OXFORD CDE

Journal of Computational Design and Engineering, 2025, **12**, 94–120

DOI: 10.1093/jcde/qwaf033

Advance access publication date: 21 March 2025

Research Article

Evaluating ChatGPT on Korea's BIM Expertise Exam and improving its performance through RAG

Youngsu Yu [1], Sihyun Kim [2], Wonbok Lee [2] and Bonsang Koo [2,*]

[1]Urban Construction Engineering & Management, Seoul National University of Science and Technology, 232 Gongreung-ro, Nowon-gu, Seoul 01811, Korea

[2]Department of Civil Engineering, Seoul National University of Science and Technology, 232 Gongreung-ro, Nowon-gu, Seoul 01811, Korea

*Correspondence: bonsang@seoultech.ac.kr

Abstract

This study aimed to evaluate ChatGPT-3.5 and ChatGPT-4's understanding of the building information modelling (BIM) knowledge domain by testing on the multiple-choice section of Korea's BIM Expertise Exam (KBEE), a professional license exam tailored for BIM. ChatGPT-4 achieved passing scores across all tested years and averaged 85%, 20% higher than its predecessor. Nevertheless, ChatGPT-4 had difficulty in certain areas including the 'BIM guidelines' subcategory, as the problems required access to documents specific to Korea's BIM policies and stipulations. By supplying the required documents using Retrieval Augmented Generation (RAG), GPT-4's score further improved to 88.6%, an improvement of 25.7%. The results provide evidence that ChatGPT-4, albeit within the context of the KBEE, is overall knowledgeable in its understanding of the BIM domain. The results of RAG demonstrate that partial gaps in its knowledge can be bolstered by providing the appropriate documents. The study verified that this capability was particularly valuable for categories that varied depending on local and regional contexts. The findings suggest that ChatGPT, when integrated with RAG, holds significant potential as a comprehensive and adaptive knowledge model in the BIM domain, and thereby alleviates the barriers for BIM adoption, such as the lack of experts, the cost of training, complexity of the BIM models, and additional workloads required during the BIM delivery process.

Keywords: building information modelling, large language model, Generative Pre-trained Transformer, Retrieval Augmented Generation, professional exam, knowledge assessment

1. Introduction

ChatGPT, a generative language model based on OpenAI's Generative Pre-trained Transformer (GPT) architecture, has advanced the field of natural language processing (NLP) by enabling machines to understand and generate human-like text. Trained on diverse data sets, ChatGPT demonstrates proficiency in contextual understanding and complex query response, positioning it as a valuable tool across various fields, including engineering, education, healthcare, and more (Liu *et al.*, 2023).

To examine its capabilities, recent studies have emerged that tested its performance on major professional qualification exams, with the goal of assessing its level of understanding in their respective knowledge domains. Examples include applications to the bar exams (Bommarito & Katz, 2022; Martínez, 2024), the medical board exams (Kasai *et al.*, 2023; Kung *et al.*, 2023; Strong *et al.*, 2023; Takagi *et al.*, 2023), and collegiate entrance exams (OpenAI, 2023). The studies showed ChatGPT capable of passing these exams, although it had difficulty in highly specialized areas (Lecler *et al.*, 2023). Its success in these exams has created opportunities to use ChatGPT as a virtual assistant (George & George *et al.*, 2023) or tutoring agent (Baidoo-Anu & Ansah, 2023).

To date, however, no such tests have been performed with regard to building information modelling (BIM). BIM has become a critical technology in the architecture, engineering, and construction (AEC) industry, with its own set of policies, concepts, processes, technical tools, and nomenclature unique to its domain (Koo *et al.*, 2021, Yu *et al.*, 2022). Determining ChatGPT's level of understanding would verify its adaptability to BIM's domain-specific requirements and also provide legitimacy as a potential application of the domain.

Thus, the immediate goal of this study was to provide an objective evaluation of ChatGPT's knowledge of the BIM domain. A subsequent goal was to investigate how its performance could be improved if and when any shortcomings were to be identified. The first goal was achieved by testing ChatGPT on a professional exam, i.e. Korea's BIM Expertise Exam (KBEE). The exam consists of a broad range of topics related to BIM, including software utilization, general concepts and processes, environmental requirements, potential applications, and Smart construction technologies. Examining its performance across the different categories would provide insight into its relative strengths and weaknesses. The second goal was achieved by using Retriever Augmented Generation (RAG) to provide domain-specific documents to GPT-4 for one of its weakest subcategories. RAG was deployed to retrieve relevant chunks of information for problems previously answered incorrectly and thus improve ChatGPT's responses.

To the best of our knowledge, measuring ChatGPT's level of understanding with respect to BIM has thus far not been investigated. Most studies integrating ChatGPT with BIM have focused on leveraging its capabilities for semantic search (Rane *et al.*, 2023),

Received: December 16, 2024. **Revised:** March 10, 2025. **Accepted:** March 11, 2025

information retrieval (Zheng & Fischer, 2023; Jang *et al.*, 2024), design visualization (Jo *et al.*, 2024; Lee *et al.*, 2024), and task planning (Amer *et al.*, 2021; Prieto *et al.*, 2023; Singh *et al.*, 2023; Hwang *et al.*, 2024). Score results from the KBEE provide an objective measure of its relative strengths, which should be taken into account when utilizing ChatGPT. The use of RAG demonstrates how gaps in its knowledge could be improved and thus deploy ChatGPT with higher confidence.

More interestingly, this study posits that ChatGPT provides a comprehensive and adaptive knowledge management framework previously unavailable for use in the BIM domain. Compared to existing expert systems, ChatGPT and more broadly LLM's, present a unique opportunity to deploy a 'top-down' approach, where the general knowledge of a domain is provided *a priori* and customizations (e.g. fine-tuning or RAG, etc.) for specific subdomains are performed contingent upon a need's basis. Hence, having an understanding of its knowledge expertise and customization techniques, provides critical insights into its utility, benefits, and potential applications.

Commensurately, the study assesses how ChatGPT can be incorporated to mitigate the existing barriers hampering BIM adoption, such as the lack of experts, the cost of training, complexity of the BIM models, and additional workloads required during the BIM delivery process (Ullah *et al.*, 2019). Using the results of the tests, we also provide recommendations on how it should by deployed as a general-purpose knowledge model for the BIM domain.

The rest of this paper is structured as follows. Section 2 introduces studies that have tested GPT on professional exams, an overview of the KBEE exam as well as its eligibility as a test for determining BIM expertise. The section also explores existing approaches to customize and augment the response of LLMs, Section 3 outlines the overall methodology for the two goals of the study. Section 4 summarizes the approach and results of testing on the KBEE, while Section 5 presents how the RAG framework was utilized for a subcategory of the exam. Section 6 discusses the implications of the results, limitations, and recommendations for ChatGPT deployment.

2. Research Background

2.1 GPT: a brief overview

The GPT, developed by OpenAI, is a leading large language model (LLM) known for excelling in NLP tasks such as translation, summarization, and text generation (OpenAI, 2023). Trained in extensive publicly available text from websites, books, and articles, GPT produces coherent and contextually relevant outputs.

Earlier versions like GPT-1 and GPT-2 faced limitations, often generating biased or inaccurate outputs due to their reliance on next-word prediction and lack of intent comprehension (Gehman *et al.*, 2020; Bender *et al.*, 2021). GPT-3.5 addressed these issues by incorporating Reinforcement Learning from Human Feedback, enhancing its alignment with human preferences and ability to perform complex tasks (Chan, 2023). GPT-4 introduced significant advancements, including an expanded context window for processing longer inputs and tackling sophisticated tasks (Hadi *et al.*, 2023). With an estimated 170 trillion parameters—roughly 1000 times more than GPT-3.5—and a larger, more recent data set extending to April 2023, GPT-4 offers improved accuracy and superior performance, particularly in non-English NLP tasks (Baktash & Dawodi, 2023; Takagi *et al.*, 2023).

ChatGPT, a specialized application of GPT, is optimized for interactive, conversational tasks. While GPT excels in general language generation, ChatGPT is tailored for dynamic interactions, such as answering questions, engaging in dialogue, and generating creative content like stories or code (Antaki *et al.*, 2023). Building on the advancements of GPT-3.5 and GPT-4, ChatGPT delivers more adaptive and versatile responses, making it particularly effective for conversational and interactive applications. This study differentiates the models by using the terms, GPT-3.5 and GPT-4 when denoting the LLM, and ChatGPT to refer to conversational AI.

2.2 Testing the performance of ChatGPT on professional license exams

Many studies have tested ChatGPT using professional license exams in multiple domains of expertise, including law (Bommarito & Katz, 2022; Choi *et al.*, 2022; OpenAI, 2023; Katz *et al.*, 2024; Martínez, 2024), finance (Callanan *et al.*, 2023; Zhang *et al.*, 2023a; Kim *et al.*, 2024), medicine (Kasai *et al.*, 2023; Kung *et al.*, 2023; Strong *et al.*, 2023; Takagi *et al.*, 2023), and math and science (Boiko *et al.*, 2023; OpenAI, 2023; Wang *et al.*, 2024; Lam *et al.*, 2024). This section introduces these studies with a focus on the bar exam, the medical license exam, and also major secondary and tertiary entrance exams.

- Bar exam

Several studies have tested versions of the GPT model on bar examinations. Most notably, Bommarito & Katz (2022) applied the GPT-3.5 model to the United States (US) bar exam, which is administered by the National Conference of Bar Examiners. While the content and structure of the examination vary from state to state, most states have recently adopted the Uniform Bar Examination. The examination format includes three sections: (1) multiple-choice, (2) essay, and (3) scenario-based sections. The multiple-choice section, the focus of the study, accounts for 50% of the total grade. This section consists of 200 questions, with 25 questions for each of the eight categories.

The test was administered to GPT-3.5 using two different scenarios: (1) to only select the correct answer, and (2) to select and rank the top three answers. In the first scenario, GPT-3.5 achieved a 50.3% correct response rate, which is significantly higher than the 25% expected from random chance but lower than the 68% average score of human test-takers. In the second condition, the average correct response rate rose to 88%, and still higher than the random chance rate of 75%. The lowest scores were observed in the Criminal Law section, which is reputed to have the most specialized legal terminology of the eight areas. This indicates a relative weakness of GPT-3.5 in highly technical and specialized areas.

- Medical license exam

Several studies have also tested GPT models on medical license exams. In particular, Kasai *et al.* (2023) evaluated and compared the performance of GPT-3, GPT-3.5, and GPT-4 on the multiple-choice section of the Japanese National Medical Examination. The exam consists of a total of 400 questions, divided into six categories (28 subcategories in total), each consisting of 50–75 questions. The questions are also distinguished into three types with varying minimum passing scores: 70% for 'general', 80% for 'required', and no more than three incorrect answers for the 'prohibited' type. The last type asks questions related to life-threatening and fatal consequences. The exam, which has a high average pass rate of 91.7%, is administered to medical school graduates from Japan who are eligible to apply for residency upon passing.

Table 1: Problem categories of the KBEE.

Category	Definition
BIM software	Encompasses understanding different BIM authoring tools, focusing on software functionalities that support collaboration, and documentation. Also involves creating and managing BIM libraries to ensure modelling that fits the project requirements
BIM general knowledge	Covers the foundational principles and theoretical frameworks of BIM, including core concepts, standards, guidelines, and procurement methodologies crucial for the effective implementation of BIM in construction projects
BIM environment	Involves the establishment of environmental requirements for BIM application in design and construction, including BEPs, classification systems, standards such as Industry Foundation Classes (IFC), Geographic Information System (GIS) integration, and LOD to ensure data consistency and interoperability
BIM applications	Focuses on the practical use of BIM, including parametric modelling, 3D visualization, effectiveness assessment, and quality control. This category emphasizes BIM's role in enhancing design accuracy, construction processes, and facility management
Smart construction technologies	Examines the integration of advanced technologies with BIM, such as 3D scanning, AR (Augmented Reality)/VR (Virtual Reality), IoT (Internet of Things), Artificial Intelligence (AI), and construction automation, to improve productivity, accuracy, and sustainability in the construction industry.

In the study, five years' worth of the exam questions were input into GPT models using the original Japanese language. The results showed that GPT-4, followed by GPT-3.5, and GPT-3, recorded the highest performance, and only GPT-4 successfully passed all five years of the exams. Moreover, only GPT-4 was able to succeed in meeting the criterion for the 'prohibited' questions. Despite passing the exam, GPT-4 scored lower than the average of real-world test takers.

