ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0439
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

Retrieval-Augmented Generation-based Question Answering Technology for Construction Safety

Minwoo Jeong[1], Taegeon Kim[2], Seokhwan Kim[3] and Hongjo Kim[4*]

[1]*Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea,* E-mail: minuj9855@yonsei.ac.kr
[2]*Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea,* E-mail: ktg9655@yonsei.ac.kr
[3]*Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea,* E-mail: yyksh2019@yonsei.ac.kr
[4*]*Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea,* E-mail: hongjo@yonsei.ac.kr (corresponding author)

Abstract: This study investigates the potential of Retrieval-Augmented Generation (RAG)-based Question Answering (QA) technology for accurate and relevant responses of Large Language Models (LLMs) to construction safety-related queries. Despite LLMs' advancements, their application, especially a Q&A Chatbot faces challenges due to hallucination and lack of domain-specific details. This study explores RAG's potentials to mitigate these issues by making LLM refer to external databases, such as the OSHA Field Safety and Health Manual, for generating precise and factual contents. A comparative analysis of different RAG technologies—Naïve-RAG, Rerank-RAG, and Iterative Retrieval-Generation—demonstrates their effectiveness over traditional LLM approaches. The findings highlight RAG's significance in producing structured, fact-based responses, underscoring its superiority in addressing the domain-specific informational needs regarding construction safety practices. This research marks a step forward in the application of generative AI technologies to enhance safety standards and practices within the construction industry.
Key words: Retrieval-Augmented Generation (RAG), Large Language Model (LLM), Construction Safety, Question-Answering (QA)

1. INTRODUCTION

Recently, efforts to introduce artificial intelligence technology in all industries have been active, and the construction industry is also trying to follow suit [1,2]. Despite the successful integration of various AI technologies into the construction domain including computer vision [3], the application of Large Language Models (LLMs) faces significant limitations. One challenge is the issue of hallucination; LLMs tend to generate content that is nonsensical or inconsistent with the user's query or established facts [4,5]. Another barrier to adopting LLMs in the construction industry is that the content produced by LLMs often lacks the detail necessary for direct application to construction-specific tasks, such as generating safety contents [6,7] for construction site workers based on safety standards [5]. Because LLMs are generally pre-trained on large datasets, they possess broad knowledge but lack detailed or technical understanding in specific domains.

To overcome these significant challenges, the technology of Retrieval-Augmented Generation (RAG) [8] has emerged as a key solution. In essence, RAG enables LLMs to generate responses based on an external database rather than solely relying on their pre-existing knowledge. For instance, in construction safety management, the external database could include documents containing construction-specific information, such as the OSHA (Occupational Safety and Health Administration) Field Safety and Health Manual [9], which offers extensive safety rules for workers in various fields. By generating responses based on domain-specific knowledge, it is anticipated that the outputs from LLMs will be more accurate and fact-oriented, compared to those not utilizing RAG [10].

This paper introduces different types of RAG technology that are expected to have a high potential for construction applications. First, literature review related to LLM applications on construction domain and efforts to reduce the hallucination problem is conducted. Next, the proposed method is presented. This includes detailed explanations regarding the workflow of RAG, and evaluation metrics. Then, details about the experiment is following in the next section. Lastly, conclusion and further studies will be presented.

2. RELATED WORKS

There have been massive amounts of studies on NLP applications in the construction domain with a rise of pre-trained LLMs such as GPT [11], PaLM [21] and LLaMA [22]. The applications have been extensively utilized across various applications, encompassing Building Information Modeling (BIM), project management, construction safety, and more. Amer et al. [17] proposed a solution for automatic matching between look-ahead planning tasks to master schedule activities, utilizing fine-tuned GPT-2 to update existing master schedule activities with corresponding look-ahead plans. Here, GPT-2 is trained to generate look-ahead planning task descriptions. Zheng et al. [18] presented a prompt-based BIM 3D visualizations with retrieved information from BIM according to user's Natural Language (NL) query.

