Generative project question answering system: triangulating three approaches for project authoring

Yunshun Zhong, (yunshun.zhong@mail.utoronto.ca)
Faculty of Civil and Mineral Engineering, University of Toronto, Canada

Tamer El-Diraby, (tamer.diraby@utoronto.ca)
Faculty of Civil and Mineral Engineering, University of Toronto, Canada

Keywords: Generative AI, Large language models, Text mining, Project authoring, Knowledge management

Abstract

In the planning and design of new projects, team members often rely on existing technical documents to make informed decisions. A question-answering system can significantly aid in this process using Large Language Models (LLMs) like GPT-4 for specialized fields such as Architecture, Engineering, and Construction (AEC). However, they face challenges such as inaccurate information retrieval, limited access to domain-specific data, inadequate prompt customization, and hallucinations. This paper addresses these issues by developing a generative question-answering system for technical documents, using retrieval-augmented generation (RAG) and prompt engineering. The system leverages a vectorized database with domain-specific embeddings and employs semantic similarity for efficient passage retrieval. By integrating retrieved domain knowledge into the LLM through prompts, the system aims to enhance the accuracy and relevance of generated information. The expert evaluation compared the developed advisor to traditional search methods including Google and document databases, showing superior performance in all LLM-Assisted evaluation metrics, especially search efficiency.

1. Introduction and Literature Review

In the planning and design of new projects, team members often rely on existing technical documents to make informed decisions. However, retrieving information from these documents can be challenging due to their complex and unstructured nature. Technical documents are divided into various sections, each serving a different purpose—from establishing context and reviewing existing scholarship to outlining challenges and proposing resolutions (B. Zhong et al., 2020). This complexity makes it difficult for stakeholders and technicians to quickly and efficiently find the information they need (Lin et al., 2012). Traditional information retrieval (IR) techniques have been employed to address these difficulties. For instance, Zhang developed a deep neural network-based method using transfer learning strategies to support automated compliance checking, significantly improving the precision of retrieving relevant documents and data (Zhang & El-Gohary, 2021). However, while these methods excel at extracting pertinent information, they often struggle to synthesize this information into cohesive answers and handle queries without direct matches in the data.

To overcome these shortcomings, question-answering (QA) systems based on pre-trained transformer models like BERT have emerged as a significant improvement. For example, Wang's application of a BERT-based QA system for extracting information from building

models demonstrates progress in this area (N. Wang et al., 2022). Despite these advancements, QA systems still face challenges in dynamically adjusting responses based on user feedback, integrating external knowledge effectively, and synthesizing large volumes of information into concise answers.

In the era of Industry 4.0, researchers and companies are increasingly developing AI assistants or QA systems using large language models (LLMs) to support various tasks (N. Wang et al., 2022), such as OpenAI's GPT-4, Google's Gemini, and Meta's LLAMA. LLMs offer advanced generative capabilities, broader contextual understanding, versatility across a range of topics, and adaptability to different question types or languages without extensive task-specific training (Ouyang et al., 2022). These capabilities make them particularly suited to navigating the dense and complex landscape of technical documents essential in project authoring. However, integrating LLMs into specialized domains such as Architecture, Engineering, and Construction (AEC) presents unique challenges, including inaccurate information retrieval, limited access to domain-specific data, inadequate customization of prompts (L. Wang et al., 2024), and hallucinations (Martino et al., 2023). Consequently, identifying relevant answers directly becomes a difficult task for LLMs with frozen weights during inferencing.

To enhance the domain-specific understanding of LLMs, several methods can be employed, including fine-tuning (H. Wang et al., 2023) and prompt engineering (Zuccon & Koopman, 2023). However, fine-tuning an LLM, such as GPT-4, demands extensive GPU memory resources, making it an impractical solution for many applications. Therefore, this research adopts the prompt engineering method as a viable alternative. Prompt engineering involves the strategic formulation of instructions that set the context for the LLM's operation, highlighting the critical information and specifying the desired format and content of its output (White, Fu, et al., 2023). This method not only directs the LLM towards focusing on pertinent information but also tailors its responses to meet the specific needs of analyzing agency's technical documents.

This research aims to develop a generative project QA system that integrates a curated technical document database (TDD) with advanced LLMs like GPT-4. By leveraging a vectorized database with domain-specific embeddings, retrieval-augmented generation (RAG), and prompt engineering, the system seeks to enhance the accuracy and relevance of generated information.