- Tertiary education competency exams

OpenAI conducted an independent analysis of the performance of GPT-3.5 and GPT-4 on the Scholastic Assessment Test, the Graduate Record Examination, and Advanced Placement (AP) subject tests in the US (OpenAI, 2023). The results demonstrated that GPT-4 outperformed GPT-3.5 in these tests, with the image-readable version of GPT-4 achieving the best performance, scoring in the top 20% in most subjects. Among AP subjects, GPT-4 scored over 90% in psychology and US history, while scoring poorly in AP calculus and the American Mathematics Competitions exam. This suggested that the GPT models have weaknesses in mathematics and scientific reasoning, which has been supported by research such as Wang *et al.* (2024). Addressing these deficiencies may involve integrating additional systems or techniques specifically designed to enhance the model's performance in specific domains.

As discussed, a comparable study for the BIM knowledge domain is currently absent. GPT models potentially provide an unprecedented opportunity as a knowledge management tool for the BIM domain, that is unique from existing legacy knowledge models. However, a crucial drawback is in its lack of transparency and reliability as its sources are not made explicit, making it challenging to trust its responses, especially from an engineering knowledge standard point. There is a lack of research that provides evidence in this regard, which in turn makes it difficult to argue with confidence in what capacity and form it should be used. Thus, conducting the test provides legitimacy to its capabilities and a basis for arguing its utility, benefits, and potential applications in the BIM domain, which are consequently also absent in the current literature.

Methodologically, this study adopted several of the approaches used in these studies in designing our performance test with respect to the GPT models' understanding of the BIM domain. Specifically, our research also tested GPT models on multiple-choice problems, used categories to identify and distinguish relative performance, and used the original native language as input. However, these studies did not provide or suggest a way to improve the weaknesses of GPT, whereas this study used RAG for performance and response improvement.

2.3 Korea's BIM Expertise Exam and its eligibility

2.3.1 *KBEE overview*

The Korean government has made strong initiatives to incorporate BIM as an essential technology for increasing productivity and reducing costs in the lifecycle of construction projects. In 2023, the Ministry of Land, Infrastructure, and Transport (MOLIT) amended its by-laws to stipulate the use of BIM as mandatory for public projects over 100 billion won (approximately US $71 million). These projects are required to submit designs, perform quality checks using BIM, and propose use cases in the construction phase.

With these mandates, the government has identified the need to ensure that engineering firms and contractors have the capability to implement BIM-based projects correctly and efficiently (Choi *et al.*, 2023). Commensurately, the government has endorsed a license examination, KBEE, to test their competency in using BIM processes and software skills. Engineers and contractors are encouraged to take these exams and get certified as a way to show their proficiency in using BIM for public projects. Although currently not mandatory, it is comparable to the Fundamentals of Engineering exam, which is a prerequisite exam to being licensed as a Professional Engineer in the US.

The Korean BIM Evaluation Agency0F,[1] which administers the test on behalf of MOLIT, states the exam tests the ability to (1) plan and execute BIM tasks for civil and architectural projects, (2) coordinate and manage projects by establishing phased and discipline-specific work processes, (3) define public construction and data exchange protocols, and (4) facilitate communication among project participants.

Category-wise, the KBEE tests for knowledge pertaining to the following five major categories, i.e. BIM software, BIM general knowledge, BIM environment, BIM applications, and Smart construction technologies. Their definitions are provided in Table 1.

The KBEE consists of three levels, i.e. levels 1–3, with level 1 being the most advanced. The level 1 exam consists of a 50-question, multiple-choice exam, and a subsequent oral interview exam.

[1] https://www.bimkorea.or.kr/main/

Table 2: Mappings between BIM professional's competencies and KBEE categories.

BIM professional's competencies \ KBEE Categories		BIM software	BIM general knowledge	BIM environment	BIM applications	Smart construction technologies
Knowledge areas	Knowledge about construction design and contracting procedures		✓			
	Merits and demerits of BIM for design/construction/operation processes		✓		✓	
	BIM model progression and specification		✓	✓		
	Data security		✓	✓		
	Information management				✓	
	Contractual and legal aspects of BIM implementation		✓			
	BIM standardization		✓	✓		
	BIM implementation procedures		✓	✓		
Domain-specific skills	Basic BIM operating skills	✓				
	Central databases	✓				
	Interoperability from one tool to another	✓		✓		
	Storing, sharing, and accessing of information	✓		✓	✓	
	Ability to interact with a model in 3D interface	✓			✓	
	Change analysis				✓	
	Ability to extract specifications and information from the BIM model	✓				

Candidates must score over 60 points to pass both the written and oral exams. The exam is administered three times a year. This study used the questions from the level 1 exam, as it consisted of the most advanced and comprehensive questions related to BIM.

2.3.2 *KBEE's eligibility for evaluating ChatGPT's expertise in the BIM domain*

This study selected KBEE to test ChatGPT's understanding of BIM domain knowledge as the test was the official exam for testing BIM professionals in the authors' native country. More importantly, the test was deemed to be the most comprehensive for determining the expertise of BIM professionals, when compared to other major certification or license exams for BIM. Most notable exams include UK's RICS Certificate in BIM and buildingSMART's Professional Certification, which are both administered by public, non-profit institutions. Private BIM software providers, namely Autodesk and Bentley also provide certifications, but the scope of these exams is limited to the functionalities of their respective software.

To demonstrate KBEE's comprehensiveness, this study adopted the criteria identified by Saka & Chan (2020), who conducted a Delphi survey and determined eight knowledge areas, seven domain-specific skills, and nine domain-specific functionalities regarded by quantity surveyors to be competent and functional in a BIM environment. Because the latter functionalities were specific to quantity surveying, i.e. cost related, only the two former areas were utilized.

As shown in Table 2, the KBEE covers the majority of the required criteria: the eight knowledge areas are addressed by KBEE's 'BIM general knowledge', 'BIM applications', and the 'BIM environment' categories, while the seven domain-specific skills are covered mostly by the 'BIM software' category as well as the 'BIM applications', and the 'BIM environment' categories. As shown in Table 3, the UK's RICS subject areas are more conceptual and process oriented, focusing on the roles and and responsibilities throughout the lifecycle of BIM-based delivery processes. BuildingSMART's exam is focused on evaluating the competency of BIM professionals' knowledge and application of openBIM standards at different levels of expertise.

Thus, although the KBEE cannot be said to be exhaustive, it does relatively test on a broad range of areas that are regarded essential for BIM professionals in today's AEC industry.

2.4 Improving GPT responses

Despite the transformational advances achieved with each upgrade, even the most recent GPT models may not always provide appropriate or satisfactory results for tasks within specific domains. When queries pertain to more recent data or specific domains, gaps in the training data can lead to 'hallucinations' where the model generates plausible but incorrect answers (Zhang *et al.*, 2023a; Zhang *et al.*, 2023b2023b). This issue is particularly pronounced in knowledge-heavy domains such as medical, legal, and other technical fields, and is a widespread challenge faced by LLMs, not just GPT (Yang *et al.*, 2023).

The construction of LLMs, such as GPT, to fulfill domain necessities and to address hallucinations is constrained by the challenges and high costs associated with data collection. Three approaches, i.e. prompt engineering, fine-tuning, and RAG, have emerged to effectively adapt pre-trained LLMs, i.e. foundation models, for specific domains and tasks (Dubois *et al.*, 2023).

- Prompt engineering is the process of designing and refining input prompts to get desired outputs from language models like GPT (Zhou *et al.*, 2022). Since language models generate responses based on prompts, the quality and structure of the input directly influence the results. Chain of thought (Wei *et al.*, 2022) is a typical approach used in engineering prompts. It involves providing the model with a multi-step framework or reasoning path within the prompt, rather than simply asking for an answer outright. This method enhances the accuracy and depth of responses, particularly in tasks requiring logical reasoning or complex decision-making.
- Fine-tuning is the process of taking a pre-trained model and adjusting it to perform better on a specific task or data set

Table 3: Subject areas of UK's RICS certificate in BIM and buildingSMART's professional certification.

Certification or license exams for BIM	Objective	Subject areas or exam levels
RICS' Certificate in BIM	Exam is conceptual and process oriented, focusing on the roles and responsibilities of BIM professionals throughout the lifecycle of BIM-based delivery processes.	Transition to ISO 19650/Assessment and need/Invitation to tender and tender response/Appointment and mobilization/ Information production and planning/Information model delivery and project close–out/BIM for operations
BuildingSMART's Professional Certification	Exam is focused on evaluating the competency of BIM professionals' knowledge and application of openBIM standards at different levels of expertise.	Entry level/Foundation level/Management level/Practitioner level

(Jeong, 2024). This involves training the model further on domain- or task-specific data to make it more specialized. Typically, this can be achieved by training the models on new or updated information within a specific domain, using a collection of prompt-completion pairs.

- RAG is a method where the foundation model combines information retrieval and generation (Lewis *et al.*, 2020). In this approach, the model retrieves relevant documents or data from an external knowledge base and then uses that information to generate a response. This helps to provide accurate and up-to-date responses, especially in knowledge-heavy domains (Ram *et al.*, 2023). Without RAG, the model may generate responses based on its training data, which may be outdated. With RAG, the LLM can retrieve relevant, up-to-date papers or articles from an external database and incorporate the information into its response. This leads to a more accurate and contextually relevant answer (Sahoo *et al.*, 2024).

The goal of this study was to enhance the responses of GPT for BIM-related domain knowledge by providing needed documents that may not have been available to GPT trained in general knowledge. Prompt engineering enhances the response of an LMM by guiding the reasoning steps it should take. However, it does not provide a way to provide additional external knowledge from which it can refer to provide better responses. Fine-tuning does enable training of LLM to improve its responses. However, it requires the preparation of query-response pairs and is most suited when such task-specific, labelled training sets are available. Most common use cases thus include text classification (e.g. spam/non-spam email) or QA systems (e.g. FAQ data sets) (Huang *et al.*, 2023). Thus, RAG was judged to be the most suited approach as it allows external sources to be provided in a way that allows an LLM to retrieve relevant information and respond within the given context. The detailed process for setting up an RAG pipeline is introduced in the following section.

2.5 The RAG framework

The general steps of incorporating RAG are illustrated in Figure 1.

1. *Source data collection*. Constructing an RAG starts with gathering the required documents of a specific domain of knowledge. The collected data can include both structured and unstructured formats, with textual documents being a primary source of information. The source data should be curated to ensure relevance to the intended application or domain, as well as to minimize noise and redundancy.
2. *Chunk splitting*. Splitting is the process of dividing the collected source documents into 'chunks', which are small, manageable units suitable for efficient processing. This step is essential to ensure that the data can be effectively embedded and retrieved during later steps. The chunks should be sufficiently large to maintain context and semantic continuity, but small enough to facilitate efficient retrieval. The maximum chunk size should also take into account the token limits of the retriever and LLMs (Yepes *et al.*, 2024). Juvekar & Purwar (2024) suggest chunk sizes between 512 and 1 024 tokens. Overlapping sections are incorporated between chunks to maintain semantic flow and improve coherence (Crossley *et al.*, 2019).
3. *Embedding and storing in a vector database*. Embedding is the process of vectorizing the chunks into a high-dimensional vector space where chunks with similar semantics are represented more closely to each other. Transformer-based models are commonly used due to their ability to generate context-sensitive vectors that adapt based on the surrounding text. Google's Bidirectional Encoder Representations from Transformers (BERT, Devlin *et al.*, 2018), and OpenAI's text-embedding-ada models (OpenAI, 2022), are widely used. The choice of an embedding model is critical and must be carefully tailored to factors like the language of the source data and the complexity of the retrieval task (Tang & Yang, 2024). The embeddings are stored in a vector database which organizes the high-dimensional vectors to facilitate quick similarity calculations. Databases like Chroma DB1F[2] and Pinecone2F[3] are widely adopted for this purpose.
5. *Constructing the retriever*. A retriever ranks and retrieves the most relevant chunks stored in the vector database in response to a given query (Arabzadeh *et al.*, 2021). Two types of retrievers exist. Sparse retrievers, like BM25 (Robertson & Zaragoza, 2009) and Term Frequency-Inverse Document Frequency (TF-IDF), use keyword-based methods and are efficient but lose semantic information by focusing on exact word matches. Dense retrievers, such as Dense Passage Retriever (Karpukhin *et al.*, 2020) and BERT-based models, use vector embeddings to capture the semantic meaning of the text, achieving higher accuracy, though at a higher computational cost. Sparse and dense retrievers can be combined as an ensemble retriever. By combining keyword matches with semantic context, ensemble retrievers can improve overall performance and better select relevant information. These retrievers are particularly useful when dense retrievers alone may struggle to retrieve domain-specific information that includes technical jargon or specialized terms (Mandikal & Mooney, 2024).
6. *Prompt design*. In the RAG framework, the prompt typically consists of three distinct components, i.e. instruction, con-