There has been a lot of prior research on application of infrastructure as well. Kim et al. [19] used fine-tuned BERT to extract infrastructure damage information from text data [23]. Kim et al. [20] introduced a BERT-based Quality Assurance (QA) technology by retrieving information from construction specifications. In both studies, BERT was fine-tuned for specific downstream tasks such as question answering and retrieving relevant paragraphs from specifications. However, these existing applications of NLP faces a challenge. They either generate simple contents that do not require reference to professional documents, or utilize fine-tuning for detailed responses; therefore, the generated responses are not suitable to be applied directly to the tasks demanding construction expertise since there are risk of hallucination issues. In addition, even if that problem is solved by fine-tuned LLMs, fine-tuning is time and effort intensive.

To overcome the limitation, this research introduces RAG technology for construction safety information generation. Leveraging this advancement, state-of-the-art LLMs are now equipped to produce precise and comprehensive responses based on construction specifications. The principal contributions of this manuscript are twofold: firstly, the deployment of RAG-based QA technology, complemented by a hybrid evaluation of the outputs using both human and LLMs across various RAG frameworks; and secondly, the exploration of prospective applications of RAG technology within the construction sector.

3. METHODOLOGY

3.1. Dataset generation

In this study, OpenAI's GPT-4 Turbo [12] was utilized to generate questions related to construction safety regulations. This model is capable of generating responses to complex queries based on its extensive natural language processing abilities. The question dataset consists of 50 questions about hazard factors at construction sites and the safety regulations that must be observed on-site. This dataset will be used as an input to different language models and to evaluate the responses. The details will be addressed in Section 4. The request to the model consisted of two messages. The first part of the prompt (role: "system") specified the task: "Your task is to generate 50 questions that can elicit answers based on OSHA guidelines, but the questions themselves should not obviously appear to be crafted directly for OSHA material. Provide detailed and accurate information in a clear and understandable manner." The second part of the prompt (role: "user") presented a concrete request to "Generate 50 questions about construction safety regulations." Based on the above settings, the GPT-4 model was employed to submit the request, which resulted in the generation of 50 questions concerning construction safety regulations. Part of them is shown in Table 1.

3.2 Framework of RAG technologies

As mentioned before, RAG enables LLMs to generate responses solely based on an external database without relying on their prior knowledge. RAG begins with user's input query. Then, a retriever module searches for the relevant documents from the external database. An external database comprises a vast collection of text chunks indexed from various data types, including PDFs, DOCs, PPTs, APIs, and more [13]. When relevant documents are retrieved, they are augmented to the query and then input to LLMs, in other words, generators. Finally, the generators generate a response to the input and the RAG is done. This is the brief and the most basic version of RAG: Naïve-RAG [10].

With the basic framework, now Naïve-RAG can be developed into Advanced-RAG [10] by optimizing either the retriever or the generator, or both. In this study, two of the Advanced-RAG technologies are adopted: Rerank-RAG [14] and Iterative Retrieval-Generation (Shown in Figure 1) [15]. Rerank-RAG focuses on postprocessing the retrieved relevant documents. After retrieving a set of relevant documents, a module called reranker is used to rerank the retrieved documents, but of smaller numbers. By going through this two-stage retrieval system, rerankers can outperform other common embedding models.

Table 1. Examples of question dataset

No.	Question dataset
1	What are the guidelines for safely operating heavy machinery on construction sites?
2	How should hazardous materials be handled and stored on construction sites?
3	What is the recommended procedure for reporting accidents or unsafe conditions on a construction site?
4	What training is required for workers to safely perform high-risk tasks on construction sites?
5	How are workers protected from electrical hazards on construction sites?
6	What precautions should be taken when working at heights on construction sites?
7	How should construction sites prepare for emergencies or natural disasters?
8	What are the guidelines for proper ventilation in enclosed construction spaces?
9	How should noise levels be managed on construction sites to protect workers' hearing?
10	What are the requirements for personal protective equipment (PPE) on construction sites?

