowledge. Consequently, existing maintenance support systems often fail to deliver timely and context-aware guidance when multiple unstructured documents must be considered simultaneously. Recent advances in natural language processing have brought attention to generative AI technologies, particularly Large Language Models. While LLMs demonstrate strong capabilities in natural language understanding, they are prone to hallucination when responses are generated without sufficient grounding in factual documents [8]. Retrieval-Augmented Generation has been proposed to mitigate this limitation by grounding responses in retrieved documents [9, 10]. However, conventional RAG architectures typically rely on a single unified document repository [14]. In domains characterized by unstructured document sets such as infrastructure maintenance involving multiple facility types and standards this approach may lead to domain mixing and reduced contextual relevance. Retrieved documents that are only partially related to the query may negatively affect response quality. To address this limitation, recent research has highlighted the need for structured retrieval strategies that separate document spaces according to domain or purpose [15]. Multi-head retrieval architectures enable parallel or domain specific retrieval, reducing irrelevant document interference and improving contextual precision. Based on this perspective, this paper explores the application of a Multi-Head RAG(MRAG) framework to infrastructure maintenance decision support, aiming to enhance both retrieval reliability and response consistency. While existing multi-repository or hybrid retrieval systems combine multiple data sources within a unified retrieval space, the proposed MRAG framework introduces a domain-partitioned retrieval architecture with query-level head routing, reducing cross-domain interference and improving contextual consistency. In addition, the integration of Fusion-in-Decoder (FiD) and Corrective RAG (CRAG) enables both multi-evidence reasoning and evidence-level verification.

3 Methodology

This study proposes a document-driven decision support framework for infrastructure maintenance based on a Multi-Head Retrieval Augmented Generation (MRAG) architecture. The proposed methodology is designed to overcome the limitations of conventional inspection support systems by performing parallel retrieval across multiple maintenance document collections with different roles and scopes, and by integrating the retrieved information into a generative reasoning process. Rather than functioning as a simple question-answering system, the framework is intended to support document-based decision making that reflects how maintenance engineers interpret and utilize technical documents in practice.

The overall methodology consists of document preprocessing, Multi-Head RAG Pipeline and Implementation Details, as illustrated in Fig. 1. Each stage is designed to preserve contextual relationships among maintenance standards, guidelines, and inspection records, thereby improving the consistency and reliability of maintenance related decision support.

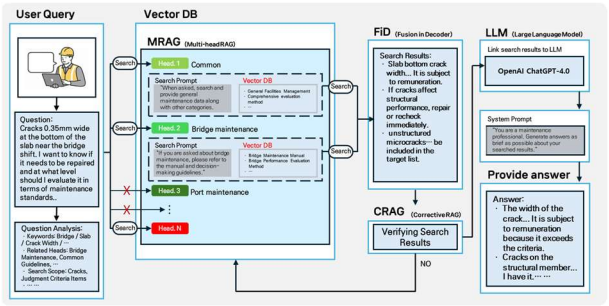

Figure 1. Overview of the proposed Multi-Head RAG pipeline

3.1 Document Preprocessing

The maintenance-related knowledge leveraged in this study is collected from various technical documents, including maintenance standards, inspection manuals, guidelines, and historical inspection reports. These documents are primarily provided in unstructured formats such as PDF or HWP files and contain diverse information types, including descriptive text, tabular data, and procedural instructions. All documents are converted into machine readable text through a document parsing process. The extracted text is then segmented into semantically coherent paragraph level units to preserve contextual consistency. Each segment is associated with metadata such as document name, page number, facility type, and functional role (e.g., standard definition, inspection criterion, or procedural guidance). After segmentation, each document segment is encoded into a dense vector representation using a text-embedding-3-large model. This embedding process maps document segments into a shared semantic space, enabling semantic comparison with user queries during retrieval. By representing both document segments and user queries in the same embedding space, the proposed framework supports similarity-based retrieval that goes beyond exact keyword matching and captures contextual relevance across unstructured maintenance documents. Fig. 2 illustrates the offline preprocessing and embedding workflow used to construct the vector database prior to online retrieval and response generation.