The structure of this paper is organized as follows: Section 1 provides a comprehensive review of the existing literature pertinent to developing an LLM-based generative project QA system. Section 2 details the development procedures, including the creation of a vectorized database, implementation of IR methods, and question-answering with domain-specific prompt engineering (Giray, 2023). Section 3 presents the methodology for evaluating the the system. Finally, Section 4 discusses the advantages, contributions, conclusion, potential limitations, and future research directions of this study.

2. Development Procedure

An overview of the proposed methodology for developing an agency-specific project authoring advisor is shown in Figure 1. The core architecture, highlighted in orange, comprises the following five main modules:

1) Data preprocessing: Technical documents are converted from PDF to plain text, cleaned of non-English text and irrelevant content, and segmented into sentences to prepare high-quality data for the next step.

2) Vectorized database creation with domain-specific word embedding: The cleaned text is tokenized, structured into chunks, and converted into dense vectors that capture semantic meanings using domain-specific embeddings. These vectors are stored

in a database, optimized for the model's token capacity to ensure efficient retrieval and processing.
3) Information retrieval: Top-ranked passages are identified through vectorized embeddings and similarity comparison measures, selecting the most contextually pertinent chunks for answering queries.
4) Prompt Engineering: The LLM is tailored with domain-specific prompts to accurately interpret user queries and generate contextually relevant answers from the selected passages.
5) Question Answering: input retrieved data to LLM to provide comprehensive responses.

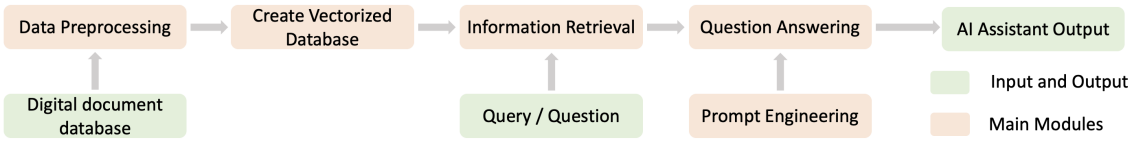

Figure 1: Overview of Methodology

2.1. Database Description and Data preprocessing

To develop an effective agency-specific project authoring advisor for the AEC domain, a robust and well-prepared dataset is essential. This section outlines the database description and preprocessing steps undertaken to ensure data quality and relevance.

2.1.1. TRCA's Technical Documents Database Description

The TRCA Technical Documents Database is a comprehensive repository that supports environmental protection and sustainable development. It includes records and analyses related to erosion control, shoreline maintenance, environmental assessments, and project planning within the TRCA jurisdiction. Key projects documented include initiatives at Humber Bay Park East and German Mills Settlers Park. The database contains detailed project briefs, work plans, design briefs, conceptual alternative reports, inspection records, and erosion hazard assessments. These documents provide insights into planning, decision-making, and the effectiveness of erosion control structures. Additionally, environmental and geomorphic reports offer empirical data on ecological impacts, including archaeological assessments, aquatic inventories, tree surveys, and species-at-risk screenings.

2.1.2. Data Cleaning and Preprocessing

Technical documents from the agency will be used in this study. A detailed description of the database is shown in Section 2.1.1. The database will be cleaned in the following steps: 1) Convert PDF to Text: Use Adobe Automation to convert PDFs to plain text. 2) Remove Website Links: Eliminate website links to focus on substantive content (Patil & Pawar, 2018). 3) Retain English Text: Use regular expressions to retain only English text, discarding unrecognizable characters and non-English text (Li et al., 2008). This serves to discard unrecognizable characters and any non-English text that could potentially disrupt the pre-training process. 4) Sentence Segmentation: Divide paragraphs into distinct sentences, removing paragraphs without terminal punctuation to exclude remnants of formulas and tables. 5) Filter Short Sentences: Remove sentences that are too short to ensure the quality and relevance of the data. 6) Remove References: Filter out references that may not contain useful context, preparing datasets both with and without references for pre-training. 7) Remove Duplicates: Eliminate duplicate sentences to enhance the uniqueness and diversity of the dataset. These preprocessing steps ensure

that the data is clean, structured, and ready for creating a vectorized database suitable for information retrieval and question-answering tasks.

