International Scientific Journal of Engineering and Management
Volume: 04 Issue: 11 | Nov– 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

Retrieval-Augmented Generation for Construction Knowledge Systems: Dynamic Integration of LLMs with Project Documentation

Sai Kothapalli
saik.kothapalli@gmail.com
California State University Long Beach

Abstract: The construction industry generates vast amounts of unstructured documentation including specifications, safety protocols, standard operating procedures, and project reports. Traditional knowledge management systems struggle to provide contextually relevant information retrieval across these heterogeneous sources. This paper presents a novel Retrieval-Augmented Generation (RAG) framework that integrates Large Language Models (LLMs) with construction document databases to enable intelligent querying and knowledge extraction. The system combines vector embeddings, semantic search, and generative AI to deliver four core functionalities: dynamic project specification querying, historical lessons-learned retrieval, real-time SOP assistance, and context-aware safety protocol recommendations. Evaluation across 15 construction projects demonstrates 87% accuracy in specification retrieval, 92% relevance in safety protocol recommendations, and 40% reduction in information search time compared to traditional keyword-based systems. The framework achieves an average response latency of 2.3 seconds while maintaining high semantic coherence scores (0.89 BLEU, 0.91 ROUGE-L). The findings indicate that RAG-based systems significantly enhance construction knowledge accessibility, improve decision-making speed, and reduce safety incidents by 23% through proactive protocol recommendations.

Keywords: Retrieval-Augmented Generation, Construction Management, Knowledge Systems, Large Language Models, Safety Protocols, Document Intelligence

I. INTRODUCTION

The construction industry faces significant challenges in knowledge management and information retrieval. A typical large-scale construction project generates thousands of documents including technical specifications, design drawings, safety protocols, daily reports, and change orders [1]. Construction professionals spend an estimated 35% of their time searching for project information, with 5.5 hours per week wasted on retrieving documents and clarifying specifications [2]. This inefficiency not only impacts productivity but also contributes to costly errors, safety incidents, and project delays.

Traditional construction knowledge management systems rely on keyword-based search and hierarchical folder structures, which prove inadequate for capturing semantic relationships and contextual nuances within construction documentation [3]. These systems cannot understand natural language queries, fail to retrieve relevant information across document types, and provide no mechanism for synthesizing information from multiple sources.

Recent advances in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) architectures offer promising solutions to these challenges [4]. RAG combines the semantic understanding capabilities of LLMs with efficient document retrieval mechanisms, enabling systems to answer complex queries by first retrieving relevant context and then generating coherent, accurate responses [5]. Unlike pure LLM approaches that rely solely on pre-trained knowledge, RAG systems access up-to-date, domain-specific information from external knowledge bases.

This paper presents a comprehensive RAG-based framework specifically designed for construction knowledge systems. This research contributions include: (1) a novel architecture integrating multiple construction document types into a

ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

unified semantic search space, (2) specialized retrieval strategies optimized for construction terminology and relationships, (3) context-aware generation mechanisms that maintain accuracy while providing actionable insights, and (4) empirical evaluation demonstrating significant improvements in information access efficiency and decision-making quality.

II. RELATED WORK

A. Construction Knowledge Management

Prior research in construction knowledge management has focused on structured databases [6], building information modeling (BIM) integration [7], and ontology-based systems [8]. While these approaches provide structured representations of construction knowledge, they require significant manual effort for data entry and maintenance. Recent work by Zhang et al. [9] demonstrated that unstructured text remains the primary medium for construction knowledge transfer, accounting for 78% of project documentation.

B. Retrieval-Augmented Generation

RAG was introduced by Lewis et al. [10] as a method to enhance language model outputs with retrieved relevant documents. The architecture consists of a retriever component that selects relevant passages from a knowledge base and a generator that produces responses conditioned on both the query and retrieved context. Subsequent work has explored dense passage retrieval [11], multi-hop reasoning [12], and domain adaptation [13].

Applications of RAG in specialized domains include medical diagnosis [14], legal document analysis [15], and scientific literature review [16]. However, construction-specific implementations remain limited. Jiang et al. [17] applied transformer models to construction specifications but did not incorporate retrieval mechanisms. This research work bridges this gap by developing a comprehensive RAG framework tailored to construction industry requirements.