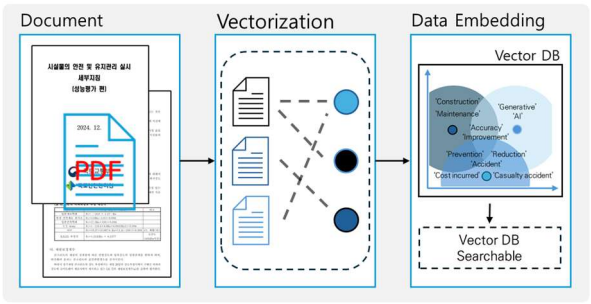

Figure 2. Data embedding process

3.2 Multi-Head RAG Pipeline

This paper adopts a Multi-Head Retrieval Augmented Generation (MRAG) pipeline to support document-driven decision making in infrastructure maintenance. Fig. 1 illustrates the overall architecture and information flow of the proposed Multi-Head RAG pipeline. Unlike conventional RAG architectures that rely on a single unified document repository, the proposed pipeline separates document collections into multiple domain specific heads according to facility type and document purpose. This design aims to reduce irrelevant document interference and improve retrieval reliability in environments characterized by complex corpuses of maintenance documents.

Each head maintains an independent vector database constructed from semantically chunked maintenance documents. Common maintenance standards and general guidelines are stored in a shared Common head, while facility specific heads (e.g., bridge, port, dam) contain documents dedicated to individual infrastructure types. When a user submits a query, the system first identifies the relevant facility context based on query intent and metadata cues and then performs parallel retrieval across the corresponding domain specific head and the Common head. Retrieved document segments from each head are ranked independently based on semantic similarity, and only evidence that satisfies predefined relevance thresholds is forwarded to the generation stage. If

sufficient facility specific evidence is unavailable, the pipeline selectively relies on the Common head to provide generalized guidance, while explicitly indicating the absence of dedicated domain evidence. This fallback strategy prevents cross domain contamination and mitigates the risk of generating unsupported responses.

The retrieved evidence is then passed to a Large Language Model, which synthesizes the information into a structured and context aware response. The generation process is constrained to the retrieved document contents, ensuring that responses remain grounded in verifiable sources. Through domain separation, parallel retrieval, and evidence aware generation, the proposed MRAG pipeline enhances the consistency and transparency of document-based maintenance support.

3.2.1 Query Analysis and Head Selection

The composition and roles of the document heads used in the MRAG pipeline are summarized in Tab. 1. The first stage of the proposed MRAG pipeline focuses on query analysis and head selection. When a user submits a natural language query, the system analyzes the query to identify key elements such as the target facility type, structural component, damage related terms, and maintenance context. This analysis enables the system to determine the most relevant domain-specific document heads to be activated for retrieval. Based on the identified facility context, the pipeline selectively activates the corresponding facility specific head (e.g., bridge, port, or dam, etc.) together with the Common head. The Common head contains general maintenance standards and evaluation guidelines that are applicable across multiple infrastructure types, while facility specific heads store documents dedicated to individual facility categories. Heads that are not relevant to the detected facility context are explicitly excluded from the retrieval process to prevent unnecessary document interference. This selective head activation mechanism allows the MRAG pipeline to narrow the search scope before retrieval is performed, reducing noise from unrelated document domains. By routing each query to an appropriate subset of heads, the system ensures that subsequent retrieval stages focus on documents that are contextually aligned with the user's intent. As a result, the pipeline improves retrieval precision while maintaining flexibility to support both facility specific and cross-domain maintenance queries.

3.2.2 Parallel Retrieval across Multi-Heads

After relevant document heads are selected, the MRAG pipeline performs parallel retrieval across the activated heads. Each head maintains an independent vector database constructed from semantically segmented maintenance documents. This separation allows retrieval to be conducted within a constrained document space that is aligned with the facility context of the query. For a given query, semantic search is executed independently within each activated head. Retrieved document segments are ranked according to their relevance to the query, and the results from different heads are kept separate during this stage. By avoiding

Table 1 Composition of document heads in the MRAG pipeline

Head	Domain Category	Included Documents	Purpose	Head	Domain Category	Included Documents	Purpose
Head 1	Common	Common manual, Seismic performance evaluation	Provides cross-facility maintenance principles and evaluation concepts	Head 4	Dam & River Infrastructure	Dam manual, Weir manual, Levee manual	Provides hydraulic and river infrastructure maintenance standards
Head 2	Bridge & Tunnel	Bridge manual, Tunnel manual	Supplies maintenance criteria for road-related structures	Head 5	Water Supply & Sewerage	Water supply manual, Sewerage manual, Pump station manual	Addresses underground and utility infrastructure maintenance
Head 3	Port & Coastal Infrastructure	Port manual, Harbor-related manuals	Covers port and coastal facility maintenance guidelines	Head 6	Slope & Retaining Structures	Retaining wall manual, Cut slope manual	Covers slope stability and earth-retaining structure maintenance
Head	Domain Category	Included Documents	Purpose				
Head 7	Auxiliary Facilities	Noise barrier manual, Building-related manuals	Supports auxiliary and facility-adjacent structures				