2.2. Vectorized Database Creation With Domain-Specific Embeddings

Following the data cleaning phase, creating a vectorized database encompasses a meticulously designed sequence of procedures as shown in Figure 2. Initially, the cleaned text undergoes tokenization and is structured into chunks. Next, each text chunk is transformed into dense vectors using domain-specific word embeddings for Construction Management Systems (CMS) domain (Y. Zhong and Goodfellow, 2024). These embeddings are trained on a comprehensive CMS domain corpus, which includes academic publications such as journal papers, conference papers, articles, whitepapers, and books, totaling 5.7 million words and 7.7 million tokens. This vector conversion captures nuanced domain semantics using pretrained transformer models, enhancing the database's depth and context awareness. The vectorized chunks are then systematically stored in a structured database, which forms the backbone of the system's retrieval capabilities. This setup facilitates quick and precise access to information, improving the system's ability to provide contextually relevant responses based on similarity measures and query relevance.

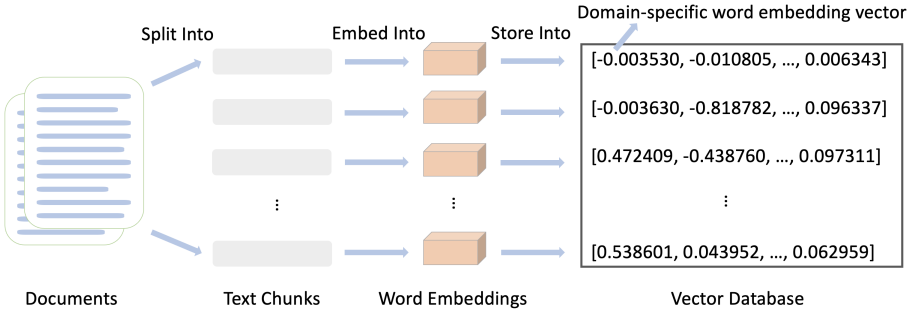

Figure 2: Vector Database Creation Procedure

2.3. Information Retrieval

In the proposed agency-specific project authoring advisor, the retrieval of relevant passages is a critical component. This process begins by transforming the user's question or query into a domain-specific embedding. This embedding is then utilized for cosine similarity calculations with vectorized database, serving as a basis for comparison with pre-processed embeddings of document chunks in the database.

Semantic cosine similarity (Rahutomo et al., 2012), a measure of the cosine of the angle between two word embedding vectors in a multidimensional space, is used to quantify the similarity between the question's embedding and that of each document chunk. Passages that exhibit higher semantic cosine similarity scores are deemed more relevant to the question, and thus, are prioritized. These relevant passages are then ranked and arranged as candidates based on the magnitude of their similarity scores.

To streamline the subsequent processing, a predetermined number of top-ranked passages (denoted as Top-D) are selected. This selection based on similarity scores considers the contextual completeness and relevance to ensure the extracted passages are both pertinent and informative. These Top-D passages are then input into the next phase of the system, which involves prompt engineering. In this phase, the selected passages are utilized with context-rich prompts that facilitate the generation of accurate and detailed answers by the LLMs.

This passage retrieval methodology ensures that the question-answering system efficiently narrows down the vast corpus to the most relevant sections, thereby enhancing the overall precision and effectiveness of the response generation process in the agency's technical documentation context.

2.4. Prompt Template Design

As part of our initiative to enhance LLMs for utilizing domain-specific knowledge from agency's technical documents, we introduce a meticulously crafted prompt template for each of the Top-D extracted passages. The template is strategically divided into five key sections: goal, database, requirement, and example as shown in the list below and Figure 3. Each segment is specifically designed to augment the LLM's capability in processing and generating targeted outputs. This prompt template incorporates several design patterns as outlined by White (White, Hays, et al., 2023), including the Persona Pattern, Format Template Pattern, Question Refinement Pattern, Cognitive Verifier Pattern, and Domain-Specific Language Creation Pattern. Additionally, the template leverages advanced techniques such as the Chain of Thought (CoT) method (L. Wang et al., 2023) and Few-shot Learning (Logan IV et al., 2021) to enhance its effectiveness. These elements are strategically integrated to optimize the template's performance across various conversational scenarios.

1. **Goal (System Message)**: This portion explicitly articulates the LLM's task to accurately address the specific query in question. It ensures the system's objectives are perfectly attuned to the information requisites of the agency.

2. **Database (Knowledge Embedding)**: Here, essential information from the top-D passages in the technical document repository is integrated. A few-shot prompting technique (Reynolds and McDonell, 2021) acquaints the LLM with the intricate semantics inherent in our domain-specific materials, enhancing its understanding and interaction with the technical content.