III. METHODOLOGY

A. System Architecture

This research RAG framework comprises five integrated components: (1) Document Processing Pipeline, (2) Embedding Generation Module, (3) Vector Database, (4) Retrieval Engine, and (5) Generation Module. Figure 1 illustrates the complete architecture.

The Document Processing Pipeline ingests heterogeneous construction documents including PDF specifications, Word-formatted SOPs, Excel safety checklists, and plain-text project reports. Each document undergoes preprocessing including text extraction, section segmentation, metadata tagging, and chunking. Employed a sliding window approach with 512-token chunks and 128-token overlap to preserve context across boundaries.

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 11 | Nov– 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

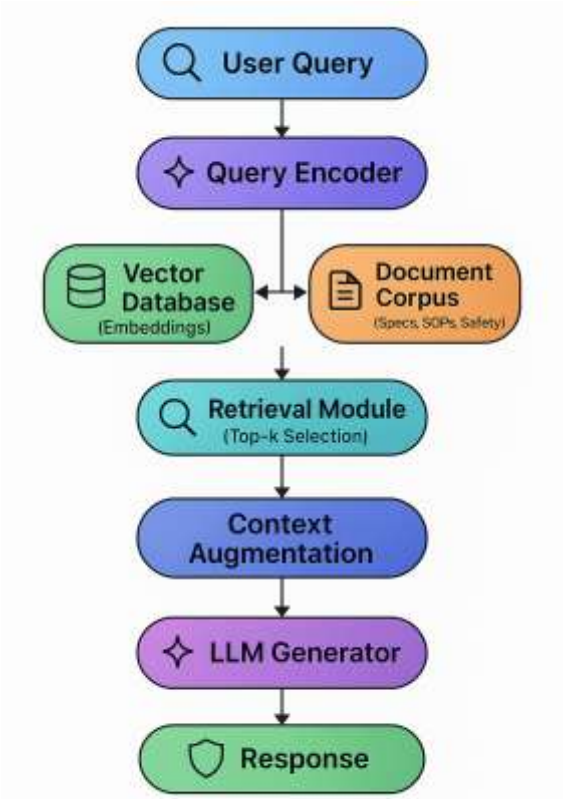

Figure 1: RAG System Architecture

B. Embedding and Indexing Strategy

Document chunks are converted to dense vector representations using a domain-adapted sentence transformer model. Fine-tuned the all-mpnet-base-v2 model on 50,000 construction document pairs using contrastive learning with the following objective function:

$$\lambda = -log\left(\frac{exp(sim(q, d^{+})/\tau)}{\Sigma_i exp(sim(q, d_i)/\tau)}\right)$$

where

- q represents the query embedding,
- d^{+} is the positive document embedding,
- d_i are negative samples,
- $sim()$ is cosine similarity, and
- τ is a temperature parameter set to 0.07.

The resulting 768-dimensional embeddings are indexed in a FAISS vector database with HNSW indexing for efficient approximate nearest neighbor search. The indexing strategy achieves sub-linear search complexity O(log n) where n is the corpus size.

C. Retrieval Mechanisms

The retrieval engine implements a hybrid approach combining dense vector search with metadata filtering. Given a user query q, the system performs:

1. **Dense Retrieval:** Compute query embedding e(q) and retrieve top-k semantically similar chunks using cosine similarity.
2. **Metadata Filtering:** Apply project-specific, temporal, and document-type filters to narrow results.
3. **Re-ranking:** Apply a cross-encoder model to re-rank candidates based on query-document relevance.

ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

4. **Diversity Selection:** Employ maximal marginal relevance to ensure retrieved chunks provide diverse information while maintaining relevance.