early fusion of diverse document sets, the pipeline prevents irrelevant or weakly related documents from influencing retrieval outcomes. Parallel retrieval enables the system to simultaneously consider both facility specific knowledge and common maintenance standards when applicable. At the same time, heads that were excluded during the query analysis stage do not participate in retrieval, ensuring that unrelated domains do not introduce noise. This design improves retrieval efficiency and enhances the contextual precision of the retrieved evidence.

The retrieved results from each head are subsequently passed to the evidence filtering and generation stages, where relevance thresholds and consistency checks are applied before response synthesis. Through this parallel and domain separated retrieval strategy, the MRAG pipeline establishes a robust foundation for reliable document grounded generation.

3.2.3 Filtering and Controlled Generation

After parallel retrieval across the activated heads, the MRAG pipeline applies evidence filtering and controlled generation stage to ensure that responses remain grounded in reliable document sources. Retrieved document segments from each head are first evaluated based on predefined relevance thresholds, and segments that do not satisfy the minimum relevance criteria are excluded from subsequent processing. To synthesize information from multiple retrieved segments, the pipeline adopts a Fusion-in-Decoder (FiD) strategy[18]. Unlike conventional RAG, FiD encodes retrieved passages separately and integrates them during decoding, allowing the model to synthesize distributed evidence while preserving source-level distinctions. Based on this integration, an initial response is generated by the model using the aggregated evidence. In addition, a Corrective Retrieval-Augmented Generation (CRAG) mechanism [19] is employed to verify the adequacy of retrieved evidence before response generation. Rather than directly filtering individual passages, CRAG operates as a generation level control mechanism that evaluates whether the generated response is sufficiently supported by consistent and relevant evidence. If the generated response indicates insufficient evidence, the pipeline refrains from producing unsupported outputs and triggers a corrective retrieval process (fallback to the Common head) followed by response regeneration. If sufficient evidence is still unavailable after the corrective retrieval process, the system explicitly outputs that no supporting evidence exists, thereby avoiding speculative or misleading responses. Through the combined use of relevance-based filtering, FiD-based evidence fusion, and CRAG-based verification, the proposed MRAG pipeline enforces evidence aware generation. This design reduces the risk of unsupported responses and enhances the transparency and reliability of document-driven maintenance support. Compared with conventional RAG, the proposed MRAG framework restricts retrieval to relevant domain-specific heads, integrates distributed evidence through FiD, and suppresses unsupported outputs through CRAG, thereby improving response relevance and reliability.

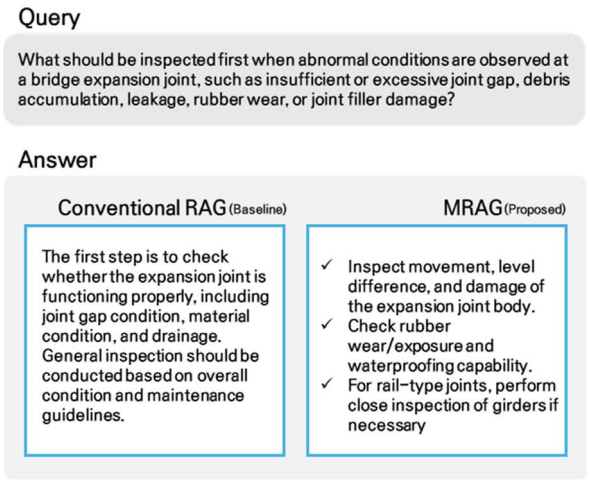

Figure 3. Response comparison for a bridge expansion joint query

As shown in Fig. 3, The conventional RAG produces a generalized and descriptive response without specifying concrete inspection items. In contrast, the MRAG provides explicit and structured inspection criteria based on domain-specific evidence.

