Efficient RAG(Retrieval-Augmented Generation) for Construction in Low-Resource Language

Kichang Choi[1], **Seungwon Baek**[1], **Jonewon Ma**[2], **Hongjo Kim**[1]

[1]School of Civil and Environmental Engineering, Yonsei University
[2]Department of Building, Civil, and Environmental Engineering, Concordia University

amki1027@yonsei.ac.kr, baeksw@yonsei.ac.kr, jongwon.ma@concordia.ca, hongjo@yonsei.ac.kr

Abstract -

The Retrieval-Augmented Generation (RAG) framework has gained attention as a fast and cost-effective method for enhancing the performance of large language models (LLMs). However, its performance remains limited in minority languages such as Korean, and this issue is exacerbated in specialized fields like construction. To address these limitations, this study proposes a dataset construction method that allows low-cost fine-tuning of embedding models originally trained on English-based data. By applying this method in the construction domain, we achieved a top-1 document retrieval accuracy of 58.65%, surpassing the performance of a commercial embedding model provided by OpenAI. We further analyzed how improvements in the embedding model influence the overall RAG pipeline and present both a dataset creation approach and an appropriate evaluation strategy for testing RAG's performance. Our findings suggest that this method can significantly enhance technical efficiency by providing a foundation for diverse language users to effectively utilize RAG in the construction domain.

Keywords -

Retrieval-Augmented-Generation; Embedding model; Construction; LLM; Retrieval; Fine-tuning

1 Introduction

Efforts to enhance productivity across various fields, including the construction industry, are continuously advancing. Successful examples include the application of deep learning models for optimizing schedules in large-scale construction projects [1] and detecting pavement cracks [2]. However, applying pretrained models to specialized domains such as construction requires domain adaptation through fine-tuning, which demands significant costs and resources [3, 4]. This is considered a major issue in utilizing deep learning technologies in the construction industry and is one of the challenges that must be addressed.

To mitigate this issue, many attempts have been made to adopt RAG [5]. RAG allows LLMs to generate answers by retrieving relevant information from external databases instead of relying solely on internally stored knowledge. This approach reduces the need for costly LLM fine-tuning. However, the embedding models and LLMs, which are core components of RAG, demonstrate high performance only in English or general-purpose language contexts [6, 7]. Their performance drops significantly in low-resource languages such as Korean or specialized domains like construction. Among these components, embedding models play a critical role in external database generation and retrieval. When domain adaptation is not applied, embedding models exhibit low retrieval accuracy, which negatively impacts the performance of the entire RAG pipeline, as confirmed through our experiments.

To address this challenge, we constructed a fine-tuning dataset specialized for the Korean construction domain and fine-tuned the embedding model to improve RAG performance, as shown in Figure 1. Publicly available construction standards were utilized to extract high-quality sentences, and we proposed a method to generate fine-tuning datasets using LLMs at a low cost. This approach enabled the creation of high-quality fine-tuning datasets for under $10. Using this dataset, we trained a small embedding model that outperformed OpenAI's commercial embedding model, ”text-embedding-3-large” [8]. Our final approach achieved a top-1 document retrieval accuracy of 58.65%. This demonstrates that the proposed method makes LLMs and RAG more accessible and practical for low-resource construction domains.

2 Related Work

The development of LLMs has brought significant advancements across various fields [9]. However, as the size of these models increases, training costs rise substantially [10], and limitations such as hallucination remain prevalent [11]. To address these challenges, RAG has been introduced [12]. RAG enhances LLM responses by retrieving relevant information from external databases rather than relying solely on internal knowledge. This approach is particularly effective for tasks requiring domain-specific knowledge or up-to-date information [13].