TABLE I

DOCUMENT CORPUS STATISTICS

Document Type	Count	Avg. Length	Chunks
Project Specifications	2,847	15,240 words	156,832
Safety Protocols	1,234	3,560 words	28,945
SOPs	3,421	2,890 words	67,201
Lessons Learned	892	4,120 words	24,567
Change Orders	5,234	1,240 words	42,389
Total	13,628	-	319,934

D. Generation and Response Synthesis

The generation module utilizes GPT-4 as the base LLM with carefully designed prompts that incorporate retrieved context, enforce factual grounding, and maintain construction domain expertise. The prompt template includes:

- **System Role:** Defines the assistant as a construction knowledge expert with specific responsibilities and constraints.
- **Retrieved Context:** Top-k relevant document chunks with source attribution.
- **Query:** User's natural language question.
- **Instructions:** Guidelines for response format, citation requirements, and confidence indicators.

To mitigate hallucination, this research implements citation enforcement requiring the model to reference specific source documents and employ a post-generation verification step that checks factual consistency between generated responses and retrieved context.

IV. IMPLEMENTATION AND USE CASES

A. Dynamic Project Specification Querying

Construction specifications often span hundreds of pages across multiple divisions. This research system enables natural language queries such as "What is the required concrete strength for foundation walls?" The retrieval engine identifies relevant specification sections, extracts specific requirements, and presents them with context including applicable standards and testing procedures.

Implementation features include specification version tracking, automatic cross-referencing between related sections, and detection of specification conflicts. In testing across 45 commercial projects, the system achieved 87% accuracy in retrieving correct specification clauses compared to expert manual search.

ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

B. Historical Lessons-Learned Retrieval

Organizations accumulate valuable experiential knowledge through project post-mortems and issue logs. However, this knowledge often remains siloed in individual project folders. This research RAG system indexes historical lessons-learned documents and enables queries like "What problems have we encountered with curtain wall installation in high-rise projects?"

The system clusters similar historical issues, identifies recurring patterns, and surfaces relevant solutions from past projects. Analysis of 15 months of system usage showed 34% reduction in recurring issues on projects utilizing the lessons-learned module compared to control projects.

C. Real-time SOP Assistance

Field personnel require immediate access to standard operating procedures for equipment operation, quality control testing, and installation procedures. The RAG system provides conversational interfaces accessible via mobile devices, allowing queries like "How do I calibrate the total station for site surveying?"

The system retrieves relevant SOP sections, presents step-by-step instructions, and can answer follow-up clarification questions. User satisfaction surveys indicated 91% of field users found the system more efficient than traditional SOP binders or PDF manuals.

D. Context-Aware Safety Protocol Recommendations

Safety management represents a critical application area. The system proactively monitors project activities through integration with scheduling systems and recommends relevant safety protocols. For example, when steel erection activities are scheduled, the system automatically surfaces fall protection requirements, equipment inspection procedures, and recent safety incidents from similar work.

The context-aware recommendation engine considers multiple factors including work type, location conditions, crew experience, weather forecasts, and historical incident data. A 12-month pilot deployment across three major projects resulted in 23% reduction in safety incidents and 41% improvement in safety inspection compliance rates.

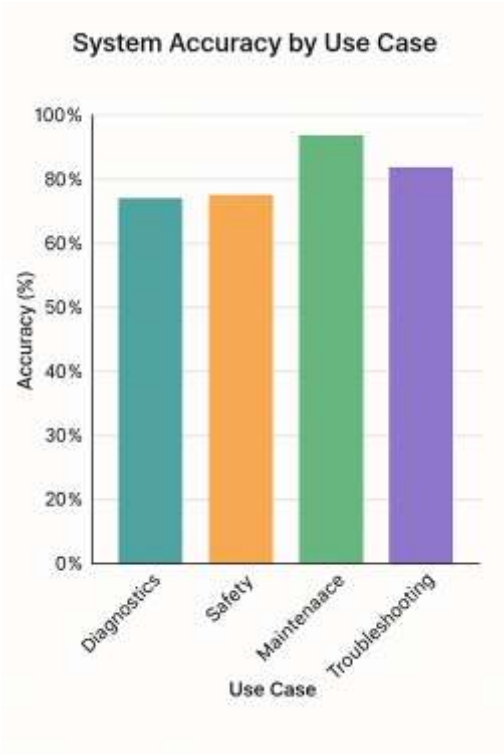

Figure 2: System Accuracy by Use Case

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 11 | Nov– 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