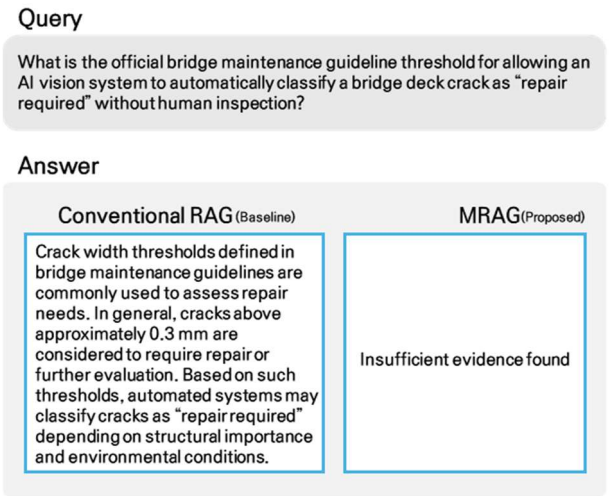

Figure 4. Comparison under insufficient evidence conditions

As shown in Fig. 4, The conventional RAG generates a plausible but unsupported response by introducing crack-width thresholds and extending them to automated decision-making. In contrast, the MRAG suppresses such responses and explicitly indicates insufficient evidence. These examples demonstrate that the proposed MRAG framework improves both response specificity and reliability compared to conventional RAG approaches. Domain-separated retrieval enables more precise evidence selection, while FiD-based integration allows effective synthesis of distributed information across documents. Furthermore, CRAG-based verification prevents unsupported generalization, ensuring that responses remain grounded in verifiable evidence.

3.3 Implementation Details

This section describes the implementation details of the proposed MRAG pipeline, focusing on document preprocessing, vector embedding, and similarity-based retrieval. All maintenance documents were first converted into machine readable text and segmented into semantically coherent chunks to preserve contextual meaning while enabling efficient retrieval [5, 17]. Chunk sizes were determined empirically to balance semantic completeness and retrieval granularity. Each document chunk was transformed into a dense vector representation using a text embedding model. The embedding process maps textual content into a shared vector space, allowing semantic similarity between user queries and document segments to be measured numerically [6, 7]. By representing both queries and documents in the same embedding space, the system supports context aware retrieval beyond exact keyword matching. For similarity-based retrieval, cosine similarity was employed to quantify the semantic closeness between query embeddings and document embeddings [16, 17]. Given two vectors A and B their inner product and Cosine Similarity are calculated as follows:

$$Cosine\ Similarity(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \qquad (1)$$

where A and B denote the embedding vectors of the query and the document segment, respectively, and ||A|| and ||B|| represent their vector norms. In this study, cosine similarity values are typically observed in the range of 0 to 1, where higher values indicate greater semantic similarity between the query and document embeddings. The closer the value is to 1, the smaller the angle between the two vectors, meaning that the semantic directions of the two texts are similar.

Retrieved candidates were ranked according to similarity scores, and only the top ranked segments that exceeded a predefined relevance threshold were selected as retrieval evidence. This threshold-based filtering reduces the influence of weakly related documents and improves the reliability of retrieved results.

The vector databases corresponding to each document head were managed independently, enabling parallel retrieval operations across multiple heads. Retrieval parameters such as the number of returned candidates and similarity thresholds were consistently applied across heads to ensure fair comparison and stable system behavior. Through this implementation strategy, the MRAG pipeline achieves efficient and scalable document retrieval while maintaining consistency across various infrastructure maintenance domains.

4 System Operation and Discussion

Unlike conventional RAG-based systems that retrieve documents from a single unified knowledge base, the proposed MRAG framework explicitly separates retrieval spaces according to facility domains and controls fallback behavior through a Common head. This section demonstrates the operational behavior of the proposed MRAG framework focusing on how architectural differences influence system behavior during real user interactions. Rather than focusing on quantitative performance metrics, the emphasis is placed on illustrating how the system processes user queries, retrieves relevant documents, and generates controlled responses grounded in maintenance documentation.

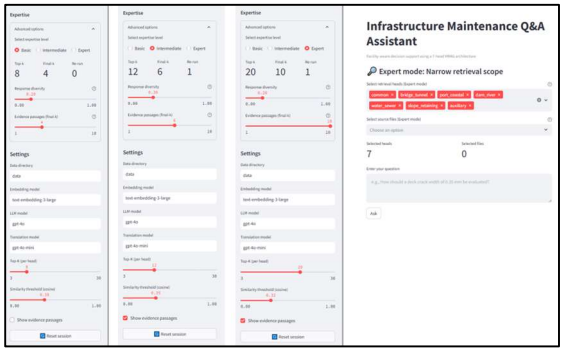

Figure 5. Expertise-level presets and adaptive retrieval parameter configuration.