[2] https://www.trychroma.com/
[3] https://www.pinecone.io/

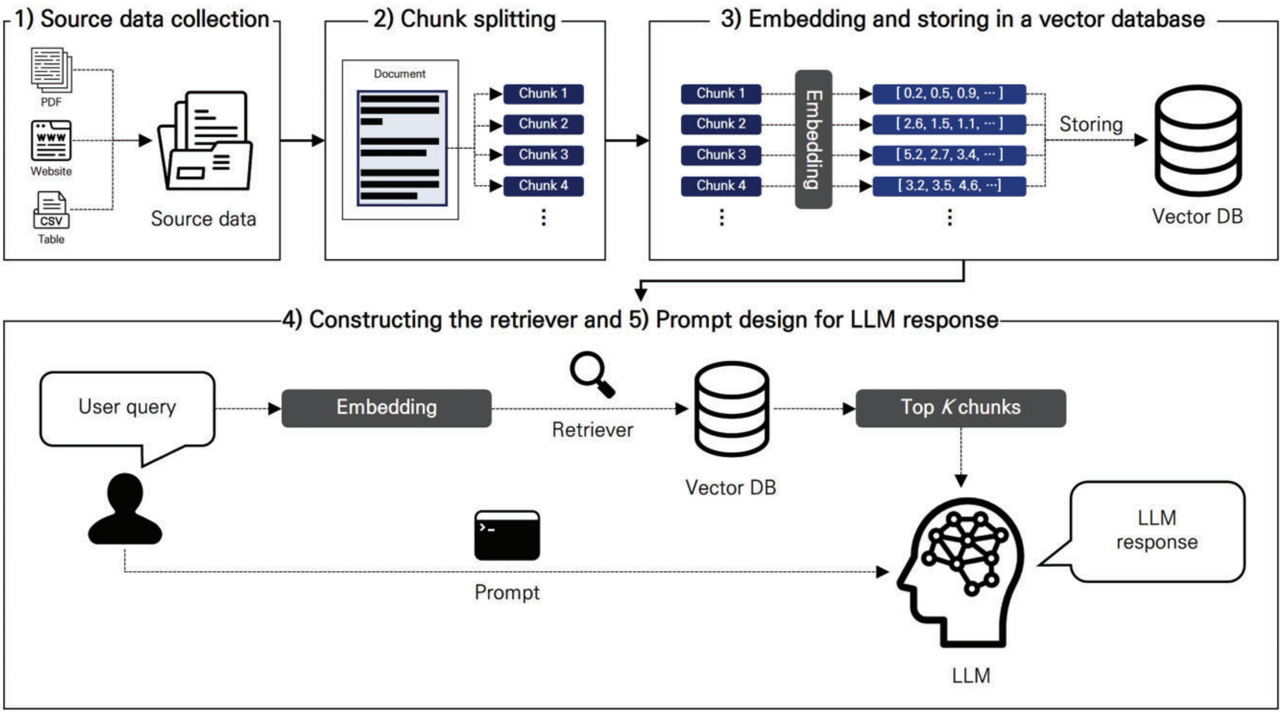

Figure 1: The RAG framework.

text, and query. The instruction provides guidelines for generating responses, including format, tone, and constraints, to ensure alignment with task objectives. The context refers to the retrieved chunks from the vector database that serve as the factual basis for the model's response. The query represents the user's original request and guides both the retrieval process and the generation of the response. By aligning these components, the query enables the LLM to generate accurate and contextually relevant responses using RAG techniques.

3. Research Methodology

As shown in Figure 2, this study was performed in two distinct but related phases. In Phase 1, the objective was to evaluate the competency of ChatGPT with respect to its comprehensive understanding of the BIM domain. Specifically, we aimed to quantify GPT's knowledge in a similar fashion to the other benchmark studies performed for the legal and medical domains, as introduced in Section 2.2. Moreover, we also wished to identify and distinguish GPT's areas of relative strengths and weaknesses.

Part 1 of Phase 1 in Figure 2 shows the process used to meet these goals. First, ChatGPT-3.5 and ChatGPT-4 were tested on the multiple-choice section of the annual KBEE's level 1 exam spanning over 5 yr (i.e. years 2019–2023). The goal was to show whether the two versions of ChatGPT were able to pass the exams while identifying discrepancies in their performance.

For part 2 of Phase 1, the 5 yr's worth of questions were further categorized into 20 subcategories and also classified as short-phrase or full-sentence problems. Duplicate questions were removed to ensure the integrity of the data set. This detailed categorization would enable a more precise evaluation of the knowledge level of the ChatGPT in each subdomain. The experiment would also help identify areas where additional knowledge and improvement are most required.

The objective for Phase 2 was to deploy RAG with the goal of improving GPT's performance, in particular in the weakest subcategories, previously identified in Phase 1. This involved collecting relevant documents related to the specific area, creating embeddings, and storing their vectors in a vector database. An ensemble retriever was employed to extract relevant information from the vector database, thus assisting GPT in better answering the questions it previously got wrong. Subsequently, the scores were compared and evaluated to determine their performance improvement quantitatively. The approach would demonstrate how RAG enables better performance of GPT by enabling domain-specific BIM knowledge to be tailored in its responses.

Because Phase 2 is dependent on the results of Part 2 of Phase 1, the detailed research approach and implementation, as well as each of the phases' results are explained sequentially in Sections 4 and 5.

4. Phase 1: Testing ChatGPT-3.5 and 4 on the KBEE Level 1 Exam

4.1 Test set up for Phase 1

4.1.1 *Compiling the exam problems*

As discussed, five years' worth of multiple-choice problems of the KBEE were used. As shown in Figure 3, a multiple-choice problem includes a question followed by four choices. Both the question and multiple-choices in the figure are originally in Korean and have been translated into English by the authors for readability.

Table 4 shows which exams were used for each part of Phase 1. For part 1 of Phase 1, only the questions from the multiple-choice section of the level 1 exam were used. Exams are administered three times a year, but only the exams from the first trimester were used. Exams conducted in the past 5 yr (2019–2023) were used, resulting in a total of 250 questions. For part 2 of Phase 1, all three exams conducted over the five years were used, resulting in

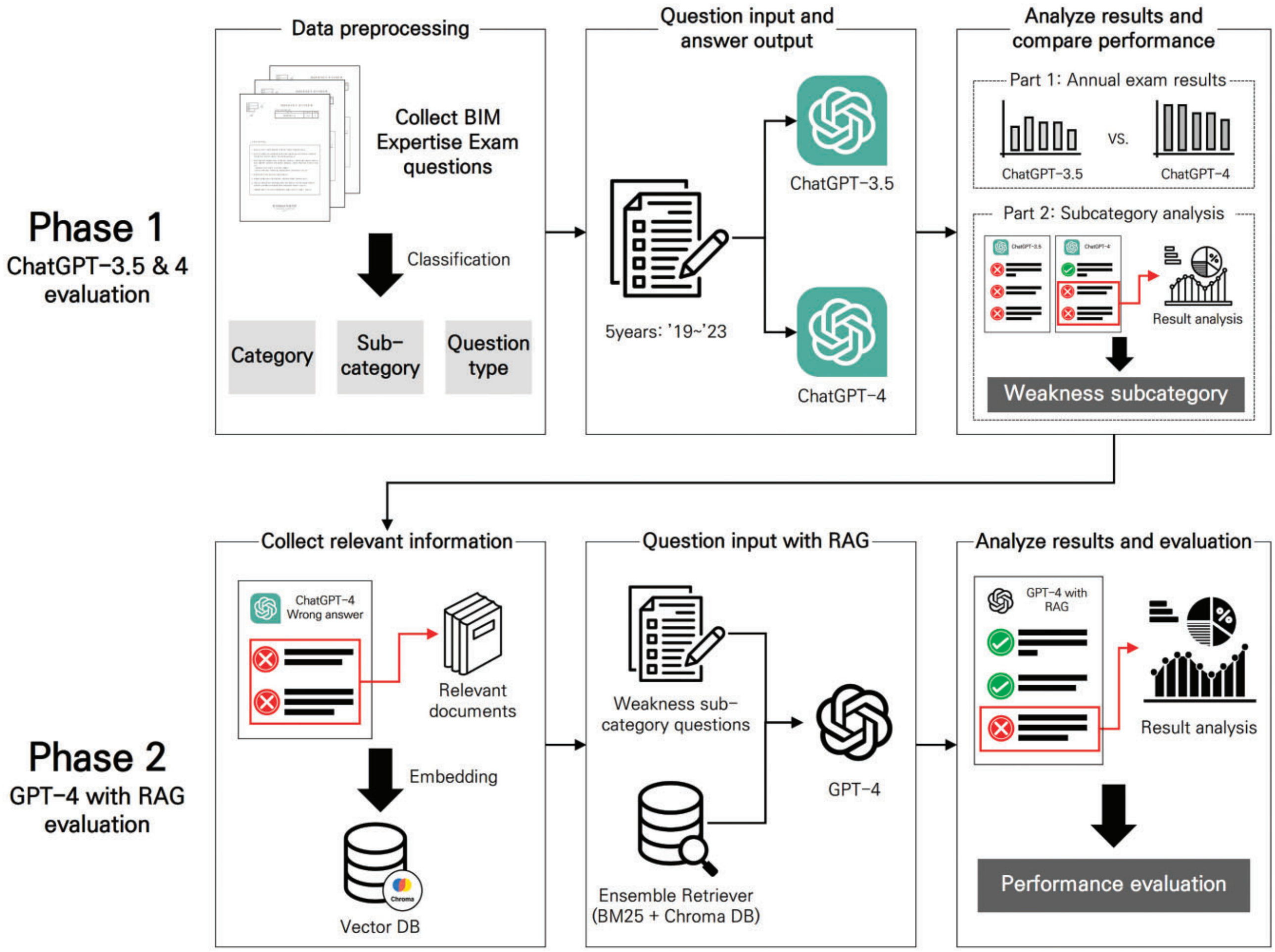

Figure 2: Research methodology.

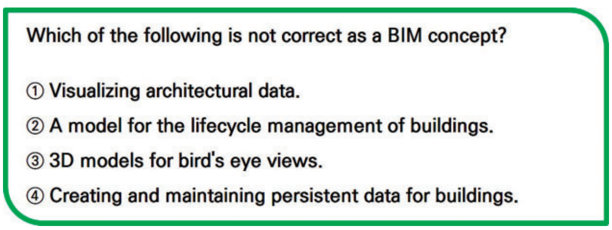

Figure 3: KBEE multiple-choice example problem.

a total of 750 questions from 15 exams. Individual inspection of these questions by the authors identified 420 duplicates. Removal of these questions resulted in a total of 330 unique questions.

4.1.2 *Subcategorization*

As described in Table 1 of Section 2.3.1, the multiple-choice problems are categorized into five major categories. Figure 4 shows the ratio of the 330 questions for each of these original five categories. The discipline of 'BIM software', with 130 out of the 330 questions, had the highest number of questions, accounting for 39.4% of the total. This was followed by 'BIM general knowledge' at 21.5% and 'BIM environment' at 18.8%. The proportion of the questions reflects the exam's focus on assessing candidates' proficiency in BIM modelling techniques as well as in the general knowledge of BIM.

As discussed, the goal of part 2 of Phase 1 was to increase the granularity of the original categories to get a finer determination of the performance differences between the two versions of ChatGPT. Thus, the original five divisions were further classified into 20 subcategories. We first referenced Lemaire *et al.* (2019), who categorized the major BIM domain knowledge topics through a literature review. Then, the problems in each category were scrutinized and subclassified within their original category. For example, the 'BIM software' category included questions regarding information sharing and collaboration tools ('Collaboration techniques'), 2D drawings and schedule generation ('Document generation'), BIM library generation and application ('Library management'), detailed functions in BIM authoring or specialty software ('Software functions'), and Software types and usage ('Software types'). Table 5 shows the definitions for the resulting 20 new subcategories together with the number and ratio of the initial questions included in each subcategory. Among the subcategories, 'Software functions' had the largest proportion of questions, accounting for 27.6% of the questions, followed by 'BIM guidelines' at 10.6%.

4.1.3 *Short-phrase and full-sentence problem categorization*

Each of the 330 problems was also classified into either short-phrase or full-sentence problems, depending on the length of the multiple-choice answers. Specifically, the short-phrase problems had choices that consisted of a succinct and concise response, such as a number or a term, while the full-sentence problems consisted of choices with longer sentences that provided a detailed

Table 4: Overview of KBEE usage for each part of phase 1.