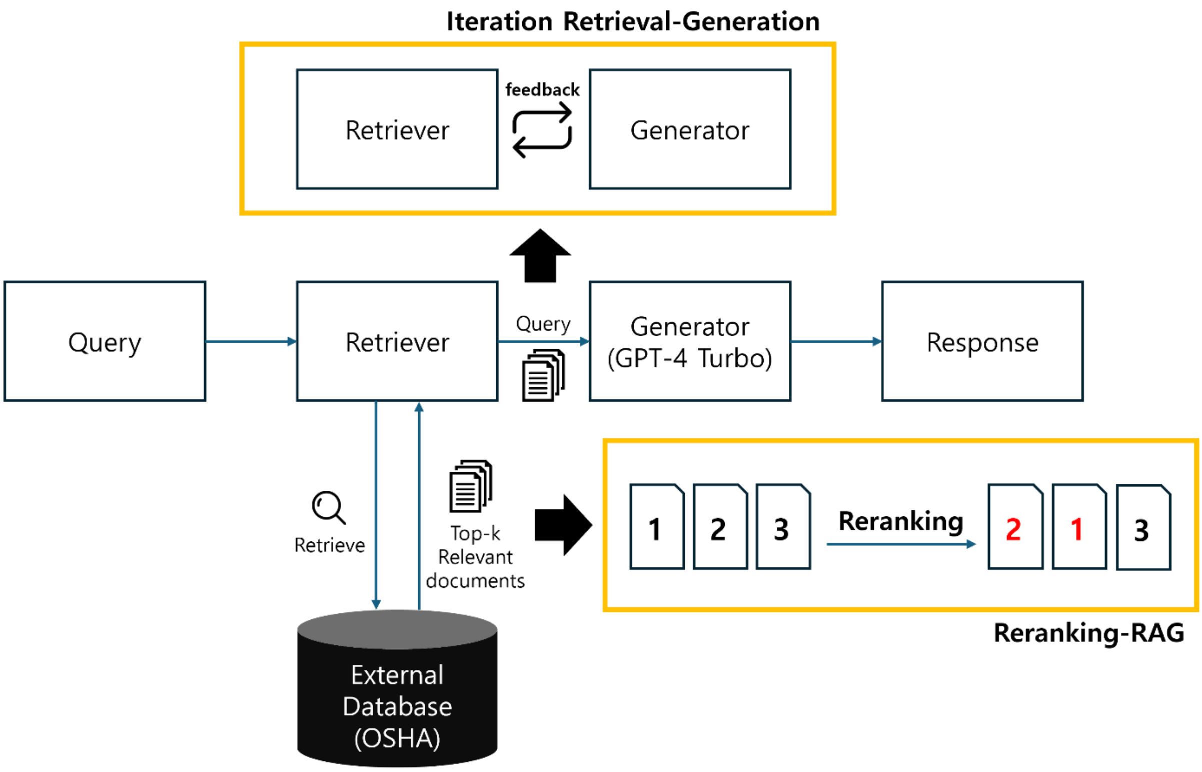

Figure 1. Several types of RAG framework adopted for this paper; It illustrates how Naïve-RAG(in the middle) can evolve into Reranking-RAG and Iteration Retrieval-Generation.

Iterative Retrieval-Generation, as it can be known by its name, repeats retrieval-generation cycles for improved reasoning. When the retriever gives the generator user's query and the relevant documents as an input, generator evaluates the input and decides whether the input contents has failed the evaluation. If so, the generator gives feedback to the retriever so that it can modify its input contents. With these three different RAG technologies, this paper conducted a pilot experiment on how RAG can be applied to the construction domain.

3.3. Evaluation metrics

3.3.1 LLM evaluation

For the evaluation of responses from different RAG technologies, the metrics compare the relevance and faithfulness between the Query (Question), Context (Retrieved documents), and Response (Generated answer). These two evaluation metrics are conducted by the LLMs [25]. The first evaluation metric is faithfulness, where the LLM determines how faithful the response is to the context with a given prompt. This study introduces a methodology that supports this evaluation by requiring a binary response of either "YES" or "NO" to ascertain whether any part of the provided context supports the information in question. This approach emphasizes the importance of direct relevance and support within the context for validating the faithfulness of specific information. An affirmative ("YES") is warranted if there is any segment of the context that corroborates the information, highlighting the method's principle that the presence of supportive evidence within the context is crucial for the validation of responses generated by RAG technologies.