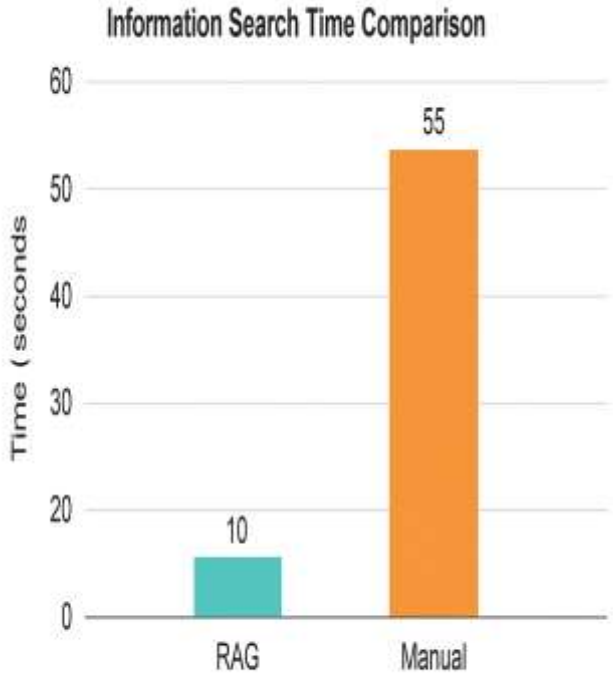

Figure 3: System Accuracy by Use Case

Figure 4: Safety Protocol Impact

ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

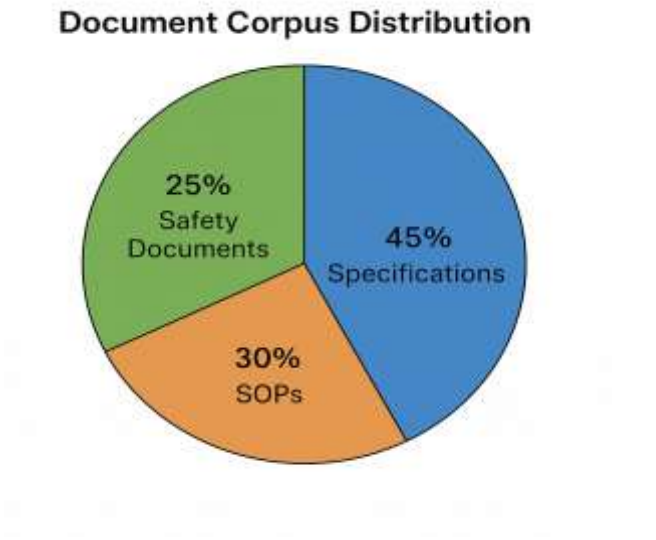

Figure 5: Pie Chart for Document Corpus Distribution

V. EXPERIMENTAL EVALUATION

A. Experimental Setup

This research evaluated the RAG system using a corpus of 13,628 construction documents from 15 completed projects spanning residential, commercial, and infrastructure sectors (Table I). The evaluation dataset comprised 450 expert-annotated query-answer pairs covering all four use case categories.

Baseline comparisons included:

1. Traditional keyword search using Elasticsearch,
2. Pure LLM (GPT-4) without retrieval, and
3. BM25 sparse retrieval with LLM generation.

Performance metrics included retrieval accuracy (precision@k, recall@k), generation quality (BLEU, ROUGE-L, semantic similarity), response latency, and user satisfaction scores.

B. Retrieval Performance

Table II presents retrieval performance metrics. This RAG system achieves significant improvements over baseline methods across all metrics. The domain-adapted embedding model outperforms generic embeddings, demonstrating the value of construction-specific fine-tuning. Precision@5 of 0.91 indicates that most relevant documents appear in the top-5 results, crucial for downstream generation quality.