Step \ Year	2019			2020			2021			2022			2023			
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	Total
Part 1	✓			✓			✓			✓			✓			5 exams (250 questions)
Part 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	15 exams (330 unique questions)

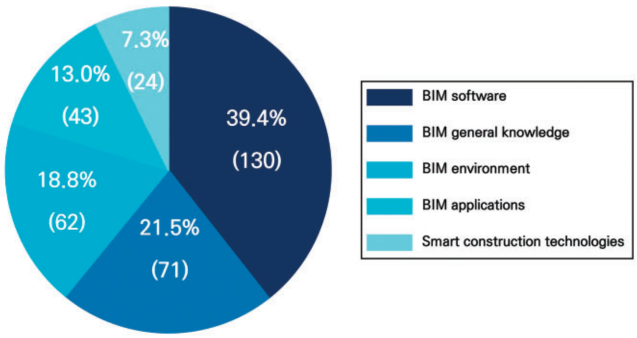

Figure 4: Distribution of questions by category in KBEE.

explanation or the meaning of a term. This classification allowed for the relative evaluation of the ChatGPT models' capabilities in comprehending complex explanations of BIM-related domain knowledge. Figure 5 illustrates examples of short-phrase and full-sentence type problems. Based on this criterion, 235 (71.2%) of the total 330 problems were classified as short-phrase problems, while 95 (28.8%) were classified as full-sentence problems (Figure 6).

4.1.4 *Prompt input*

Each problem was individually input into the two versions of ChatGPT, together with the instruction, 'Choose the best answer to the following question', preceding the multiple-choice problem itself (Figure 7). Both the instruction and problem were input in Korean, the original language of the exam. No other directions were provided, which was in line with the previous studies conducted for evaluating ChatGPT for other knowledge domains and deemed necessary to assess its performance objectively (Oztermeli & Oztermeli, 2023). Each problem was also entered into a new chat session, to ensure that ChatGPT operates in isolation, i.e. without relying on the context or data from previous dialogues. This format was used for both parts 1 and 2 of Phase 1 of this study.

4.2 Test results for Phase 1

4.2.1 *Part 1: annual exam results*

Table 6 and Figure 8 show the scores of ChatGPT-3.5 and ChatGPT-4 for the exams taken between the years 2019 and 2023. ChatGPT-3.5 achieved an average score of 65 points across the five exams. Specifically, ChatGPT-3.5 passed the exam for the years 2019 through 2022, scoring over the 60-point minimum score requirement. However, it failed for the year 2023, with a score of 58 points. Comparatively, the ChatGPT-4 passed the exam for all five years, with an average of 85 points.

Both models scored the lowest in the 2023 exam, suggesting the test was particularly hard in that year. The 20-point difference in the averages suggests that ChatGPT-4 performs substantially better than its predecessor, which is consistent with the results of the other exam-based studies introduced in Section 2.2 (Kasai *et al.*, 2023; OpenAI, 2023). A paired *t*-test conducted using the annual scores resulted in a *t*-value of 9.55 and a *p*-value of 0.00067, showing that the difference is statistically significant.

Although it is difficult to pinpoint the reason behind the superior performance, we attribute it to ChatGPT-4's notable enhancements in comprehension and reasoning, based on its training and parameter size improvements (Baktash & Dawodi, 2023). Moreover, as the exam problems were provided in Korean, ChatGPT-4's superior multi-lingual capabilities also seem to have played a role in its performance (Koubaa, 2023).

4.2.2 *Part 2: exam results by subcategory*

Table 7 and Figure 9 present the results of ChatGPT-3.5 and ChatGPT-4's performance across the subcategories. Overall, ChatGPT-4 consistently outperformed ChatGPT-3.5 in all subcategories. ChatGPT-3.5 was limited to answering 200 out of the 330 unique questions correctly, while ChatGPT-4 correctly answered 263, representing a 19.1% superior performance. A paired *t*-test conducted based on the 20 subcategories resulted in a *t*-value of 9.87 and a *p*-value of 6.49e−09, showing that ChatGPT-4's higher performance is statistically significant.

The table also provides the number of answers improved and the related percentage point increased per subcategory. Most notably, the 'Software functions' and the 'BIM guidelines' provided the highest contributions in improving the overall score of ChatGPT-4. Yet, these categories still had the highest number of incorrectly answered. Thus, they were investigated in more detail.

- Investigation of the 'Software functions' subcategory

The 'Software functions' subcategory had the largest contribution to the overall improved score at 6.97% and yet still incorrectly answered 19 of the 91 total questions. Investigation of the problems in this subcategory indicated a difference in the exposure between the two ChatGPT models to the relevant documents needed for understanding BIM software functionalities.

Figure 10 provides an example problem. The question asks which module or panel is not a part of Autodesk's Revit User Interface (UI). ChatGPT-3.5 was incorrect, while ChatGPT-4 correctly chose answer number 4, the 'Options Window'. Although difficult to verify, it seems ChatGPT-4 has better access to Revit's user manuals, which are publicly accessible on the Autodesk website3F.[4] Alternatively, the user manuals could have been included in its training data. Conversely, the error seems to indicate that

[4] https://help.autodesk.com/view/RVT/2024/ENU/

Table 5: Subcategorization result of the KBEE.

Category (Count)	Subcategory	Description	Count	Ratio (%)
BIM software (130)	Collaboration techniques	Functions for collaboration and information sharing in the BIM model	5	1.5
	Document generation	Generation of project documents such as drawings, and reports within the BIM environment	7	2.1
	Library management	Repositories of BIM objects for standardization	10	3.0
	Software functions	Functions that enhance efficiency in BIM software	91	27.6
	Software types	Types of BIM software for specific functions like design or management	17	5.2
BIM general knowledge (71)	BIM concepts	Fundamental concepts for BIM data	18	5.5
	BIM guidelines	Standards and best practices for BIM implementation	35	10.6
	Procurement methods	Strategies for acquiring BIM services, tools, and defining contractual relationships	18	5.5
BIM environment (62)	BEP	Pertains to the creation of documentation for the execution of BIM projects	3	0.9
	Classification systems	Systems for categorizing BIM elements across projects	5	1.5
	Data standards	Standards for managing and exchanging BIM data	23	7.0
	GIS	Integration of Geographic Information Systems (GIS) with BIM	10	3.0
	IFC	Open data standards for BIM software interoperability	11	3.3
	LOD	Levels of development (LOD) in BIM models	10	3.0
BIM applications (43)	BIM modelling techniques	Creation techniques of digital models for construction	19	5.8
	BIM benefits	Benefits of BIM in project outcomes and processes	14	4.2
	Quality control	Concerns about the application of quality control methods in a BIM-based project	10	3.0
Smart construction technologies (24)	3D scanning	3D scanning technologies to digitally capture existing conditions for construction	8	2.4
	BIM-based technologies	Technologies that utilize BIM data for advanced design and construction	7	2.1
	Construction automation technologies	Automation of construction tasks using BIM data and technologies	9	2.7
		Total	**330**	**100.0**

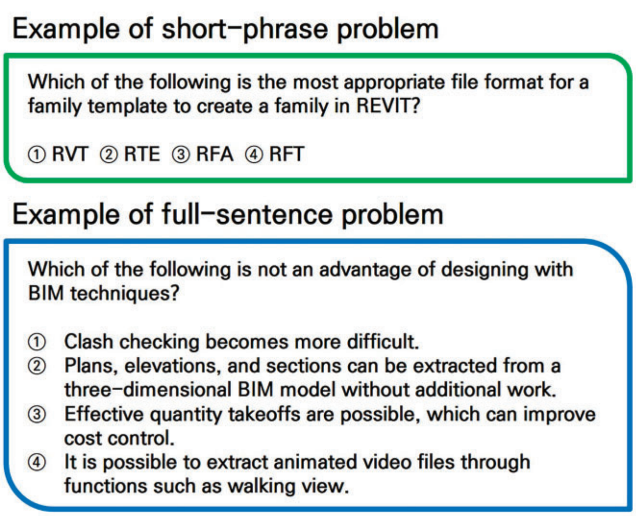

Figure 5: KBEE example problems of short-phrase and full-sentence.

ChatGPT-3.5 does not have access to this information, suggesting limited exposure to such documents.

However, despite its improved performance, ChatGPT-4 still responded incorrectly to 19 of the questions. Investigation of these problems suggested that the translation of the function names into Korean may be attributed to these errors. Function names such as ‘Insert Material type’ are provided as ‘ (jaejil sabyib)’ in Korean, which may be more difficult for ChatGPT-4 to recognize, even with access to the software's original user manuals.

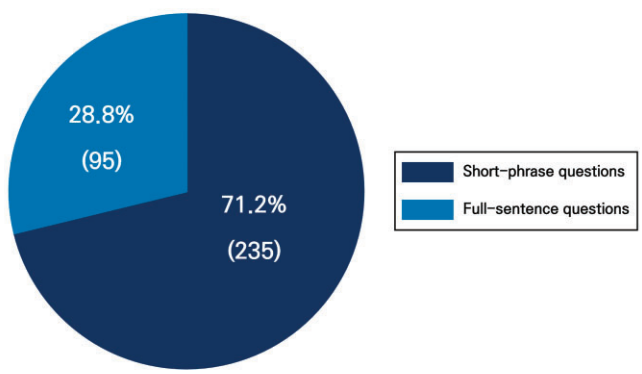

Figure 6: Distribution of problems by type in KBEE (short-phrase versus full-sentence problems).

- Investigation of the ‘BIM guidelines’ subcategory

The ‘BIM guidelines’ subcategory had the second largest contribution to the overall improved score at 1.82%. Yet ChatGPT-4 still incorrectly answered 13 of the 35 total questions. A deeper inves-

Figure 7: Prompt used for ChatGPT-3.5 and 4.

tigation revealed that both ChatGPT models incorrectly answered questions specific to concepts, processes, and requirements stipulated in Korea's local BIM guidelines and regulations.

For example, Example problem no. 2 in Figure 11 requires understanding regulations released by Korea's Public Procurement Service Agency, specifically with regard to the building information level (BIL) in the planning and design phases. Example problem no. 3 requires having knowledge with regards to the BIM standards and guidelines published by various Korean government institutions.

The reason for the poor performance seems to be manifold. The most apparent reason seems to be the lack of ready access to documents that stipulate BIM mandates specific to Korea. The documents being in the Korean language, and stipulations being different from internationally recognized standards may also have contributed. In any case, the provision of specific documents to the models seems to be needed, and the use of RAG is described in Section 5.

4.2.3 Part 2: *exam results for short-phrase/full-sentence categorization*

Table 8 and Figure 12 present the results for evaluating ChatGPTs' performance with respect to the length of the multiple-choice answers. ChatGPT-4 consistently outperformed ChatGPT-3.5 in both problem types. For short-phrase problems, ChatGPT-3.5 achieved 144 correct answers out of 235 (61.3%), while ChatGPT-4 correctly answered 188 questions (80.0%), showing an 18.7% improvement. For full-sentence problems, ChatGPT-4 answered 75 correctly out of the 95 questions (78.9%), indicating a 20.0% improvement over ChatGPT-3.5. A paired t-test conducted resulted in a t-value of 29.77 and a p-value of 0.0214, showing that the difference between the two models is statistically significant.

Both models tested slightly higher on the short-phrase problems, though the difference was not significant. Overall, ChatGPT-4 demonstrated superior performance, and the results, as stated in Section 2.1, are attributed to its expanded context window and improved accuracy in non-English NLP tasks.

5. Phase 2: Enhancing GPT-4 With RAG

The results of part 2 of Phase 1 showed that both versions of ChatGPT had difficulty in the 'BIM guidelines' subcategory, and the reason most likely lies in the inaccessibility of guidelines and protocols specific to the Korean industry. This section introduces how RAG was used with the goal of improving ChatGPT's response exclusively to this subcategory. As ChatGPT-4 performed better than its predecessor, the test was conducted for only this version. The test and implementation steps are first introduced followed by the results.

5.1 Test set up for Phase 2

As introduced in Section 2.5, implementing RAG involves collecting relevant document sources, creating vectorized representations of those documents through chunking and embedding, and storing them in a vector database. The RAG pipeline was implemented using the LangChain[5] framework and executed on a system equipped with an Intel Xeon Silver 4210R CPU, an RTX A6000 GPU, and 128 GB of RAM. The process is outlined as follows.

5.1.1 *Source data collection*

As shown in Table 7 of Section 4.2.2, ChatGPT-4 incorrectly answered 13 questions related to the 'BIM guidelines' subcategory. Thus, we investigated these questions in detail and collected relevant documents that we believed could be used in the RAG architecture (the 13 problems are provided in Table A1 of Appendix A).

Specifically, 13 documents deemed most relevantly needed were collected (Table 9). Eight of the documents were published by MOLIT of Korea. These documents included Open-BIM standard guidelines for architectural and infrastructure sectors (MOLIT, 2016, 2017), BIM strategy and roadmaps for public projects (MOLIT, 2018; 2020a, b), and more recently the 'Basic Guidelines for BIM in the Construction Industry' (MOLIT, 2020c) and the 'BIM Stipulation Guidelines' (MOLIT, 2022a). These two latter documents mandated the employment of BIM for all public projects exceeding 100 billion Korean won. They also specified the standards and details such as Levels of Development (LOD), quality control, roles and responsibilities, etc. One more specific document to Smart Construction, 'S-Construction 2030' by MOLIT (2022b), was added.