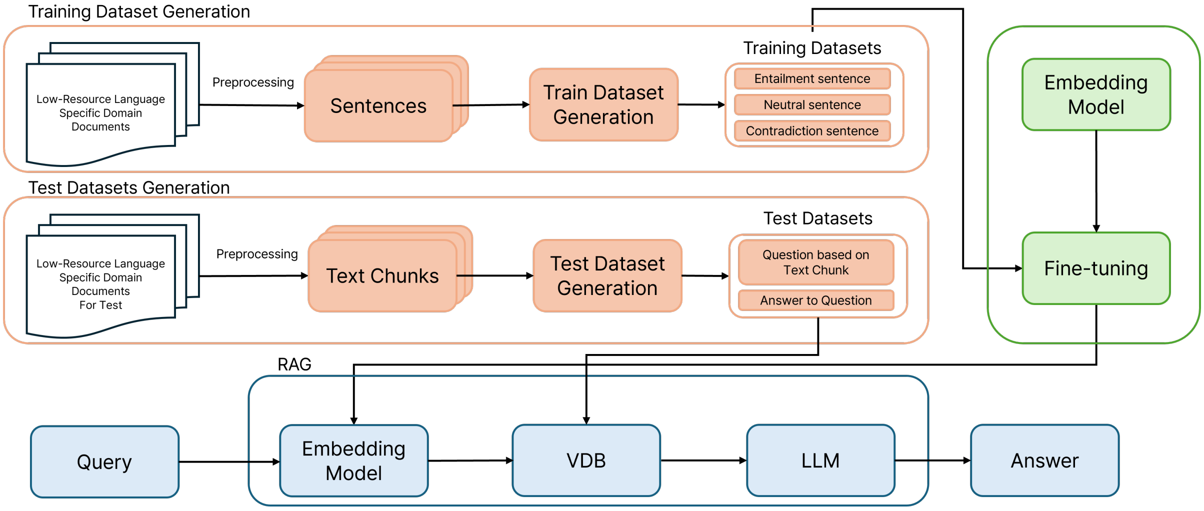


Figure 1. Workflow of embedding model fine-tuning and RAG pipeline construction. This figure illustrates the process of generating a fine-tuning dataset for the embedding model and integrating the fine-tuned model into the RAG pipeline for improved retrieval performance.

The performance of RAG heavily depends on the quality of its Retriever, which retrieves relevant external data. Dense retrieval methods, in particular, rely on embedding models to convert text into high-dimensional vectors and measure semantic similarity between queries and documents [14]. These embedding models are critical to the effectiveness of the RAG pipeline, as their ability to capture semantic meaning, contextual relationships, and syntactic nuances directly impacts retrieval accuracy. Earlier embedding techniques, such as Latent Semantic Analysis (LSA) [15] and Latent Dirichlet Allocation (LDA) [16], have been superseded by methods like Word2Vec [17] and Transformer-based architectures, including BERT [6]. Today, advanced embedding models are available as commercial APIs, such as OpenAI's high-performance models [8].

Recently, LLMs have also been applied in the construction industry. Examples include waste recognition [18], project schedule management [19], and generating captions for site images to monitor progress [20]. However, these applications remain in the early stages and primarily rely on general-purpose language processing capabilities. Discussions about embedding models tailored to construction-specific tasks are rare. To improve RAG performance in low-resource languages and specialized domains, embedding models fine-tuned on domain-specific data are essential.

Fine-tuning pre-trained models for specific tasks has become a common approach for achieving high performance in domain-specific applications [21]. In the construction domain, fine-tuning has been successfully applied to tasks such as automated compliance checking of building codes [22], pavement defect detection [23], and construction management system development [24]. However, constructing high-quality training datasets for fine-tuning in low-resource languages or niche domains is often costly, as it requires expert annotations [25].

Several methods have been proposed to address these challenges and improve the efficiency of dataset generation. One approach involves using LLMs, such as GPT-3 [26], to generate synthetic datasets automatically, which is particularly useful in low-resource language settings [27]. Another method is pseudo-labeling, where unlabeled data is annotated using pre-trained models, effectively expanding the dataset size and diversity [28]. Finally, small curated datasets have been shown to yield high performance with minimal data, as demonstrated by LIMA [29].

In this study, we propose an automatic dataset generation method using LLMs to improve RAG performance in the Korean construction domain. This approach enables the cost-effective creation of high-quality datasets while enhancing the efficiency of domain-specific embedding models.