Second evaluation metric is called relevance that assesses the alignment between a query, its context, and the generated response. This evaluation is conducted through a binary "YES" or "NO" mechanism, determining whether the response accurately reflects and aligns with the given context information. The relevance metric hinges on the principle that a response must not only answer the query but do so in a manner that is faithful to the context provided. This ensures that the generated responses are not only relevant but also contextually appropriate, enhancing the reliability and effectiveness of RAG technologies in producing accurate and pertinent information. Through this approach, the framework emphasizes the critical importance of the interplay between query, response, and context in the evaluation of generated content, promoting a nuanced understanding of relevance within the domain of RAG technology assessments.

3.3.2 Human evaluation

Human evaluation plays a crucial role in understanding the performance of NLP models [24]. This reliance on human evaluation stems from the challenge of accurately measuring certain characteristics of texts through automated evaluation metrics alone. In this study, human evaluation is conducted for the assessment of the responses from not only RAG-based LLM models, but also from LLM-only model, which in this study is GPT-4 Turbo [12]. The evaluation encompasses two criteria: factuality and relevance [26]. Factuality, defined as a measure indicating the extent to which generated answers are based on external sources and include specific information, involves a rigorous process where each response is thoroughly examined against OSHA standards through manual review. Relevance, on the other hand, evaluates the relevance of the responses to the content of the questions posed. Both dimensions employ a Likert scale for rating, ranging from 1 to 5. This evaluative process was conducted by the researchers themselves, ensuring a thorough and nuanced understanding of the model outputs in relation to the construction safety domain.

4. EXPERIMENTS

4.1. Experiment setup

In the experiment, responses generated by four different generators were evaluated. Three of them are RAG technologies, each of which are Naïve-RAG, Rerank-RAG, and Iterative Retrieval-Generation. The other one is LLM-only model that RAG is not applied, and therefore have no access to the external database. Since all the other comparison targets could retrieve the external database, additional prompt was generated for the LLM-only model so that it can generate responses based on the OSHA guidelines as much as possible. The prompt, designated with the role "system," outlines the task at hand: "You are an AI assistant with deep knowledge in OSHA Field Safety and Health Manual. Your objective is to deliver detailed and accurate information in a concise and comprehensible manner, based on the OSHA Field Safety and Health Manual. Responses should be formulated within 150 tokens." This instruction establishes the framework for the AI's operation, emphasizing the need for precision, clarity, and adherence to the constraints of the OSHA guidelines while ensuring the responses remain concise.

For the external database, OSHA Field Safety and Health Manual was selected. Its text contents were split into a chunk size of 1024 tokens. For the LLM that used as a generator, determinator in evaluation process, GPT-4 Turbo model is employed.