Four of the documents were published by the major public agencies of Korea, including the Korea Expressway Corporation (Korea Express Corporation, 2020, 2021), the Seoul Institute of Technology (Seoul Institute of Technology, 2022), and the Public Procurement Service Agency (PPS, 2022). These documents detailed the BIM implementation requirements for public projects in their jurisdiction and had additional stipulations and specifications compared to those provided by MOLIT.

Lastly, a document was added about the BIM educational resource published by the Korean Institute of BIM (KIBIM, 2021), a leading academic organization.

5.1.2 *Chunk splitting*

The collected Korean BIM guideline documents were imported using a Python library, PyPDF,[6] and split into smaller chunks through the splitter with the RecursiveCharacterTextSplitter6F[7] library in LangChain. This library ensured that chunks maintained semantic relevance and consistency by splitting text at natural breaks, such as sentences or paragraphs. As discussed in Section 2.5, a chunk size was specified to be 512 tokens for optimal performance. The splitter adjusted dynamically to ensure coherent boundaries, resulting in smaller chunks when a natural break was reached before exceeding the size limit (LangChain, 2024). Unfortunately, the final number of chunks could not be clearly determined because the chunks were split by the splitter and passed directly to the next embedding step. Therefore, a quantitative database (DB) size was presented in the next step instead.

[5] https://www.langchain.com
[6] https://pypdf.readthedocs.io/en/stable/
[7] https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/

Table 6: ChatGPT scores for annual KBEE's level 1 exam for years 2019–2023.

	ChatGPT-3.5			ChatGPT-4			Score difference
Model Year	Correct	Incorrect	Score	Correct	Incorrect	Score	(ChatGPT-4 - ChatGPT-3.5)
2019	31	19	62	44	6	88	▲26
2020	37	13	74	44	6	88	▲14
2021	32	18	64	42	8	84	▲20
2022	33	17	66	42	8	84	▲18
2023	29	21	58	41	9	82	▲24
Average	**32**	**18**	**65**	**43**	**7**	**85**	**▲20**

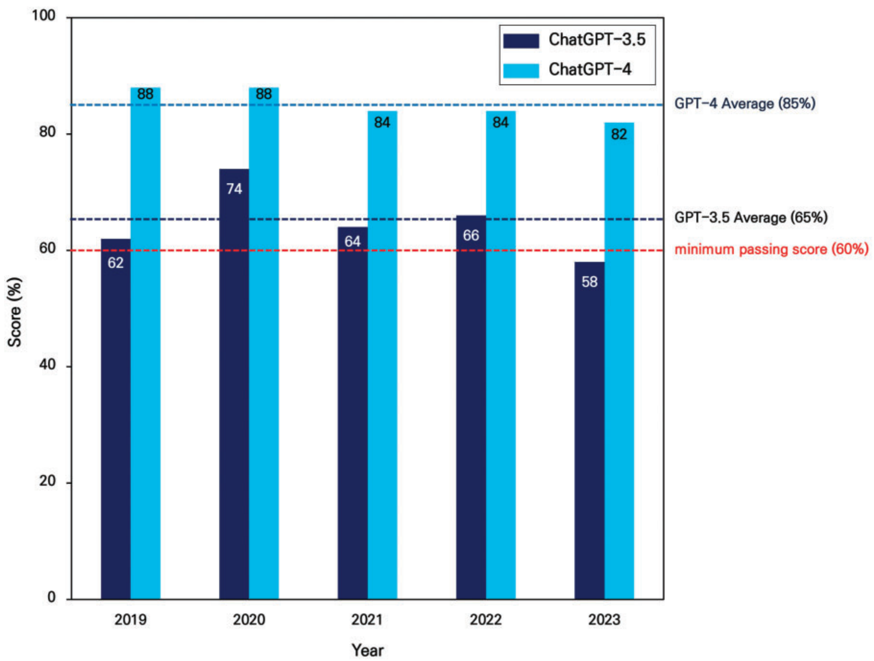

Figure 8: Comparison of ChatGPT-3.5 and 4 scores in annual KBEE's level 1 exam (2019–2023).

Additionally, an overlap of 10% (51 tokens) was applied between chunks to ensure contextual continuity, for balancing context preservation and efficient storage.

5.1.3 *Embedding and storing in a vector database*

The segmented chunks were embedded using OpenAI's text-embedding-ada-002 model7F,[8] which has demonstrated strong performance for embedding tasks (Balikas, 2023). This model converted each chunk of text into a 1536-dimensional vector, ensuring a rich semantic representation. For vector storage, Chroma DB, an open-source vector database, was utilized. Chroma DB is designed to integrate seamlessly with LangChain, enabling efficient storage, retrieval, and chaining of vectors (Chroma, 2024).

The 13 documents totaling 177 MB (2231 pages) were reduced to 55 MB in size in the vector database after chunking and embedding. This size reduction reflects the efficient compression and storage achieved and also ensures that the data remains retrievable without excessive computational overhead.

5.1.4 *Constructing the retriever*

The Chroma DB by default employs a dense retriever that uses continuous vector representations of text for search and retrieval. However, as discussed in Section 2.5, relying solely on a dense retriever can result in inaccurate information retrieval. Since dense retrievers are optimized for English data, they may struggle to capture the unique language structure and nuances of the Korean language.

Thus, an ensemble retriever approach was implemented, combining Chroma DB's dense retriever with a BM25 sparse retriever, with the goal of leveraging both keyword-based and semantic matches. Again, this method has proven to be particularly effective when dealing with documents containing domain-specific technical terms and definitions (Mandikal & Mooney, 2024) and has been adopted by studies focused on identifying the best RAG architecture for technical documents (Arslan *et al.*, 2024; Cenić et al, 2024; He *et al.*, 2025).

As illustrated in Figure 13, both the sparse and dense retrievers were configured to retrieve the top two most relevant chunks. The four retrieved chunks were then weighted equally and incorporated into the context section of the prompt for further processing.

5.1.5 *Prompt design*

Figure 14 illustrates the process of using the RAG framework to assist GPT-4 in answering a given question. Each question is embedded as a query and fed into Chroma DB in its original Korean

[8] https://platform.openai.com/docs/guides/embeddings/

Table 7: Performance of ChatGPT-3.5 and 4 on the KBEE by subcategory.

Subcategory (# of questions) \ Model	ChatGPT-3.5			ChatGPT-4			# of answers improved (percentage points increased)	Percent contribution to overall improvement*	Rank by percentage contribution
	# of correct answers	# of incorrect answers	Correct rate	# of correct answers	# of incorrect answers	Correct rate			
Collaboration techniques (5)	4	1	80.00%	5	0	100.00%	▲1 (20.0%)	0.30	12
Documentation generation (7)	5	2	71.40%	6	1	85.70%	▲1 (14.3%)	0.30	12
Library management (10)	7	3	70.00%	8	2	80.00%	▲1 (10.0%)	0.30	12
Software functions (91)	49	42	53.80%	72	19	79.10%	▲23 (25.3%)	6.97	1
Software types (17)	11	6	64.70%	14	3	82.40%	▲3 (17.6%)	0.91	5
BIM concepts (18)	14	4	77.80%	16	2	88.90%	▲2 (11.1%)	0.61	8
BIM guidelines (35)	16	19	45.70%	22	13	62.90%	▲6 (17.1%)	1.82	2
Procurement methods (18)	9	9	50.00%	14	4	77.80%	▲5 (27.8%)	1.52	3
BEP (3)	2	1	66.70%	3	0	100.00%	▲1 (33.3%)	0.30	12
Classification systems (5)	3	2	60.00%	4	1	80.00%	▲1 (20.0%)	0.30	12
Data standards (23)	14	9	60.90%	18	5	78.30%	▲4 (17.4%)	1.21	4
GIS (10)	7	3	70.00%	9	1	90.00%	▲2 (20.0%)	0.61	8
IFC (11)	8	3	72.70%	10	1	90.90%	▲2 (18.2%)	0.61	8
LOD (10)	5	5	50.00%	7	3	70.00%	▲2 (20.0%)	0.61	8
BIM modeling techniques (19)	14	5	73.70%	17	2	89.50%	▲3 (15.8%)	0.91	5
BIM benefits (14)	10	4	71.40%	11	3	78.60%	▲1 (7.1%)	0.30	12
Quality control (10)	5	5	50.00%	8	2	80.00%	▲3 (30.0%)	0.91	5
3D scanning (8)	5	3	62.50%	6	2	75.00%	▲1 (12.5%)	0.30	12
BIM-based technologies (7)	6	1	85.70%	6	1	85.70%	-	0	20
Construction automation technologies (9)	6	3	66.70%	7	2	77.80%	▲1 (11.1%)	0.30	12
Total	**200**	**130**	**60.60%**	**263**	**67**	**79.70%**	**▲63 (19.1%)**	**19.09**	**-**

*(ChatGPT-3.5's # of incorrect answers – ChatGPT-4's # of incorrect answers) * 100 / # of total

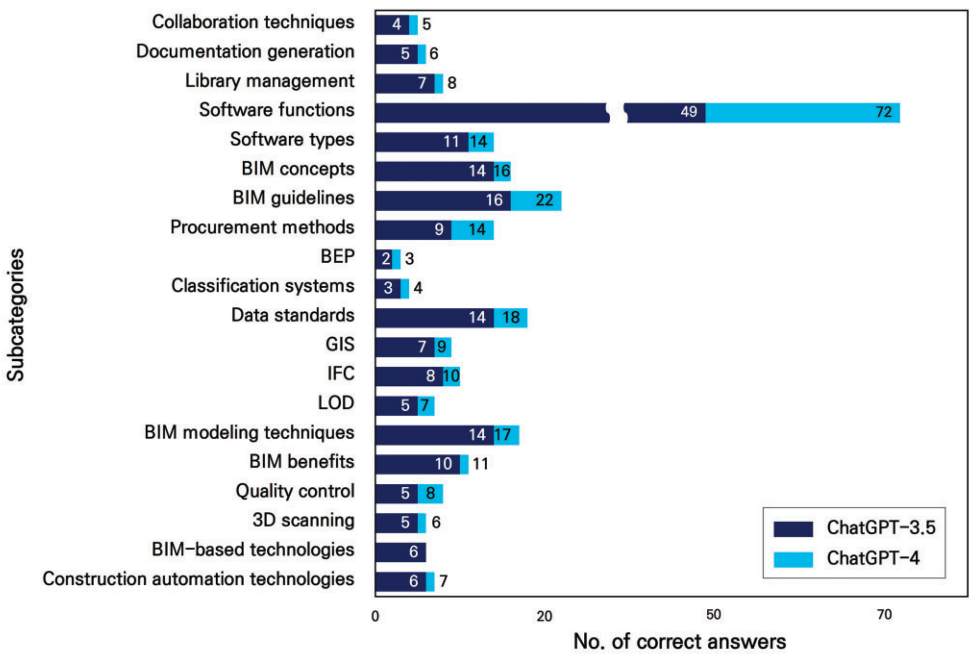

Figure 9: Comparison of correct answers for ChatGPT-3.5 and 4 by subcategory.

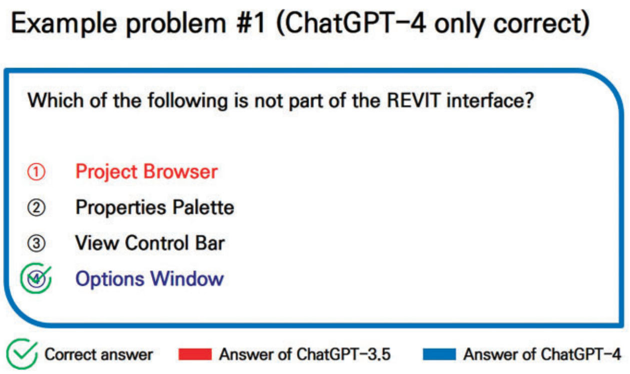

Figure 10: Example problem in the 'Software functions' subcategory (response of ChatGPT-3.5 and 4).

language. The ensemble retriever retrieves the four most relevant chunks from the vector database. The original query, along with the retrieved chunks, is sent to the prompt. An instruction is provided as part of the prompt, specifying the use of the chunks for answering the given query. The prompt is then sent to GPT-4, which responds based on the provided context. This process was applied to all 35 questions in the 'BIM guidelines' subcategory, including the 13 questions previously answered incorrectly.