3 Methodology

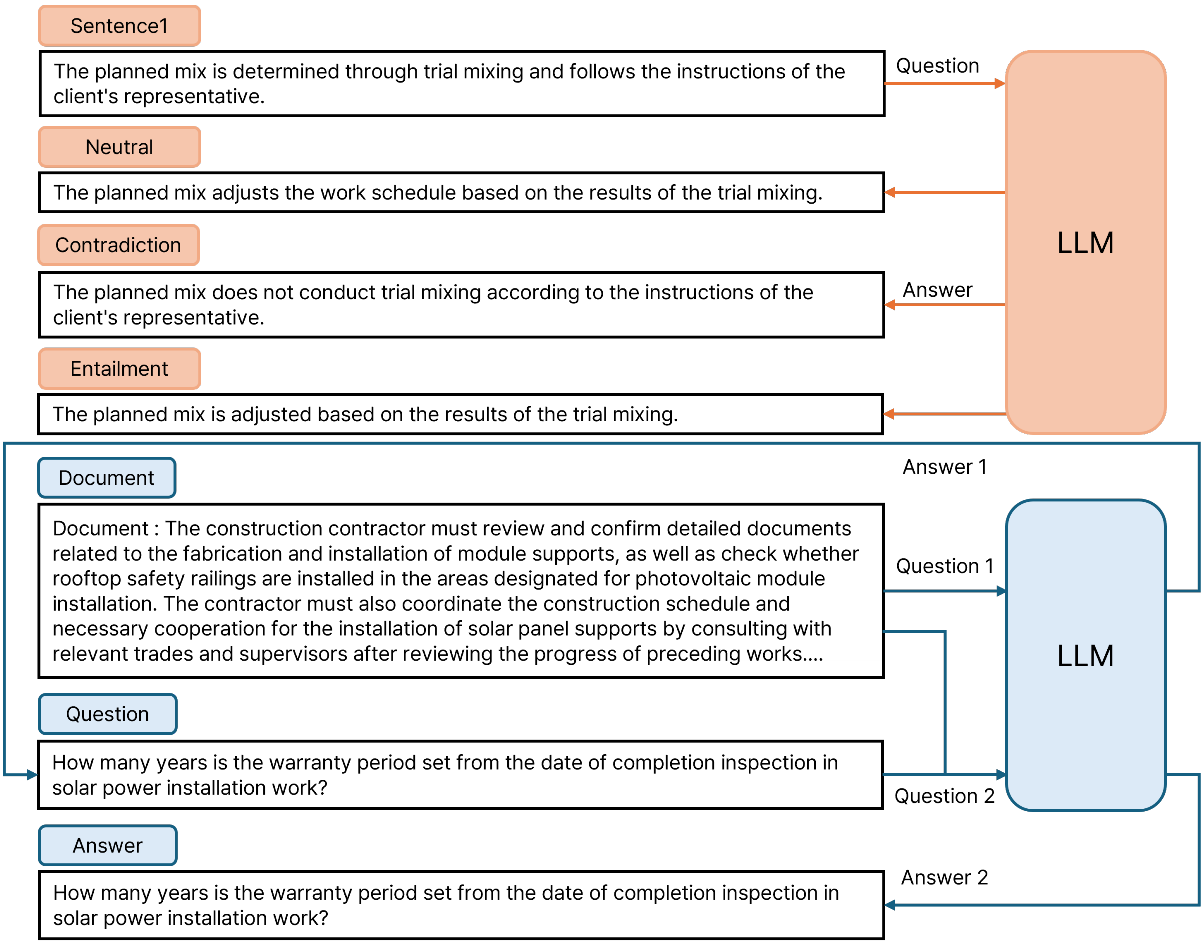
3.1 Dataset Generation

This study proposes a cost-efficient dataset generation method aimed at reducing the costs of data creation while improving the quality and applicability of the datasets to RAG systems. Figure 2 illustrates the proposed dataset generation pipeline.

To construct the training dataset, we first collected high-quality sentences written in Korean and relevant to the con-

Table 1. Training dataset structure for fine-tuning. This table presents the composition of the training dataset, where each original sentence S_1 is paired with three corresponding sentences S_2 representing Entailment, Neutral, and Contradiction relationships.

S_1	S_2 Entailment	S_2 Neutral	S_2 Contradiction
The manufacturer specializes in producing materials specified in the design documents. They evaluate production records, supply performance, and cases of product defects to select a suitable manufacturer.	They select manufacturers with a proven track record of high quality and consistently reliable deliveries.	They seek manufacturers with high production capacity, reliable supply performance, and competitive pricing.	The manufacturer only uses material types specified in the design documents, and no quality issues occur.
The concrete for the wall section is poured so that each part always maintains almost the same height.	The height is consistently maintained when pouring concrete for the wall section.	When pouring concrete for the wall section, they aim to keep the height uniform for each layer.	While pouring concrete for the wall section, some areas were made lower than others.
The allowable deviation of the target air content must be within ±1.5%.	The target air content must have an allowable deviation within ±1.5%.	The allowable deviation of the target air content should be maintained within ±1.5%.	The allowable deviation of the target air content may be ±2.0%.



The planned mix is determined through trial mixing and follows the instructions of the client's representative.