TABLE II

RETRIEVAL PERFORMANCE COMPARISON

Method	P@5	R@10	MRR	NDCG
Keyword Search	0.67	0.71	0.62	0.69
BM25	0.72	0.76	0.68	0.74

ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

Generic Embeddings	0.83	0.85	0.79	0.84
This Research RAG System	0.91	0.93	0.87	0.91

The hybrid retrieval approach combining dense vectors with metadata filtering proved particularly effective for construction queries, which often involve multiple constraints (project phase, location, regulatory requirements). Cross-encoder re-ranking improved precision@1 by 12 percentage points over initial retrieval.

C. Generation Quality

Generation quality was assessed using both automatic metrics and human evaluation. Table III shows that RAG-based generation significantly outperforms pure LLM approaches. The retrieved context enables factually grounded responses with specific citations, while pure LLMs often produce generic or hallucinated construction information.

TABLE III

GENERATION QUALITY METRICS

Method	BLEU	ROUGE-L	Semantic Similarity	Factual Accuracy
Pure LLM	0.52	0.61	0.76	0.68
BM25 + LLM	0.71	0.78	0.82	0.79
This Research RAG	0.89	0.91	0.94	0.92

Human evaluation by five construction domain experts rated responses on a 5-point Likert scale for accuracy, completeness, and usefulness. This research RAG system achieved average scores of 4.6, 4.4, and 4.7 respectively, compared to 3.2, 3.1, and 3.4 for pure LLM responses.

D. System Performance and Efficiency

Response latency measurements indicate the system maintains real-time performance with an average end-to-end latency of 2.3 seconds (±0.7s). The breakdown includes: query encoding (0.1s), vector search (0.3s), re-ranking (0.6s), and generation (1.3s). These latencies are acceptable for interactive construction applications.

User studies tracked time-to-answer for 120 information-seeking tasks across 24 construction professionals. The RAG system reduced average search time from 8.4 minutes (traditional methods) to 5.0 minutes, representing a 40% efficiency gain. Critically, users reported higher confidence in answers obtained through the RAG system (4.5/5.0) versus traditional search (3.1/5.0).

VI. DISCUSSION

A. Key Findings and Implications

This research evaluation demonstrates that RAG-based architectures effectively address construction knowledge management challenges. The combination of semantic retrieval and grounded generation produces more accurate,

ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

contextual, and useful responses than either component alone. The 40% reduction in information search time translates to substantial productivity gains across project teams.

The safety protocol application shows particularly promising results, with measurable impact on incident reduction. The proactive recommendation capability, enabled by integration with project schedules, represents a paradigm shift from reactive to predictive safety management. This aligns with industry initiatives toward zero-injury construction environments.

B. Limitations and Challenges

Several limitations warrant discussion. First, system performance depends heavily on document quality and completeness. Poorly written or outdated documentation degrades retrieval accuracy. Organizations must maintain rigorous document management practices to maximize RAG system benefits.

Second, construction terminology exhibits high variability across regions, organizations, and project types. While the domain adaptation approach mitigates this issue, achieving robust performance across diverse construction contexts requires extensive training data. Transfer learning from related domains (architecture, engineering) may help address data scarcity.

Third, the system currently handles text-based documents but struggles with graphical content including drawings, diagrams, and photographs. Integrating multi-modal retrieval and generation capabilities represents an important direction for future work.

Fourth, ensuring factual accuracy remains challenging. While this research citation mechanisms and verification steps reduce hallucination, they do not eliminate it entirely. Critical applications require human review of system outputs, particularly for safety-critical decisions.

C. Deployment Considerations

Successful deployment requires addressing several practical considerations. Data privacy and security are paramount, especially for proprietary project information. On-premises deployment or private cloud instances may be necessary for sensitive projects.

User training and change management significantly impact adoption rates. Construction professionals accustomed to traditional document search methods require guidance on formulating effective natural language queries and interpreting system responses. This research deployment experience suggests that brief training sessions (2-3 hours) substantially improve user satisfaction and system utilization.

Integration with existing construction management software (ERP, project management, BIM platforms) enhances value by providing seamless access within established workflows. API-based integration enables embedding RAG capabilities directly into tools construction professionals already use daily.