Fig. 5 illustrates the expertise-level presets provided by the proposed system and their corresponding retrieval parameter configurations. To accommodate users with different levels of experience, the chatbot offers three predefined modes-basic, intermediate, and expert-which automatically adjust key retrieval parameters such as the number of retrieved passages (top-k), the number of evidence passages forwarded to the LLM (final-k), and the similarity threshold. By encapsulating retrieval strategies into expertise presets, the system allows non-expert users to obtain stable and concise responses

without manual parameter tuning, while enabling advanced users to access broader and more detailed evidence when required. This design improves usability and ensures consistent system behavior across diverse maintenance decision making scenarios.

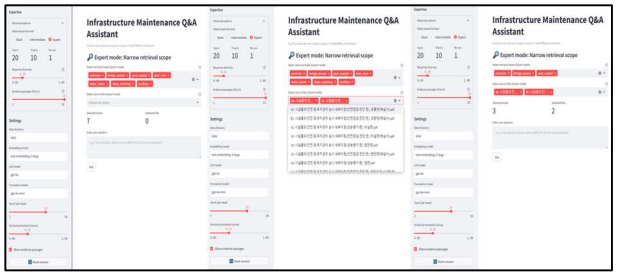

Figure 6. Expert mode interface enabling explicit selection of domain-specific retrieval heads.

Fig. 6 presents the expert mode interface of the proposed MRAG-based system. In expert mode, users can explicitly constrain the retrieval scope by selecting a subset of domain specific heads, such as common, bridge and tunnel, port and coastal infrastructure, dam and river facilities, water supply and sewerage, slope and retaining structures, and auxiliary facilities. Rather than filtering retrieved passages at the search space, the system restricts the retrieval loop itself to the selected head repositories, ensuring that only domain relevant document collections are queried. This mechanism is particularly effective when users have prior knowledge of the target facility type and seek to avoid interference from unrelated domains, thereby improving retrieval precision and response relevance.

The proposed MRAG system incorporates a fallback mechanism to ensure reliable responses under sparse facility specific data conditions. As illustrated in Fig. 7, when no relevant passages are retrieved from the routed facility head, the system triggers a Common-head fallback. The chatbot explicitly informs the user that facility specific evidence is unavailable and generates an answer solely based on shared maintenance guidelines. This design prevents misleading facility level interpretations while maintaining system usability and transparency in document grounded decision support.

Through these examples, the operational behavior of the proposed MRAG framework is demonstrated in realistic maintenance decision making scenarios. The results highlight how domain-separated retrieval, expertise aware parameter control, and transparent fallback mechanisms collectively enhance the reliability and interpretability of LLM-based infrastructure maintenance support systems. From a practical perspective, the proposed MRAG framework has implications beyond isolated query-response tasks. In infrastructure asset lifecycle management, evidence-grounded and domain-consistent recommendations can improve the reliability of inspection and maintenance planning processes. Furthermore, the framework can be integrated with BIM or digital twin systems, where structured inspection knowledge and real-time data are combined to support continuous monitoring and decision-making. This highlights the potential of MRAG as a reliable decision-support tool in safety-critical engineering environments.

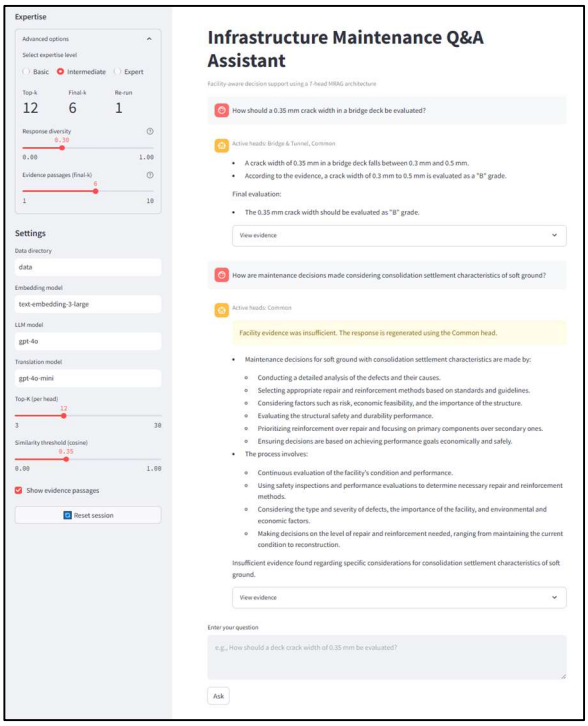

Figure 7. Common-head fallback behavior when facility-specific evidence is unavailable.