The planned mix adjusts the work schedule based on the results of the trial mixing.

The planned mix does not conduct trial mixing according to the instructions of the client's representative.

The planned mix is adjusted based on the results of the trial mixing.

Document : The construction contractor must review and confirm detailed documents related to the fabrication and installation of module supports, as well as check whether rooftop safety railings are installed in the areas designated for photovoltaic module installation. The contractor must also coordinate the construction schedule and necessary cooperation for the installation of solar panel supports by consulting with relevant trades and supervisors after reviewing the progress of preceding works....

How many years is the warranty period set from the date of completion inspection in solar power installation work?

How many years is the warranty period set from the date of completion inspection in solar power installation work?

Figure 2. Process of training and test dataset generation. This figure visually represents the sequential generation of training and test datasets using an LLM. The training dataset is constructed to fine-tune the embedding model, while the test dataset is generated to evaluate retrieval performance.

struction domain. Publicly available construction specification documents were identified as appropriate sources for this purpose. Many non-English-speaking countries, including Korea, Japan, China, Vietnam, and France, maintain construction specifications in their native languages, which align well with the objectives of this study. Specifically, we extracted 8,499 sentences from the Korean Construction Specifications. Each of these sentences, referred to as S_1, was expanded into three additional sentences, S_2, representing entailment (positive), neutral, and contradiction (negative) relationships. This resulted in a total of 8,499 S_1 - S_2 pairs. The resulting dataset, named Ko Con NLI (Korea Construction Natural Language Inference), was generated using GPT-4o API with a simple data generation prompt. The entire process of generating all 8,499 pairs through the GPT-4o API incurred a total cost of less than $10.

The final datasets are summarized as follows. The training dataset consists of 8,499 S_1 - S_2 pairs, generated from Korean construction documents, with all data used for training to maximize performance. Validation was performed by observing improvements in Retriever performance based on the embedding model's vector database. Examples of the datasets are provided in Table 1.

3.2 Embedding Model Fine-tuning

To fine-tune the embedding model, this study employed Multiple Negative Ranking Loss (MNRL), which is widely recognized for its effectiveness in retrieval-based tasks.

The MNRL loss function is defined as follows:

$$J_{\text{MNRL}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\sigma\left(f_\theta(x_i), f_\theta(y_i)\right)\right)}{\sum_{j=1}^{N} \exp\left(\sigma\left(f_\theta(x_i), f_\theta(y_j)\right)\right)} \tag{1}$$

where N is the batch size, f_θ is the sentence encoder function mapping sentences to the embedding space, and σ is a similarity function, such as cosine similarity. x_i represents the anchor sentence in the i-th pair, y_i is its corresponding positive sentence, and y_j includes all sentences in the batch, including those treated as negative samples.

MNRL optimizes retrieval performance by increasing similarity between positive pairs while minimizing similarity with negative samples, as formulated in Equation 1.

4 Experiment

4.1 Experiment Setup

This study aimed to improve the performance of RAG systems in the Korean construction domain by fine-tuning embedding models and comparing the performance of the fine-tuned models with baseline models. The dataset used for fine-tuning was the **Ko Con NLI dataset**, which was generated by extracting sentences from the Korean Construction Specifications using LLMs, as described earlier. This dataset consists of sentence pairs representing positive, neutral, and negative relationships and was specifically designed to be suitable for embedding model fine-tuning.

The baseline model selected for comparison was OpenAI's ”text-embedding-3-large” model, a widely recognized embedding model known for its excellent performance across multiple languages and domains. This model is commercially available. For fine-tuning, the KLUE-RoBERTa-base model[30], pre-trained on the KLUE dataset, was used. Two experimental setups were considered:

1. Using the pre-trained KLUE-RoBERTa-base model without fine-tuning.

2. Fine-tuning the KLUE-RoBERTa-base model using MNRL.

All experiments were conducted on an NVIDIA GeForce RTX 4090 GPU using CUDA 12.2 and PyTorch. The AdamW optimizer was used with a learning rate set to 2×10^{-5}. Model performance was quantitatively measured by creating a vector database (VDB) and evaluating retrieval accuracy.

The retrieval process works by utilizing a test dataset consisting of pairs of documents and corresponding questions based on each document, ensuring that each document has a correct question. All documents are embedded to create a VDB, and when a query is given, it is embedded as well to retrieve the most similar document. Retrieval accuracy is evaluated by checking whether the retrieved document matches the correct document, and several metrics were applied to verify the accuracy of the proposed method.