3. **Requirement (Format and Process Specification)**: Expectations for the response structure are set in this segment. It incorporates a CoT directive, prompting the LLM to methodically break down the query and proceed in a step-by-step fashion, thus ensuring the output is systematically reasoned and formatted as per our guidelines.

4. **Example (Reasoning Model)**: To foster an approach akin to human analytical progression, this section provides examples to LLM. These guide the LLM's reasoning through CoT, demonstrating a sequential deconstruction of similar problems, thereby instructing the LLM in a structured problem-solving approach (L. Wang et al., 2023).

2.5. Question Answering using MapReduce

In this section, we use MapReduce, a distributed computing framework, to process and synthesize answers from multiple sources, including top-ranked passages and internet search results. This approach manages long passages and contexts that exceed LLMs' input capacity by distributing the computational load across multiple LLMs, enabling scalable handling of extensive datasets and complex tasks. The procedure includes five steps as shown in Figure 4.

Figure 3: Illustration of Prompt

1. Passage Injection and Initial Answer Generation: Each top-ranked passage, identified through information retrieval, is fed into an LLM using a carefully designed prompt template. This template contextualizes the passage and guides the LLM in generating a preliminary answer, leveraging the LLM's capability to interpret and respond based on domain-specific knowledge.

2. Data Aggregation and Answer Summarization: A Reduce function aggregates outputs from the LLM-generated answers. This involves comparative analysis and synthesis, identifying overlaps, discrepancies, and complementary details to distill a coherent and comprehensive final answer.

3. Refinement and Contextual Adaptation: Before presenting the answer to the user, the response is refined for clarity, relevance, and coherence. This step adjusts the tone, style, or format to align with the user's query context, ensuring the response is informative and engaging.

By incorporating MapReduce into the question-answering framework, this methodology effectively harnesses distributed computing power to enhance scalability and responsiveness. Users receive precise, comprehensive answers synthesized from multiple sources, significantly improving the quality and reliability of the question-answering process.

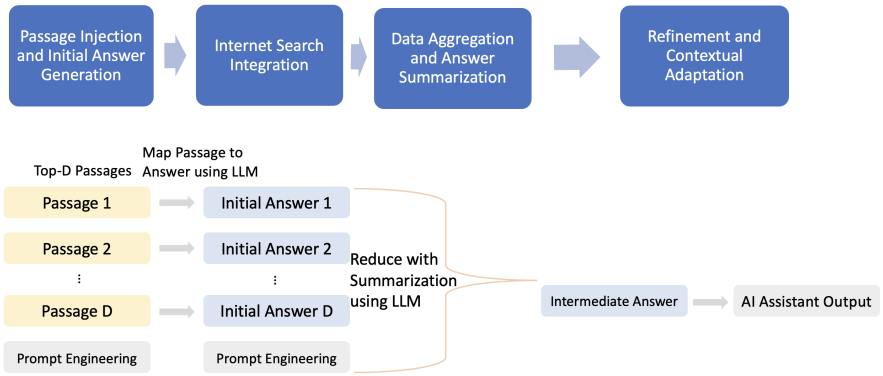

Figure 4: Question Answering Procedure

3. Model Evaluation

The human expert evaluation is designed to comprehensively assess the effectiveness, accuracy, and usability of the developed agency-specific project authoring advisor against traditional information search methods. The processing of evaluating the project authoring advisor involves its comparison with searching with two frequently used methods: (1) the commercial Google search engine (www.google.com); and (2) the agency's TDD. Comparing the project authoring advisor to these tools is crucial as they are commonly used by professionals in the AEC field. This comparison helps determine whether the project authoring advisor provides a competitive user experience, accurate and relevant responses, and efficient information retrieval.

Ten experts from the AEC field will participate in the evaluation. These experts, with extensive experience in handling the agency's technical documents and project authoring, conducted a comprehensive assessment of the project authoring advisor. Each expert was presented with six questions, divided into two categories (e.g., three questions on sustainable development and conservation, and three on future plans and upcoming projects). The same set of questions was also used to evaluate Google search and the agency's TDD for comparison.

The evaluation metrics are the same as those used in LLM-Assisted Evaluations (Comprehensiveness, Accuracy, Relevance, Clarity, and Conciseness) with an additional metric: Search Efficiency. This new metric assesses the balance between user-friendliness and the speed of retrieving relevant information, evaluating the overall effectiveness of the search experience. Such a comparison aims to determine the effectiveness of each of the three methods to answer the agency-specific questions. Experts were asked to select the best-performing method for each question in terms of Comprehensiveness, Accuracy, Relevance, Clarity, Conciseness, and Search Efficiency.