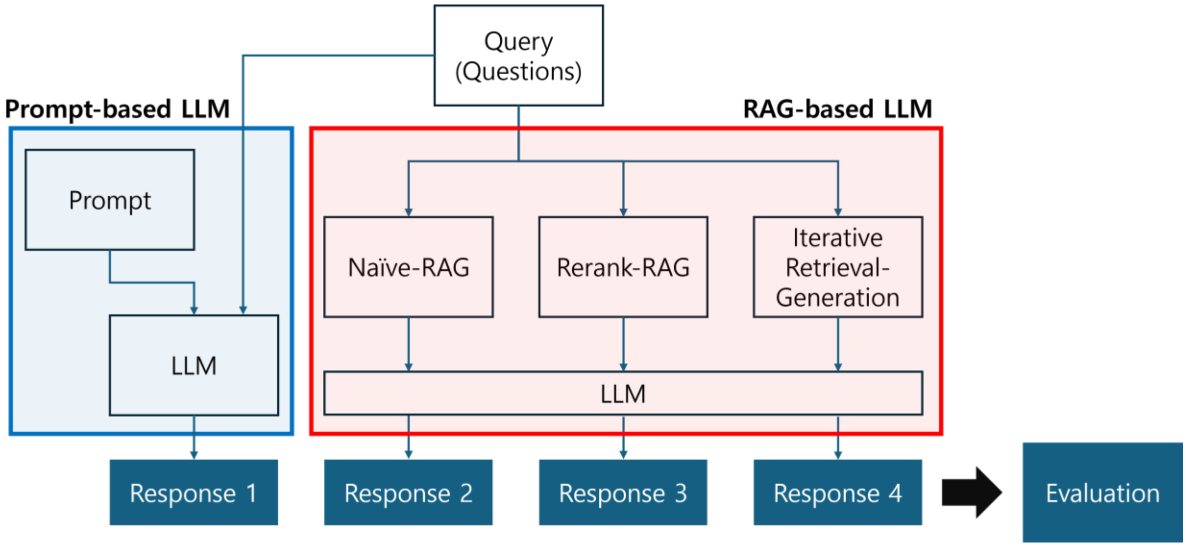

Figure 2. The overall flowchart of the experiment

4.2. Results and discussion

Table 2 and 3 shows the results of LLM evaluation and human evaluation of the proposed technologies each. First of all, LLM evaluation of the three different RAG technologies are conducted. Both faithfulness and relevance were calculated as an average out of 50 questions in total. Rerank-RAG and Iterative Retrieval-Generation achieved 94% of average faithfulness, meaning that in only 3 cases the responses were not faithful enough to the retrieved context. Iterative Retrieval-Generator achieved 80% of relevance, which means that in 10 cases the relevance between the query, context and the response was not consistent enough.

Human evaluation was conducted, revealing that LLM-only models tend to generate general responses, which often do not quantitatively represent the given external database. Although LLM-only models are adept at reasoning in response to the given query, leading to higher relevance compared to RAG-enhanced LLMs, they struggle with precision in numerical representation. This discrepancy is attributed to RAG-enhanced LLMs being significantly influenced by the content of retrieved documents during response generation. If the retrieved documents contain irrelevant information, it negatively impacts the scoring. Additionally, in the scoring process, responses that accurately state the absence of information not found in the content were also awarded high marks (5 points). This observation suggests that while RAG technology helps in sourcing content-related specifics, its effectiveness is contingent upon the relevance and accuracy of the underlying documents, thereby affecting the overall performance in scenarios where precision and relevance are critical.

Furthermore, qualitative evaluation between the responses of LLM-only model and Naïve-RAG is conducted. Since the goal of applying RAG to construction domain is to create fact-based detailed information based on construction-specific knowledge, these responses were compared qualitatively, with respect to OSHA guidelines.

Table 2. LLM evaluation of faithfulness and relevance of different RAG technologies

Models	Faithfulness	Relevance
Naïve-RAG	92% (46/50)	70% (35/50)
Rerank-RAG	**94%** (47/50)	78% (39/50)
Iterative Retrieval-Generator	**94%** (47/50)	**80%** (40/50)

Table 3. Human evaluation of factuality and relevance of generated responses

Models	Factuality		Relevance	
	Mean	STD	Mean	STD
LLM-only	3.3	1.1	**4.34**	0.65
Naïve-RAG	4.14	1.31	4.04	1.34
Rerank-RAG	**4.36**	1.07	4.18	1.16
Iterative Retrieval-Generator	4.14	1.32	4.04	1.36

Table 4 shows the generated responses of LLM-only and Naïve-RAG respectively. Upon initial examination, both responses to the question appear to aptly address it based on the OSHA guidelines. However, the objective is to generate answers that are both factual and organized. A closer inspection of the LLM-only model's response reveals content not actually included in the OSHA guidelines' "Electrical Safety" chapter, indicated in italicized text. Conversely, the Naïve-RAG response is meticulously generated based on the OSHA guidelines, with the necessary content from the "Electrical Safety" chapter presented in sequence. For instance, "Workers are ... unguarded sources." is categorized under Procedures, "Additionally, ... contact water." under Restrictions, and "Training ... electrical hazards." under Training, each distinctly identified. In contrast, the content from the LLM-only response that pertains to the OSHA guidelines is generated without regard to sequence.