4.2 Experiment Results

As shown in Table 2, the KLUE-RoBERTa-base model achieved a low Hit Rate @1 of 12.40% without fine-tuning. However, fine-tuning the model using MNRL significantly improved the Hit Rate @1 to 58.65%, demonstrating the effectiveness of the fine-tuning approach.

Notably, this result surpasses OpenAI's ”text-embedding-3-large” model, which achieved a Hit Rate @1 of 52.67%, by approximately 6 percentage points, despite being a lightweight model trained on a dataset generated with less than $10 in GPT-4o API costs. This demonstrates that even in low-resource language environments, high-quality domain-specific documents can be leveraged to develop cost-effective and efficient embedding models that rival large-scale commercial models.

Table 2. Retrieval performance comparison of fine-tuned and baseline models.

Models	Learning Method	Hit Rate @1 (%)	nDCG@5	MRR@5
KLUE-RoBERTa-base	None	12.40	0.1983	0.1766
KLUE-RoBERTa-base	MNRL	58.65	0.6904	0.6621
OpenAI	None	52.67	0.6784	0.6349

The effectiveness of this approach stems from two key factors: (1) the structured dataset generation process and (2) the use of MNRL. Unlike manually curated datasets, which often include precise similarity scores or fine-grained semantic relationships between sentences, the dataset generated using GPT only provides categorical labels—such as Entailment (positive), Neutral, and Contradiction (negative)—without explicit numerical similarity scores. MNRL is well-suited for both cases, as it effectively optimizes the ranking of positive and negative samples regardless of whether explicit similarity scores are available. This characteristic makes MNRL particularly effective for training on automatically generated datasets, where manually assigned scores are not available.

This synergy between the generated dataset and MNRL loss function proved to be highly effective, enabling the fine-tuned model to achieve superior retrieval accuracy in the Korean construction domain. These findings suggest that even in specialized technical fields and low-resource languages, it is possible to construct lightweight yet high-performance embedding models using carefully curated domain data and appropriate training techniques.

The improvements in retrieval accuracy directly enhance the quality of final RAG-generated responses. Since RAG relies on retrieving relevant documents before generating answers, higher retrieval accuracy ensures that the language model is provided with more contextually relevant information. This reduces the risk of hallucinations, where the model generates inaccurate or misleading responses due to incorrect retrieval. As a result, the fine-tuned embedding model not only improves document retrieval efficiency but also significantly enhances the reliability and factual consistency of the generated responses in knowledge-intensive applications.

5 Conclusion, Limitations & Future Work

This study proposed a methodology to enhance the retrieval and answer generation performance of RAG systems in low-resource language environments, focusing on the Korean construction domain. By utilizing publicly available construction standards and LLMs, a domain-specific dataset was generated at a minimal cost of less than $10. This dataset enabled the fine-tuning of the KLUE-RoBERTa-base model, significantly improving retrieval performance. Specifically, the Hit Rate @1 of the untrained model, which was only 12.40%, increased to 58.65% after fine-tuning with MNRL. Furthermore, the fine-tuned model demonstrated approximately 6% higher retrieval performance compared to OpenAI's "text-embedding-3-large" model.

Despite its promising results, this study has several limitations, which open avenues for future research. While the study focused on improving the retrieval performance of the RAG pipeline, it did not include a demonstration of final answer generation using LLMs. Although it is widely understood that retrieval performance significantly affects final answer quality, further research is needed to explore the extent of improvement in answer generation performance. Additionally, since the LLM in the RAG pipeline was not fine-tuned, there remain many unexplored areas regarding the level of quality that can be achieved in generating answers based on retrieved documents. Without LLM fine-tuning, the study could not determine the precise impact on answer generation quality.

This study relied on publicly available construction standards to generate high-quality datasets. However, in domains where domain-specific documents are scarce or unavailable, applying this methodology may present challenges. Future research should focus on developing generalized data generation techniques that can adapt to domains lacking structured or standardized documents. Furthermore, while the study focused on the Korean construction domain, the applicability of the proposed approach to other languages and technical fields remains untested.

Beyond the Korean construction domain, the proposed approach can be extended to other low-resource languages and specialized technical fields. Many languages, such as Vietnamese, Thai, and Arabic, face similar challenges in adapting retrieval models due to the lack of high-quality domain-specific training data. Leveraging publicly available documents and generative models enables low-cost fine-tuning for these languages. Furthermore, this methodology can be applied to various technical domains, including healthcare, legal systems, and engineering, where precise retrieval and accurate knowledge synthesis are critical. Future research can investigate domain-specific optimizations and cross-lingual adaptability to improve retrieval-based applications in low-resource settings.