5.2 Test results for Phase 2

Table 10 shows the results of using RAG-based GPT-4 for the 35 questions. Using RAG, GPT-4 correctly answered 31 out of 35 questions, resulting in a score of 88.6%. Compared to ChatGPT-4, which answered 13 questions incorrectly and scored 62.9%, RAG-based GPT-4 achieved a 25.7% improvement.

The results demonstrate that the RAG framework was effective in improving the performance of GPT-4. The results also correspondingly demonstrate that the documents embedded in the vector database and the design of the vector database and retriever were also appropriate.

Selected questions are provided to demonstrate in detail how RAG was used, while also showing an example where the answer was still incorrect.

5.2.1 *Exam problems answered correctly using RAG*

RAG-based GPT-4 correctly answered all 22 questions that were also answered correctly originally by ChatGPT-4, while answering 9 of the 13 questions that were previously answered incorrectly by ChatGPT-4.

Figure 15 shows one such problem. Example problem no. 4 inquiries about the required precision (scale) of digital terrain drawings in the Detailed Design stage of a project. The figure shows a snippet of the actual chunk used by GPT-4 to answer the question. The chunk is an excerpt from chapter 6 'Geological and Site Investigations' of the 'Expressway Smart Construction Guideline' (p. 223). As this document was embedded in the vector database, and the retriever identified the relevant chunk, GPT-4 was able to provide the correct answer. In contrast, ChatGPT-4 provided an incorrect response, as it lacked access to the document in question.

5.2.2 *Exam problems answered incorrectly using RAG*

Investigation into the four problems that were still incorrect revealed that the relevant information was not absent from the source documents or was not retrieved properly by the retriever. Example problem no. 5 in Figure 16 is provided to exemplify this issue (refer to Appendix Table B1, for the other three problems). The example problem requires having knowledge about the BIM-related standards or BIM-based systems employed by various countries, i.e. specifically Singapore, the UK, Japan, and Korea. Both ChatGPT-4 and GPT-4 with RAG selected choice no. 4: 'Korea Expressway Corporation's BIM brand – EX-BIM,' as the incorrect answer.

Inspection of the chunks provided by the retriever revealed the reason behind the incorrect selection. As shown in Table 11, a total of four chunks were retrieved: two each by the BM25 (sparse retriever) and Chroma DB retriever (dense retriever). The top two chunks, retrieved by BM25, contained relevant informa-

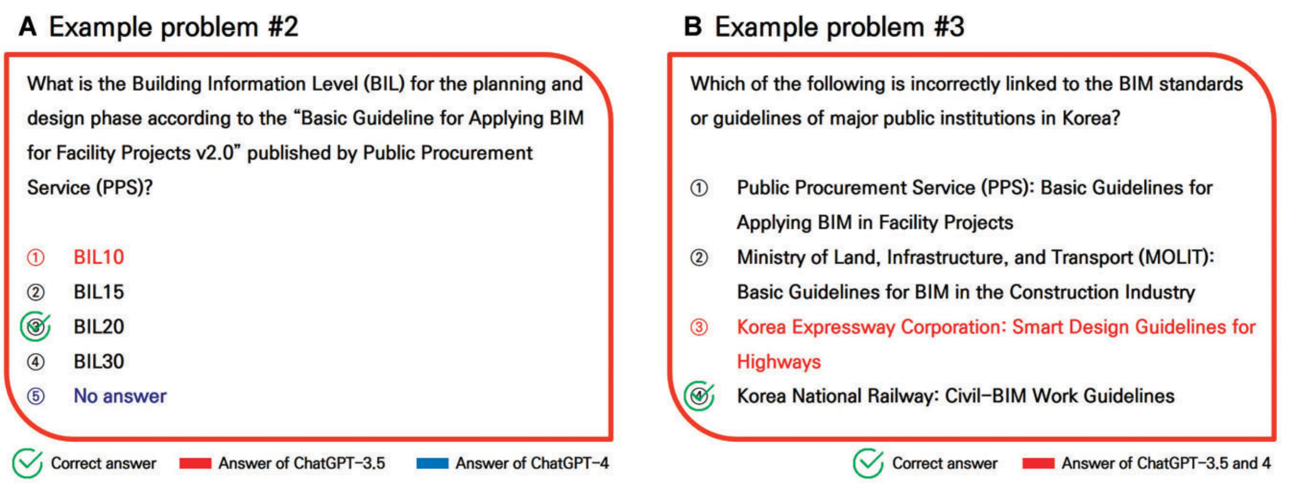

Figure 11: Example problems in the 'BIM guidelines' subcategory (response of ChatGPT-3.5 and 4): (A) and (B) both ChatGPT-3.5 and 4 are incorrect.

Table 8: Performance of ChatGPT-3.5 and 4 in KBEE by problem type (short-phrase versus full-sentence problems).

Model / Type (# of questions)	ChatGPT-3.5			ChatGPT-4			Correct rate difference (ChatGPT-4 - ChatGPT-3.5)
	# of correct answers	# of incorrect answers	Correct rate	# of correct answers	# of incorrect answers	Correct rate	
Short-phrase (235)	144	91	61.3%	188	47	80.0%	▲18.7%
Full-sentence (95)	56	39	58.9%	75	20	78.9%	▲20.0%
Total	**200**	**130**	**60.6%**	**263**	**67**	**79.7%**	**▲19.1%**

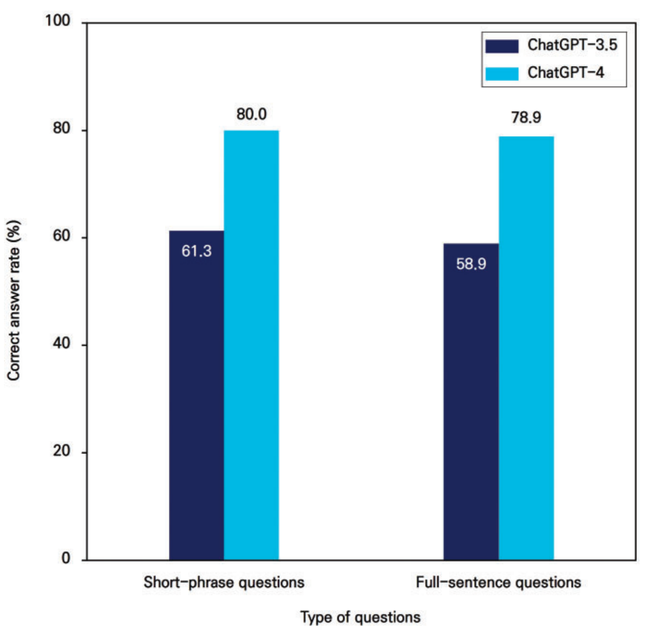

Figure 12: Comparison of correct rates for ChatGPT-3.5 and 4 by problem type (short-phrase versus full-sentence problems).

tion related to choice no. 1, i.e. definitions concerning Singapore's CORENET. However, the other chunks, retrieved by the Chroma DB retriever, did not provide sufficient information to solve the question. Specifically, the third chunk contained generic information regarding knowledge management systems in advanced countries, and the fourth chunk was entirely unrelated to the question.

A closer inspection of the source documents revealed that the information related to PAS (choice no. 2), i-Construction (choice no. 3), and EX-BIM (choice no. 4) did exist in the source documents (Figure 17). However, unlike the definition for CORENET (choice no. 1) which was presented in plain text, their definitions were provided in tabular format. Manual inspection of their chunks revealed that they did not contain a semantically coherent sequence of texts that properly retained the definitions originally stated in the tables. For example, Figure 18 shows that the definition of PAS in table format was not properly retained in the corresponding chunk, with the tabular structure (e.g. row and column headings) lost and replaced with unnecessary line breaks (i.e. '\n').

Thus, based on the 'garbled' chunk, the retriever did not provide this as a viable chunk with high relevancy and thus was not included, leading to an erroneous answer. Such difficulties in interpreting and utilizing textual data in tables for both GPT-4 and RAG have been well-documented and seem to be the reason in this case (Pan *et al.*, 2022; Allu *et al.*, 2024).

Table 9: List of BIM guidelines-related documents.

No.	Publication date	Published by	Title
1	2016.06	Ministry of Land, Infrastructure and Transport	Report of enhancement and substantiation plan for core technology of Open BIM-based architectural design
2	2017.01	Ministry of Land, Infrastructure and Transport	Appendix of Open BIM-based building design standards and infrastructure construction
3	2018.09	Ministry of Land, Infrastructure and Transport	BIM roadmap and activation strategies for public SOC projects
4	2020.09	Korea Expressway Corporation	Expressway smart design guidelines
5	2020.09	Ministry of Land, Infrastructure and Transport	Architect's scope of work and consideration for public works projects
6	2020.12	Ministry of Land, Infrastructure and Transport	2030 architecture BIM activation roadmap
7	2020.12	Ministry of Land, Infrastructure and Transport	Basic guidelines for BIM in the construction industry
8	2021.04	Korea Expressway Corporation	Expressway BIM design evaluation criteria
9	2021.09	Korean Institute of BIM	BIM educational resources
10	2022.06	Seoul Institute of Technology	A study on the guideline and roadmap for applying BIM to Seoul
11	2022.07	Ministry of Land, Infrastructure and Transport	S-Construction 2030
12	2022.11	Ministry of Land, Infrastructure and Transport	Construction industry BIM stipulation guidelines (owner, contractor, and designer)
13	2022.12	Public Procurement Service	Guidelines for applying BIM for facilities projects

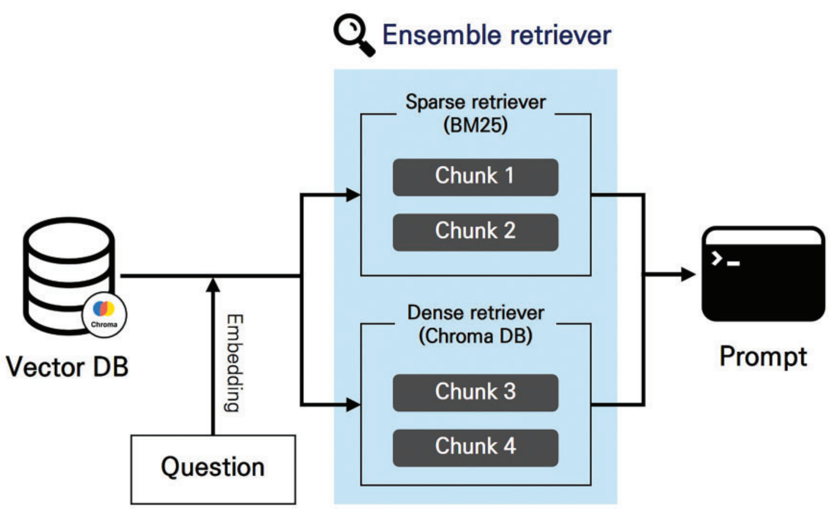

Figure 13: Process of ensemble retriever combining dense and sparse retrievers.

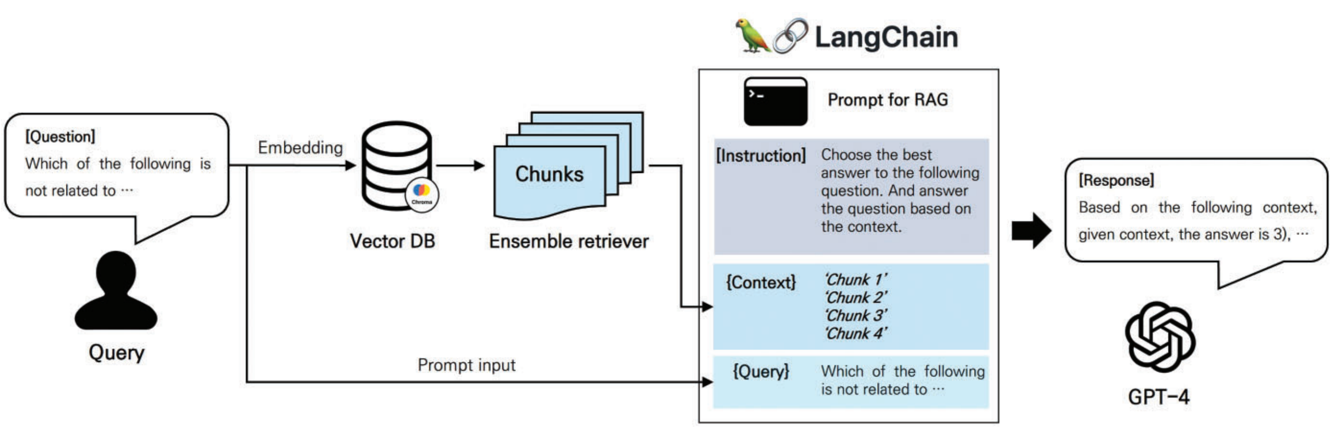

Figure 14: RAG process using ensemble retriever and GPT-4.