This occurrence of hallucination when not employing RAG technology demonstrates the necessity of integrating RAG technology into the construction industry, proving its value in ensuring accurate and structured responses.

Table 4. An example of the responses generated from LLM-only and Naïve-RAG

	Responses	
Question	**LLM-only**	**Naïve-RAG**
How are workers protected from electrical hazards on construction sites?	Workers are protected from electrical hazards on construction sites through OSHA regulations, which mandate clear guidelines such as ensuring all electrical equipment is properly grounded or *double-insulated*, implementing lockout/tagout procedures to control hazardous energy, *maintaining safe distances from overhead power lines*, providing Personal Protective Equipment (PPE) when necessary, and requiring regular safety training and awareness programs for all employees to recognize and avoid electrical hazards. Compliance with these standards helps to prevent electrical shocks, burns, and electrocutions.	Workers are protected from electrical hazards on construction sites by following specific safety measures. These measures include assessing areas for electrical hazards, inspecting locations after a disaster, stopping work and contacting the Responsible OSHA Manager when hazards are identified, and ensuring that unqualified persons do not approach uninsulated or unguarded sources. Additionally, workers are required to wear appropriate clothing and not wear conductive jewelry or clothing that may present a contact hazard. They are also instructed to report any potential exposure to electrical hazards, not conduct electrical repairs, not operate equipment if there is a hazard, and not allow electrical equipment or components to contact water. Training is also provided to workers to ensure they have the necessary knowledge and skills to work safely around electrical hazards.

5. CONCLUSION

The study delves into the utilization of Retrieval-Augmented Generation (RAG) technology to bolster safety on construction sites, comparing its efficacy with that of traditional Large Language Models (LLMs). The research underscores the capability of RAG to produce responses that are not only more accurate and structured but also deeply rooted in facts, thanks to the integration of the OSHA Field

Safety and Health Manual. A detailed comparative analysis revealed that RAG technologies, notably Rerank-RAG and Iterative Retrieval-Generation, offer greater fidelity and relevance in addressing queries related to construction safety regulations. This accentuates RAG's edge in catering to specific information requirements and mitigating the hallucination issues commonly associated with LLM-only approaches. The study points to the substantial advantages of embedding RAG within construction safety protocols, representing a significant leap towards leveraging AI to elevate industry norms and enhance worker safety. Future research will involve human evaluation conducted by experts, which would reinforce the validity of the evaluation. It is anticipated that the application of RAG technology in the construction domain could extend beyond the field of construction safety, potentially leading to the development of specialized chatbot systems for on-site assistance and administrative support systems powered by the interpretation of construction-specific documents.

ACKNOWLEGMENTS

This research was conducted with the support of the “2023 Yonsei University Future-Leading Research Initiative (No. 2023-22-0114)” and the “National R&D Project for Smart Construction Technology (No. RS-2020-KA156488)” funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation. This study referenced the codes from ‘LlamaIndex [16]’. All the related codes and information can be accessed through ‘LlamaIndex (https://github.com/run-llama/llama_index)’.