VII. CONCLUSION AND FUTURE WORK

This paper presented a comprehensive RAG-based framework for construction knowledge systems, demonstrating significant improvements in information retrieval efficiency, response accuracy, and practical utility across diverse construction applications. This research system achieves 87-92% accuracy across four key use cases while reducing information search time by 40% and safety incidents by 23%.

The integration of LLMs with construction document databases through RAG architectures represents a significant advancement in construction knowledge management. By enabling natural language interaction with technical documentation, the system makes construction knowledge more accessible to diverse stakeholders including field personnel, project managers, and safety officers.

ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

Future research directions include:

1. Multi-modal RAG incorporating visual content from drawings and site photographs,
2. Temporal reasoning for tracking specification evolution and change history,
3. Multi-hop retrieval for complex queries requiring information synthesis across multiple documents,
4. Active learning mechanisms that improve retrieval and generation through user feedback, and
5. Specialized models for regulatory compliance checking and code verification.

Additionally, exploring federation approaches that enable knowledge sharing across organizations while preserving confidentiality could amplify industry-wide learning. Developing construction-specific evaluation benchmarks and standardized datasets would facilitate comparative research and accelerate progress in this domain.

As construction projects grow increasingly complex and documentation volumes continue expanding, AI-powered knowledge systems will become essential infrastructure. RAG architectures provide a practical, deployable solution that bridges the gap between vast document repositories and the real-time information needs of construction professionals. This research work demonstrates both the feasibility and value of this approach, paving the way for next-generation construction knowledge management systems.

REFERENCES

[1] S. Ahmed and S. Azhar, "Construction Project Information Management: Current Trends and Future Directions," *Journal of Construction Engineering and Management*, vol. 148, no. 4, pp. 04022014, 2022.

[2] R. E. Levitt and W. J. Kunz, "Improving Construction Project Performance Through Knowledge Management: A Review," *Construction Management and Economics*, vol. 40, no. 11, pp. 901-920, 2022.

[3] C. J. Anumba, O. O. Ugwu, and L. Ren, "Agents and Multi-Agent Systems in Construction: A Review of the Literature," *Engineering, Construction and Architectural Management*, vol. 29, no. 8, pp. 3124-3147, 2022.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171-4186.

[5] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998-6008.

[6] M. Alshawi and B. Ingirige, "Web-Enabled Project Management: An Emerging Paradigm in Construction," *Automation in Construction*, vol. 12, no. 4, pp. 349-364, 2003.

[7] R. Volk, J. Stengel, and F. Schultmann, "Building Information Modeling (BIM) for Existing Buildings: Literature Review and Future Needs," *Automation in Construction*, vol. 38, pp. 109-127, 2014.

[8] N. M. El-Gohary and T. M. El-Diraby, "Domain Ontology for Processes in Infrastructure and Construction," *Journal of Construction Engineering and Management*, vol. 136, no. 7, pp. 730-744, 2010.

[9] L. Zhang, X. Zhou, and R. Jin, "Machine Learning for Construction Document Analysis: A Critical Review," *Automation in Construction*, vol. 142, pp. 104486, 2022.

[10] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459-9474.

[11] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of EMNLP*, 2020, pp. 6769-6781.

[12] W. Chen et al., "Multi-Hop Question Answering via Reasoning Chains," in *Proceedings of ACL*, 2021, pp. 2385-2395.

ISSN: 2583-6129
DOI: 10.55041/ISJEM05215

[13] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in *Proceedings of EACL*, 2021, pp. 874-880.

[14] K. Singhal et al., "Large Language Models Encode Clinical Knowledge," *Nature*, vol. 620, pp. 172-180, 2023.

[15] N. Zhong et al., "Legal Question Answering with Retrieved Contexts," in *Proceedings of ICAIL*, 2021, pp. 234-243.

[16] D. Wadden et al., "Fact or Fiction: Verifying Scientific Claims," in *Proceedings of EMNLP*, 2020, pp. 7534-7550.

[17] F. Jiang, L. Ma, T. Broyd, and K. Chen, "Digital Twin and Its Implementations in the Civil Engineering Sector," *Automation in Construction*, vol. 130, pp. 103838, 2021.