The evaluation results in Table 1 clearly demonstrate that the agency-specific project authoring advisor significantly outperforms both Google Search and the agency's TDD across all measured metrics. In terms of comprehensiveness, the project authoring advisor achieved the highest score of 16 points, providing more complete information compared to Google's 5 points and the TDD's 9 points. This indicates its ability to deliver detailed and thorough responses.

When evaluating accuracy, the advisor again excelled with a score of 14 points, surpassing Google's 6 points and being notably close to the TDD's 10 points. This high score suggests that the advisor is adept at providing precise and correct information efficiently.

The relevance of the advisor's answers was also rated highly, scoring 16 points, which is substantially higher than Google's 4 points and on par with the TDD's 10 points, demon-

Metrics	Advisor	Google Search	Agency's TDD	Best/Worst Ratio
Comprehensiveness	16	5	9	3.2
Accuracy	14	6	10	2.3
Relevance	16	4	10	4
Clarity	18	5	7	3.6
Conciseness	19	8	3	6.3
Search Efficiency	25	4	1	25

Table 1: Evaluation Score of Project Authoring Advisor, Google, and TDD

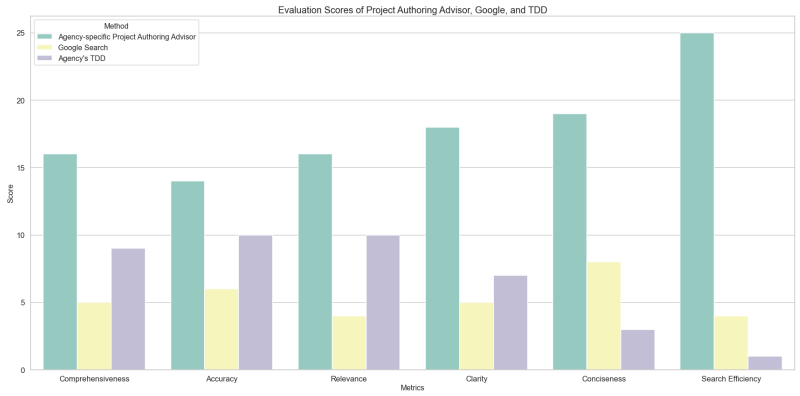

Figure 5: Evaluation Score of Project Authoring Advisor, Google, and TDD

strating its capability to generate highly pertinent responses to the queries posed.
The clarity of the responses from the project authoring advisor was another area where it stood out, scoring 18 points compared to Google's 5 points and the TDD's 7 points. This indicates that the advisor's answers are more understandable and clear. In terms of conciseness, the advisor achieved a remarkable score of 19 points, significantly outstripping Google's 8 points and the TDD's 3 points, showing its efficiency in providing succinct and to-the-point information.
The most notable performance was in search efficiency, where the advisor scored an impressive 25 points. This score is substantially higher than Google's 4 points and the TDD's 1 point, highlighting the advisor's superior balance between user-friendliness and speed in retrieving relevant information.
Overall, the agency-specific project authoring advisor demonstrates superior performance across all metrics, making it a highly effective tool for professionals in the AEC field. It offers significant improvements over traditional search methods like Google and the agency's TDD, providing a comprehensive, accurate, relevant, clear, concise, and efficient QA experience. This evaluation underscores the advisor's potential to enhance project authoring and technical document management in the AEC domain.

4. Discussion and Conclusion

This research introduces an innovative approach to improving question-answering systems for technical documents in the AEC domain. In this study, a generative project question answering system using RAG technique and domain-specific word embedding for technical documents in the AEC domain is developed. This system delivers precise, context-aware answers, thereby streamlining information retrieval and significantly enhancing the quality and accuracy of responses. This advancement empowers profes-

sionals in the engineering sector by facilitating more efficient interactions with extensive technical documentation, consequently reducing the time and effort traditionally required. Specifically, this research has the following two contributions:

Firstly, our research has established a highly automated, end-to-end pipeline that integrates several state-of-the-art techniques to fully harness the capabilities of LLMs in project authoring. Key techniques that this system used include prompt engineering, which tailors the inputs to the LLM to elicit the most relevant and accurate outputs; RAG, which enhances the LLM's response quality by dynamically incorporating information retrieved from a vast database; MapReduce framework, which constitutes the backbone of question answering procedure; and domain-specific word embeddings, which adapt the model to understand and process the unique terminology and nuances of the AEC industry. This pipeline sets a precedent for future advancements in effectively utilizing LLMs within the AEC domain.