REFERENCES

[1] P. Ghimire, K. Kim, M. Acharya, Generative AI in the Construction Industry: Opportunities & Challenges, (n.d.).
[2] M. Regona, T. Yigitcanlar, B. Xia, R.Y.M. Li, Opportunities and Adoption Challenges of AI in the Construction Industry: A PRISMA Review, Journal of Open Innovation: Technology, Market, and Complexity 8 (2022) 45. https://doi.org/10.3390/joitmc8010045.
[3] S. Paneru, I. Jeelani, Computer vision applications in construction: Current state, opportunities & challenges, Automation in Construction 132 (2021) 103940. https://doi.org/10.1016/j.autcon.2021.103940.
[4] A. Saka, R. Taiwo, N. Saka, B.A. Salami, S. Ajayi, K. Akande, H. Kazemi, GPT models in construction industry: Opportunities, limitations, and a use case validation, Developments in the Built Environment 17 (2024) 100300. https://doi.org/10.1016/j.dibe.2023.100300.
[5] V. Rawte, S. Chakraborty, A. Pathak, A. Sarkar, S.M.T.I. Tonmoy, A. Chadha, A. Sheth, A. Das, The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023: pp. 2541–2573. https://doi.org/10.18653/v1/2023.emnlp-main.155.
[6] Y. Ding, M. Liu, X. Luo, Safety compliance checking of construction behaviors using visual question answering, Automation in Construction 144 (2022) 104580. https://doi.org/10.1016/j.autcon.2022.104580.
[7] A.B. Saka, L.O. Oyedele, L.A. Akanbi, S.A. Ganiyu, D.W.M. Chan, S.A. Bello, Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities, Advanced Engineering Informatics 55 (2023) 101869. https://doi.org/10.1016/j.aei.2022.101869.
[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, (2021). http://arxiv.org/abs/2005.11401 (accessed October 23, 2023).
[9] C. Brown, OSHA Field Safety and Health Manual, (n.d.).
[10] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, H. Wang, Retrieval-Augmented Generation for Large Language Models: A Survey, (2024). http://arxiv.org/abs/2312.10997 (accessed February 5, 2024).
[11] Brown et al., Language Models are Few-Shot Learners, (2020). http://arxiv.org/abs/2005.14165 (accessed September 19, 2023).

[12] OpenAI et al., “GPT-4 Technical Report.” arXiv, Dec. 18, 2023. doi: 10.48550/arXiv.2303.08774.
[13] Llama Hub, (n.d.). https://llamahub.ai/ (accessed February 15, 2024).
[14] M. Glass, G. Rossiello, M.F.M. Chowdhury, A.R. Naik, P. Cai, A. Gliozzo, Re2G: Retrieve, Rerank, Generate, (2022). http://arxiv.org/abs/2207.06300 (accessed February 15, 2024).
[15] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, W. Chen, Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy, (2023). https://doi.org/10.48550/arXiv.2305.15294.
[16] J. Liu, LlamaIndex, (2022). https://doi.org/10.5281/zenodo.1234.
[17] F. Amer, Y. Jung, M. Golparvar-Fard, Transformer machine learning language model for auto-alignment of long-term and short-term plans in construction, Automation in Construction 132 (2021) 103929. https://doi.org/10.1016/j.autcon.2021.103929.
[18] J. Zheng, M. Fischer, Dynamic prompt-based virtual assistant framework for BIM information search, Automation in Construction 155 (2023) 105067. https://doi.org/10.1016/j.autcon.2023.105067.
[19] Y. Kim, S. Bang, J. Sohn, H. Kim, Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers, Automation in Construction 134 (2022) 104061. https://doi.org/10.1016/j.autcon.2021.104061.
[20] J. Kim, S. Chung, S. Moon, S. Chi, Feasibility Study of a BERT-based Question Answering Chatbot for Information Retrieval from Construction Specifications, in: 2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2022: pp. 0970–0974. https://doi.org/10.1109/IEEM55944.2022.9989625.
[21] Chowdhery et al., PaLM: Scaling Language Modeling with Pathways, (2022). https://doi.org/10.48550/arXiv.2204.02311.
[22] Touvron et al., LLaMA: Open and Efficient Foundation Language Models, (2023). https://doi.org/10.48550/arXiv.2302.13971.
[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, (2019). http://arxiv.org/abs/1810.04805 (accessed July 24, 2023).
[24] F. Guzmán, A. Abdelali, I. Temnikova, H. Sajjad, S. Vogel, How do Humans Evaluate Machine Translation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015: pp. 457–466. https://doi.org/10.18653/v1/W15-3059.
[25] C.-H. Chiang, H. Lee, Can Large Language Models Be an Alternative to Human Evaluations?, (2023). https://doi.org/10.48550/arXiv.2305.01937.
[26] A. Allam, M. Haggag, The Question Answering Systems: A Survey, International Journal of Research and Reviews in Information Sciences 2 (2012) 211–221