Table 10: Comparison of results for ChatGPT-4 and GPT-4 with RAG in the 'BIM guidelines' subcategory.

Model / Category (count)	ChatGPT-4			GPT-4 with RAG			Correct rate difference (GPT-4 with RAG - ChatGPT-4)
	# of correct answers	# of incorrect answers	Correct rate	# of correct answers	# of incorrect answers	Correct rate	
BIM guidelines (35)	22	13	62.9%	31	4	88.6%	▲ 25.7%

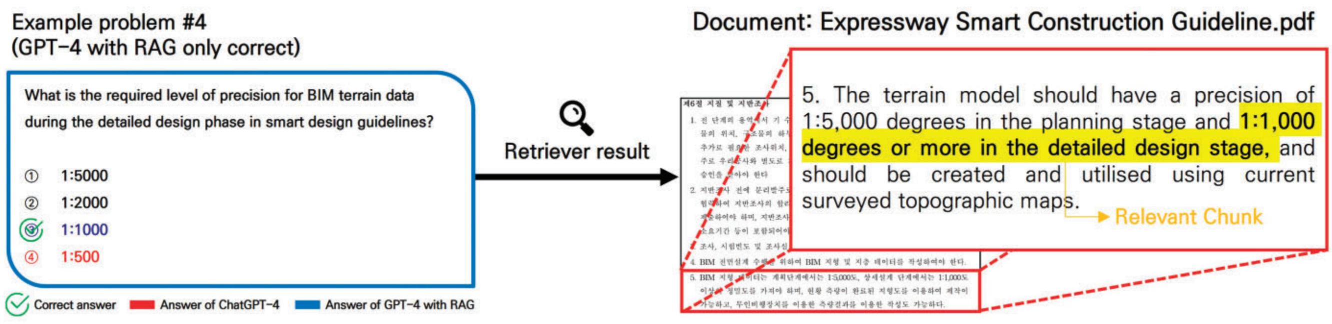

Figure 15: 'BIM guidelines' problem correctly answered by GPT-4 with RAG using relevant chunk.

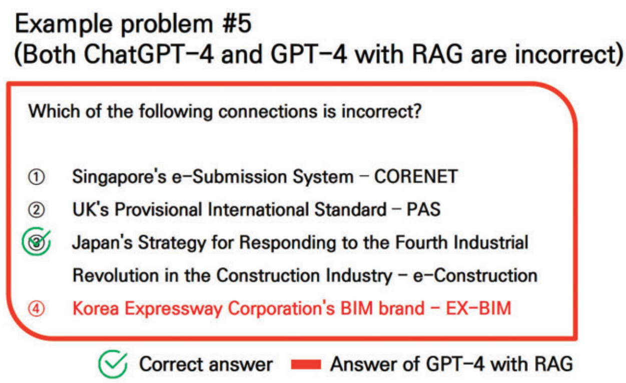

Figure 16: 'BIM guidelines' problem incorrectly answered by both ChatGPT-4 and GPT-4 with RAG.

6. Discussion

6.1 Implications of the results

6.1.1 *Performance of ChatGPT on the KBEE level 1 exam*

The results for Phase 1 of the study revealed that ChatGPT-4 overall showed superior performance to ChatGPT-3.5 in providing the correct answers to the multiple-choice section of the KBEE exam.

For the five years' worth of the annual exams, ChatGPT-4 scored an average of 85 points, 20 points higher than its predecessor. ChatGPT-4 alone passed every year of the exam. When comparing short-phrase versus full-sentence problems, ChatGPT-4 again achieved a higher score. Both versions performed better for short-phrase problems, although the difference was not particularly significant. Thus, we conclude that ChatGPT-4 is superior to its predecessor, and more importantly, also proficient in its knowledge of the BIM domain, at least enough to qualify for KBEE's level 1 exam.

However, it was difficult to pinpoint the reason behind ChatGPT-4's superior performance. As discussed, we rely on existing literature that attributes its performance to extensively larger parameters and training data sets (Baktash & Dawodi, 2023), as well as its improved multi-lingual capabilities (Koubaa, 2023).

In this regard, the subcategory analysis was most meaningful because it revealed ChatGPT-4 having varying levels of performance per subcategory. The variance in the results suggested that ChatGPT-4 did not have ample access to the required information for specific categories. Concretely, the 'Software functions' category was one of the most improved when compared with ChatGPT-3.5. However, ChatGPT-4 still incorrectly answered 19 of the 91 questions. While we suspected some of the wrong responses to translations of the technical nomenclature, inaccessibility to Korean-language manuals may also have contributed. The 'BIM guidelines' subcategory, which was the second highest contributor, also showed that limited access to domain-specific information, i.e. guidelines and roadmaps specific to Korea's BIM policies, was the underlying cause of the relatively lower performance.

These results suggested that providing relevant information to ChatGPT-4 would be required to further improve its level of understanding for specific areas of the BIM domain.

6.1.2 *RAG implementation for the 'BIM guidelines' subcategory*

Phase 2 of this study revealed RAG as a viable solution to improving GPT-4 for specific domain knowledge that may not be readily accessible. Specifically, the RAG framework could be used to enhance GPT-4's understanding of BIM regulations and standards stipulated in Korea. By providing the appropriate documents and designing an ensemble retriever to retrieve the relevant chunks assisted in improving the response quality of GPT-4 quantitatively by 25.7%.

However, the experiment also revealed the limitations of using the RAG framework. First, the source documents needed to be selected manually. It is difficult to determine *a priori* the scope of documents to be included and whether they are sufficient to represent the required knowledge of a particular section of a domain. Thus, the judgment of selecting the relevant documents may be

Table 11: Relevant chunks retrieved for example problem no. 5 by the ensemble retriever.

Relevant chunks rank	Retriever model	Documents	Chunks
1	BM25 (sparse retriever)	Appendix of Open BIM-based building design standards and infrastructure construction	The CORENET (COnstruction and Real Estate NETwork) system is a web-based construction administration processing system, overseen by Singapore's Building and Construction Authority, which connects the construction and IT sectors through the collaboration of 13 government agencies. The CORENET system is primarily composed of the CORENET e-Submission system, the CORENET e-PlanCheck system, and the CORENET e-Info system
2	BM25 (sparse retriever)	Appendix of Open BIM-based building design standards and infrastructure construction	The CORENET e-Submission system is a web-based platform that facilitates the sharing of documents related to business and government operations. The CORENET e-PlanCheck system is an automated drawing compliance checking system. Additionally, the CORENET e-Info system is an information-sharing platform that allows users to access a variety of information through a web environment
3	Chroma DB retriever (dense retriever)	Report of enhancement and substantiation plan for core technology of Open BIM-based architectural design	International trends show that, although knowledge management systems in the construction industry are less developed than in other sectors, they are more actively utilized abroad compared to domestic practices. Governments in advanced countries manage construction knowledge at a national level and promote the use of knowledge networks. Foreign companies adapt government frameworks to their needs, unlike in domestic industries, where systems focus on generating company profit. While institutions like Georgia Tech and Stanford University have compiled successful BIM cases, no comprehensive BIM database exists. Countries such as the US, Germany, the UK, and Singapore are working on developing BIM systems but have yet to systematically collect and utilize BIM knowledge
4	Chroma DB retriever (dense retriever)	Report of enhancement and substantiation plan for core technology of Open BIM-based architectural design	It is proposed to establish a knowledge management system, along with the development of a data warehouse system for knowledge generation, with the project duration set from 2019 to 2025 (Stage 2).

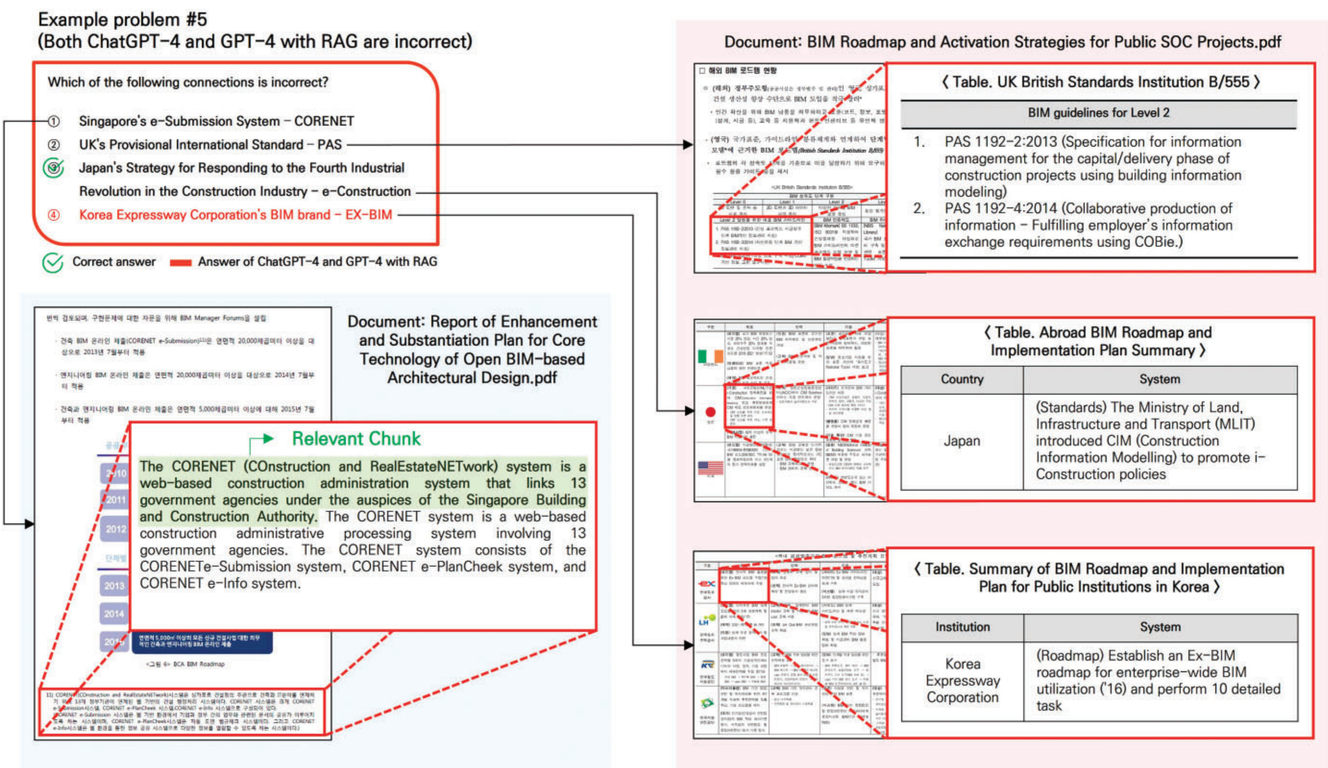

Figure 17: Documents in table format to be retrieved for each choice in example problem no. 5.

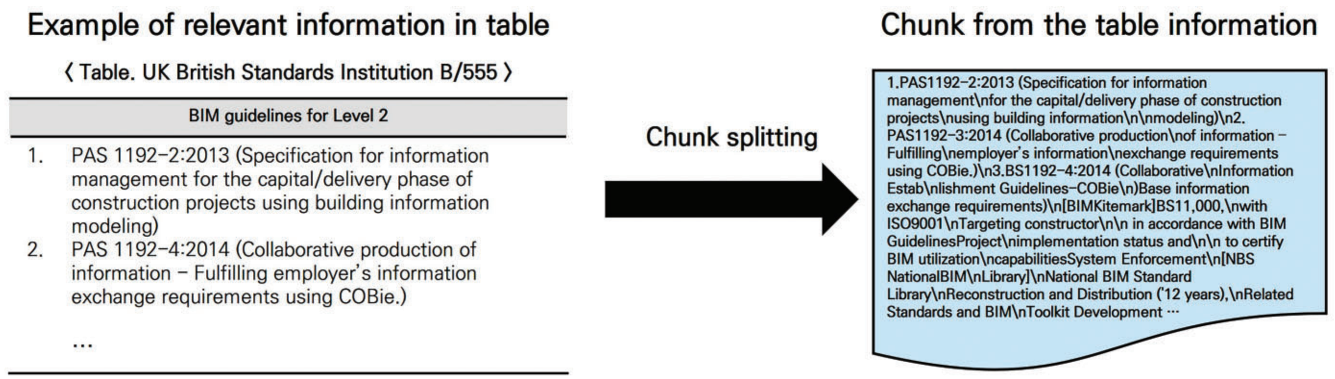

Figure 18: Example of table information converted into a chunk.

the most critical aspect for RAG adoption and this must be provided by domain specialists knowledgeable in the field.