In addition, a meticulously crafted prompt template for question answering is developed. This template integrate advanced techniques such as the CoT method and Few-shot Learning, which enhance the LLM's ability to process and generate domain-specific knowledge. By incorporating design patterns like the Persona Pattern, Question Refinement Pattern, Cognitive Verifier Pattern, and Format Template Pattern, this template effectively guides the LLM in producing precise and context-aware responses. This strategic approach streamlines the information retrieval process and significantly elevates the quality and accuracy of the answers provided.

In conclusion, this research not only achieves its objective of enhancing the efficiency and effectiveness of question-answering systems for technical documentation but also paves the way for future advancements in applying LLMs in the AEC domain. As we continue to explore and expand the capabilities of LLMs, the foundational work laid out by this project promises to spur further innovation, demonstrating the extensive applicability and benefits of artificial intelligence within the industry.

While the advancements presented in this research are noteworthy, there remain considerable opportunities for future investigations that could significantly benefit both academia and industry. Future work could explore different word embedding methods, experimenting with alternative techniques that may offer improved semantic understanding and richer contextual interpretation. By addressing these areas for future improvements, we can continue to elevate the system's performance, making it an even more powerful tool for engineering professionals navigating extensive technical documentation.

Acknowledgements

This research was funded by the Toronto and Region Conservation Authority (TRCA) and the University of Toronto. In addition, this research was made possible through the provision of datasets by TRCA, which significantly enhanced the depth and quality of our analysis. The authors gratefully acknowledge TRCA's support, especially from Jamil Benhamida and David Gingerich.

References

Giray, L. (2023). Prompt engineering with chatgpt: A guide for academic writers. *Annals of biomedical engineering*, *51*(12), 2629–2633.

Li, Y., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., & Jagadish, H. (2008). Regular expression learning for information extraction. *Proceedings of the 2008 conference on empirical methods in natural language processing*, 21–30.

Lin, H.-T., Chi, N.-W., & Hsieh, S.-H. (2012). A concept-based information retrieval approach for engineering domain-specific technical documents. *Advanced Engineering Informatics*, *26*(2), 349–360.

Logan IV, R. L., Balažević, I., Wallace, E., Petroni, F., Singh, S., & Riedel, S. (2021). Cutting down on prompts and parameters: Simple few-shot learning with language models [https://arxiv.org/pdf/2106.13353.pdf]. *arXiv preprint arXiv:2106.13353*.

Martino, A., Iannelli, M., & Truong, C. (2023). Knowledge injection to counter large language model (llm) hallucination. *European Semantic Web Conference*, 182–185.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730–27744.

Patil, S. P., & Pawar, B. (2018). Removing non-relevant links from top search results using feature score computation. *Bulletin of Pure & Applied Sciences-Mathematics and Statistics*, *37*(2), 311–320.

Rahutomo, F., Kitasuka, T., Aritsugi, M., et al. (2012). Semantic cosine similarity. *The 7th international student conference on advanced science and technology ICAST*, *4*, 1.

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7.

Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., & Liu, T. (2023). Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., & Lim, E.-P. (2023). Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *npj Digital Medicine*, *7*(1), 41.

Wang, N., Issa, R. R., & Anumba, C. J. (2022). Nlp-based query-answering system for information extraction from building information models. *Journal of computing in civil engineering*, *36*(3), 04022004.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

White, J., Hays, S., Fu, Q., Spencer-Smith, J., & Schmidt, D. C. (2023). Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design [https://www.dre.vanderbilt.edu/ schmidt/PDF/prompt-patterns-book-chapter.pdf]. *arXiv preprint arXiv:2303.07839*.

Zhang, R., & El-Gohary, N. (2021). A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking. *Automation in Construction*, *132*, 103834.

Zhong, B., He, W., Huang, Z., Love, P. E., Tang, J., & Luo, H. (2020). A building regulation question answering system: A deep learning methodology. *Advanced Engineering Informatics*, *46*, 101195.

Zhong, Y., & Goodfellow, S. D. (2024). Domain-specific language models pre-trained on construction management systems corpora. *Automation in Construction*, *160*, 105316.

Zuccon, G., & Koopman, B. (2023). Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*.