Future research should validate this framework's scal-

Table 3. Retrieval Accuracy Metrics

Metric	Description
Hit Rate @1	Measures whether the correct document appears as the top retrieved result.
nDCG@5	Normalized Discounted Cumulative Gain; evaluates ranking quality by considering both relevance and position in the top 5 results.
MRR@5	Mean Reciprocal Rank; calculates the inverse rank of the first correct document within the top 5 retrieved results.

ability across other low-resource languages and specialized fields, including healthcare, legal, and educational domains.

By addressing these limitations, future research can further enhance the applicability, scalability, and robustness of RAG systems in low-resource environments, ultimately expanding their utility across a broader range of domains and languages.

6 Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00457308). ※ MSIT: Ministry of Science and ICT
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.

References

[1] Yuan Yao, Vivian WY Tam, Jun Wang, Khoa N Le, et al. Automated construction scheduling using deep reinforcement learning with valid action sampling. *Automation in Construction*, 166:105622, 2024.

[2] Lele Zheng, Jingjing Xiao, Yinghui Wang, Wangjie Wu, et al. Deep learning-based intelligent detection of pavement distress. *Automation in Construction*, 168:105772, 2024.

[3] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–35, 2023.

[4] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3749–61, 2022.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–74, 2020.

[6] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–27, Huhhot, China, 2021.

[7] Zhilin Yang. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[8] OpenAI. New embedding models and api updates. URL `https://openai.com/index/new-embedding-models-and-api-updates/`. Accessed: 2024-10-31.

[9] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[10] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[11] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of ACL*, pages 1906–19, 2020.

[12] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *Proceedings of the International Conference on Machine Learning*, pages 3929–38. PMLR, 2020.

[13] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, et al. Freshllms: Refreshing large language models with search engine augmentation. In

Findings of the Association for Computational Linguistics: ACL 2024, pages 13697–13720, Bangkok, Thailand, 2024.

[14] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, et al. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–81, 2020.

[15] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, et al. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[17] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–43, Doha, Qatar, 2014.

[18] Ying Sun, Zhaolin Gu, and Sean Bin Yang. Probing vision and language models for construction waste material recognition. *Automation in Construction*, 166:105629, 2024.

[19] Fouad Amer, Yoonhwa Jung, and Mani Golparvar-Fard. Transformer machine learning language model for auto-alignment of long-term and short-term plans in construction. *Automation in Construction*, 132: 103929, 2021.

[20] Yoonhwa Jung, Ikhyun Cho, Shun-Hsiang Hsu, and Mani Golparvar-Fard. Visualsitediary: A detector-free vision-language transformer model for captioning photologs for daily construction reporting and image retrievals. *Automation in Construction*, 165: 105483, 2024.

[21] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, et al. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–37, New Orleans, Louisiana, 2018.

[22] Xiaorui Xue, Jiansong Zhang, and Yunfeng Chen. Question-answering framework for building codes using fine-tuned and distilled pre-trained transformer models. *Automation in Construction*, 168:105730, 2024.

[23] Tianjie Zhang, Donglei Wang, and Yang Lu. A data-centric strategy to improve performance of automatic pavement defects detection. *Automation in Construction*, 160:105334, 2024.

[24] Yunshun Zhong and Sebastian D Goodfellow. Domain-specific language models pre-trained on construction management systems corpora. *Automation in Construction*, 160:105316, 2024.

[25] Fan Yang, Lei Zhang, Sijia Yu, Danil Prokhorov, et al. Feature pyramid and hierarchical boosting network for pavement crack detection. *IEEE Transactions on Intelligent Transportation Systems*, 21(4): 1525–35, 2019.

[26] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–901. Curran Associates, Inc., 2020.

[27] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of ACL*, pages 14409–28, Toronto, Canada, 2023.

[28] Elyas Asadi Shamsabadi, Seyed Mohammad Hassan Erfani, Chang Xu, et al. Efficient semi-supervised surface crack segmentation with small datasets based on consistency regularisation and pseudo-labelling. *Automation in Construction*, 158:105181, 2024.

[29] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, et al. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–21. Curran Associates, Inc., 2023.

[30] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, et al. Klue: Korean language understanding evaluation, 2021.