5 Conclusion

This study presented a generative AI-based decision support framework for infrastructure maintenance using a Multi-Head Retrieval-Augmented Generation (MRAG) architecture. By organizing multiple maintenance documents into domain specific retrieval heads and a shared Common head, the proposed framework enables facility aware parallel retrieval while preventing interference from unrelated document domains. The explicit Common-head fallback mechanism further enhances reliability by transparently handling cases where facility specific evidence is unavailable, as demonstrated through interactive chatbot-based experiments.

The results indicate that the proposed MRAG framework improves retrieval relevance, response consistency, and transparency compared to conventional

RAG approaches, while supporting practical adoption through flexible, expertise aware user controls. Overall, this study confirms that a Multi-Head RAG architecture with explicit fallback control provides an effective and scalable foundation for AI-assisted infrastructure maintenance support. Future work will focus on extending the framework toward a chatbot-based system that supports bridge damage classification and maintenance judgment, validating its applicability in real world bridge maintenance practice.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used GPT-5.2 in order to improve the readability and language. After using this tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (RS-2024-00336270), (RS-2025-02223612).

References

[1] BROWN, Richard E.; WILLIS, H. Lee. The economics of aging infrastructure. *IEEE Power and Energy Magazine*, 2006, 4.3: 36-43.

[2] GRAYBEAL, Benjamin A., et al. Visual inspection of highway bridges. *Journal of nondestructive evaluation*, 2002, 21.3: 67-83.

[3] SAMSAMI, Reihaneh. A Systematic Review of Automated Construction Inspection and Progress Monitoring (ACIPM): Applications, Challenges, and Future Directions. *CivilEng*, 2024, 5.1: 265-287.

[4] CHO, Mingeon, et al. Development of machine learning-based construction accident prediction model using structured and unstructured data of construction sites. *KSCE Journal of Civil and Environmental Engineering Research*, 2022, 42.1: 127-134.

[5] SOIBELMAN, Lucio, et al. Management and analysis of unstructured construction data types. *Advanced Engineering Informatics*, 2008, 22.1: 15-27.

[6] ZHAO, Wayne Xin, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023, 1.2.

[7] NAVEED, Humza, et al. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2025, 16.5: 1-72.

[8] JI, Ziwei, et al. Survey of hallucination in natural language generation. *ACM computing surveys*, 2023, 55.12: 1-38.

[9] LEWIS, Patrick, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 2020, 33: 9459-9474.

[10] JEONG, Cheonsu. Generative AI service implementation using LLM application architecture: based on RAG model and LangChain framework. *Journal of Intelligence and Information Systems*, 2023, 29.4: 129-164.

[11] SHAHRIVAR, Farham, et al. AI-based bridge maintenance management: a comprehensive review. *Artificial Intelligence Review*, 2025, 58.5: 135.

[12] ZHANG, Chenhong; LEI, Xiaoming; XIA, Ye. Enhancing bridge inspection data quality using machine learning. *Automation in Construction*, 2025, 175: 106182.

[13] BARNETT, Scott, et al. Seven failure points when engineering a retrieval augmented generation system. In: *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*. 2024. p. 194-199.

[14] BESTA, Maciej, et al. Multi-head rag: Solving multi-aspect problems with llms. *arXiv preprint arXiv:2406.05085*, 2024.

[15] WAN, Yuwei, et al. Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. *Advanced Engineering Informatics*, 2025, 65: 103212.

[16] ALSHAMMARI, Ahmad Farhan. Implementation of text similarity using word frequency and cosine similarity in Python. *International Journal of Computer Applications*, 2023, 185.36: 54-59.

[17] QURASHI, Abdul Wahab; HOLMES, Violeta; JOHNSON, Anju P. Document processing: Methods for semantic text similarity analysis. In: *2020 international conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 2020. p. 1-6.

[18] Izacard, Gautier, and Edouard Grave. "Leveraging passage retrieval with generative models for open domain question answering." *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*. 2021.

[19] Yan, Shi-Qi, et al. "Corrective retrieval augmented generation." (2024).