Secondly, using the RAG framework meant that GPT-4 relied exclusively on the source documents provided for its responses (as to relying on its general knowledge). If the documents are reliable and up-to-date, GPT-4 will also produce accurate, high-quality responses. If not, it would provide the wrong or misleading answers. Case in point, the RAG-based GPT-4 still answered four problems in Figure 16 and Table B1 of Appendix B incorrectly, as the provided documents did not contain the required information for these problems. The exclusivity also meant that GPT-4 was dependent on the quality and particular format provided by the documents. Example problem no. 5 in Figure 17 showed that information in tabular formats, may not be properly embedded, leading to incorrect responses.

Despite these limitations, the results demonstrated RAG as a feasible solution for semantically searching BIM-related knowledge that may be specific to local or regional governing bodies. As countries and their institutions have different adoption timelines and policy roadmaps for BIM, using an RAG framework can ensure that correct responses can be attained for their related queries.

6.2 Limitations and future work

The main limitations of this study are recognized as follows:

- As ChatGPT was only tested on the multiple-choice section of the KBEE's level 1 exam, we do not state that the study has conclusively measured GPT's knowledge of the BIM domain. Moreover, its performance may have been placed at a disadvantage as the exam was administered in the Korean language, leading to potential losses in comprehension and mistranslations in technical terminologies. Nevertheless, the results of Phase 1 do provide a quantified measure of the two versions' performance, which may be used as a benchmark for comparing other LMMs or other BIM-related professional license exams.
- This study only tested versions 3.5 and 4 of ChatGPT, which were the two versions available at the time of the study. This approach is also comparable to the other studies that have tested ChatGPT on major professional exams, introduced in Section 2.2. More recent versions, i.e. GPT-4 Omni and o1-mini, are also online, and their performance has not been tested. Indeed, these versions are expected to perform better than their predecessors and may not even require the aid of supplying specific documents using RAG. However, the results again have meaning as they provide quantified results, and thus can be used to relatively compare and measure the performance of future versions as they become available. Such tests will still be necessary even with future versions, mainly as OpenAI does not divulge the scope of training data.
- RAG was only implemented for one of the subcategories, i.e. BIM guidelines. Other subcategories, most notably 'Software functions', were not tested. This subcategory was one of the most improved when compared between the two versions, but still had the highest number of incorrect answers. Implementing RAG for this category would have also shown the validity of improving ChatGPT's knowledge as well as providing additional insights. In this regard, this study focused more on showing RAG as a valid solution for providing GPT with domain-specific knowledge, and less on the cumulative improvement possible as a concrete contribution. Also, several design choices are available regarding the embedding model, the vector database (e.g. Chroma DB, Pinecone, etc.), and the sparse and dense retrievers. These alternatives were not evaluated in this study. Future work thus includes extending the RAG framework for other subcategories, experimenting with RAG design alternatives, as well as novel embedding and enhanced hybrid retrievers (Omrani *et al.*, 2024; Tang & Yang, 2024).

6.3 Applications of ChatGPT in the BIM knowledge domain

6.3.1 *ChatGPT as a catalyst for BIM adoption*

Despite the limitations, the study showed that overall ChatGPT was competent in its responses to the required areas for BIM expertise, and partial gaps in its knowledge could be overcome using the RAG framework. Moreover, the pace of improvement in the response accuracy of ChatGPT, and more broadly LLM's, is quickening year after year. Based on these assessments, the utility of ChatGPT in the BIM domain seems more likely to increase in the near future, and the probability of it becoming an integral tool for BIM should not be ignored.

In particular, the authors perceive ChatGPT as a potential catalyst for many of the issues that have been documented concerning the barriers to BIM adoption. Ullah *et al.* (2019) compiled a list of the barriers for BIM adoption faced in the AEC industry, and the following discusses how ChatGPT could be utilized to mitigate them.

- Lack of BIM experts

McAuley & Carroll (2017) noted the lack of BIM experts as a major barrier for BIM adoption. In practice, BIM experts, i.e. BIM managers and engineers, must be able to guide project stakeholders in the implementation of BIM-based project delivery processes throughout the lifecycle of a project. Perhaps more importantly, BIM experts also need to persuade the participants in the benefits of using BIM in comparison to existing conventions. Thus, as Saka & Chan (2020) noted, BIM experts need to be proficient in BIM-related knowledge (concepts, merits and demerits, implementation procedures, etc.) as well technical skills related to BIM tools.

One of the ways that owners, engineering firms, and contractors gain BIM implementation guidance is by hiring external BIM consultants. Consultants provide BIM guidelines, roadmaps and also assist clients throughout actual projects. However, relying on external expertise is costly and also does guarantee internalization of the BIM delivery process. The organization and its personnel's capability must be nurtured and matured which takes time and investment.

In this regard, ChatGPT, in the form of a Q&A service, could be deployed as a 'virtual consultant' for guiding the BIM delivery process. A Chatbot could be developed to respond to queries regarding BIM concepts, adoption standards and strategies, and implementation procedures. ChatGPT would utilize its understanding of the BIM categories identified in this study, i.e. BIM knowledge, applications, and the environment domain. A virtual consultant also has the benefit of being accessible by the entire organization, instead of the immediate personnel that may have direct contact with a human consultant.

- Inadequate training

Eadie *et al.* (2014) and Park & Kim (2017) noted inadequate training as a critical barrier to BIM adoption. Training personnel, together with software acquisition, is especially associated with the burden of costs in implementing BIM delivery processes (Shojaei al. *et al.*, 2023). Training costs can be significant to organizations, as ideally BIM should be adopted internally by all of its participants. Most specifically, training involves learning to use the multitude of specialized BIM software applications, each serving specific aspects of the design, construction, and management process. Many of these tools have steep learning curves making it a critical hurdle for practitioners in adopting BIM.

In this case, ChatGPT could be integrated as a virtual assistant for software technical guidance. For example, Jang *et al.* (2024) developed interfaces in Autodesk Revit in which users can directly request detailed modelling suggestions from ChatGPT. Such functionalities could be enhanced to query instructions for command sequences, alleviating the need to recall individual functions or shortcuts from memory. The assistance provided would also reduce the stress of individual training by experts conducted through lectures and/or workshops.

However, as shown in the results of this study, ChatGPT performed relatively poorly in the 'Software functions' subcategory. Thus, an RAG framework, in which manuals of the software is embedded and vectorized in a vector database, should ensure that its responses are accurate and up to date.

As a part of professional training, ChatGPT could alternatively be leveraged as a virtual tutoring assistant for the preparation of the certification of the exams addressed in Section 2.2. ChatGPT has the advantage of providing the answers as well as the sources and explanations of the questions. Furthermore, the questions from several of these exams could be used to train ChatGPT and create originally new questions. Again, customization using fine-tuning or RAG is recommended to ensure their quality and to minimize hallucinations.

- Complexity of the BIM model

Ahmed et al. (2014) noted the complexity of the BIM models as a hindrance to BIM adoption. Many studies endorse the use of BIM models as the main repository for knowledge management in their respective disciplines (Meadati & Irizarry, 2010; Banfi *et al.*, 2017). Consequently, BIM models can become data-intensive, as theoretically, it is able store or link to unlimited amounts of project data created during a project's life cycle.

In addition, the data in the form of entities, relations, and property sets in BIM models can be structured arbitrarily, making it difficult for relevant parties to determine its composition and specific location. ISO information management standards such as buildingSMART Data Dictionary and Information Delivery Specification exist but are not always abided to in practice.

Recent studies thus have introduced integrating ChatGPT with BIM models for information search and retrieval (Zheng & Fischer, 2023). The authors noted that the versatility of GPT in handling natural language (NL) queries while reducing the need for training as its major advantage.

ChatGPT's ability to generate structured query languages from NL text also makes it a powerful tool for information retrieval of databases used in conjunction with BIM models. For example, ChatGPT is able to generate SPARQL queries from text for ontologies, and similarly create Cypher queries for graph databases (Avila et al., 2024; Hornsteiner *et al.*, 2024). As the two data approaches are frequently used when converting BIM models into graph structures (Zhu *et al.*, 2024), ChatGPT provides a pivotal role in making them more readily accessible. For example, ChatGPT could be used to semantically search for similar design errors (e.g. clashes, etc.) which previously could only be done using literal or keyword searches.

- Additional workload

A practical reason for resistance to BIM is the additional set of workloads that is required of its delivery process. Notable examples include the generation of documents such as the Exchange Information Requirements (EIR), BIM Execution Plans (BEP), and BIM clash reports, 3D models with increasing LOD's in the design phase, and 4D/5D models in the construction phase. The deliverables become a burden particularly for transitional projects where the new requirements are requested in conjunction with conventional deliverables (e.g. 2D drawings and specifications). The additional workload has made countries such as Singapore and Korea provide subsidies or increase engineering fees for BIM mandatory projects (Liao *et al.*, 2021).

ChatGPT could be utilized to automate many of these deliverables. Representative examples are provided as follows:

- *Document generation*: The results of this study demonstrated ChatGPT to be knowledgeable about documentation related to BIM, and thus has the potential to be utilized to automate many of their generation. For example, BEP's are generated by the contractors based on the EIR's provided by the owner. ChatGPT could be trained to automate the generation of BEP's by first training it on data sets containing existing EIR and BEP pairs. The fine-tuned version could then be mobilized to automate the generation of a new BEP based on the EIR for a specific project.
- *Automated construction planning*: Although not tested in this study, ChatGPT has been known to be proficient in task planning, such as creating travel itineraries (Shin *et al.*, 2025) or cooking recipes (Değerli & Tatlisu, 2023). Such capabilities could be tailored for construction task planning and used to

automate the development of 4D BIM models. Indeed, Prieto *et al.* (2023) have shown ChatGPT was capable of creating a coherent schedule that follows a logical approach to fulfill the requirements of the scope indicated.

- *Automated 3D modelling*: The most intensive and time-consuming task is perhaps the actual 3D modelling of the BIM models. These models need to be created with higher levels of LOD as the project progresses through the design and construction phases. As aforementioned, ChatGPT could be incorporated into BIM authoring tools as a virtual assistant to guide users in the model creation process.

More advanced approaches are also being introduced in which ChatGPT is used to automate the modelling tasks itself. Lee et al. used LLM's to recommend the types of elements and their properties for detailed 3D modelling of interior walls. Du *et al.* (2024) went a step further by proposing an LLM-based multi-agent framework that translates natural language instructions into imperative code and invokes the BIM authoring tool's application programming interfaces (APIs), thereby generating editable 3D building models. Kim et al., introduced a novel approach for MEP routing by requesting ChatGPT to find optimal paths while specifying constraints and boundary conditions using advanced prompt engineering techniques.

These studies rely on ChatGPT's understanding of the software functions of BIM authoring tools as well as its knowledge of relations and constraints between model elements. The results of this study demonstrated ChatGPT's proficiency in this regard, albeit, requiring additional customization of its knowledge base.

6.3.2 *Application strategies (recommendations) for ChatGPT as a comprehensive and adaptive BIM knowledge framework (model)*

As discussed in the Introduction, this study posited that ChatGPT provides a novel knowledge management framework previously unavailable in the BIM domain. ChatGPT, and more broadly LLM's, differ from the existing knowledge models (e.g. expert systems) utilized for BIM or even the more recently developed deep learning or ontology-based models.

These models are developed by identifying and collecting the required data and structuring them in a way that make the data more accessible for querying and inference. This 'bottom-up' approach inevitably leads to these models being specialized for a specific subdomain. That is, they are expert systems developed to represent and solve a predefined set of targeted problems. For example, most expert systems in the BIM literature are developed exclusively for a subset of problems and are deliberate about delineating the targeted scope (Levchenko & Kashchenko, 2017). A similar assessment holds true for utilizing deep learning models with BIM. Most studies apply specific deep learning algorithms to solve singular classification tasks or auto-generate specialized outputs (Chen *et al.*, 2021; Lee *et al.*, 2022; Yu *et al.*, 2023). Even ontologies, whose primary objective is to link heterogeneous databases and thus generate richer content (Zhao & Qian, 2017), in practice are invariably used to represent and infer queries for a focused area of interest (Liu & Jiang *et al.*, 2016; Kang, 2022; Yin *et al.*, 2023).

By comparison, ChatGPT and more broadly LLM's, presupposes that information concerning a knowledge domain is already inherently learned, in effect removing the burden of collecting, structuring and training data for all possible areas within that domain. Thus, LLM's present a unique opportunity to deploy a 'top-down' approach, where the general knowledge base is provided and customizations (e.g. fine-tuning or RAG, etc.) for specific subdomains are performed contingent upon a need's basis. In other words, LLM's provide a comprehensive and adaptive knowledge framework previously unavailable for the BIM domain.

The results of this study confirm this proposition: ChatGPT was overall proficient in its understanding of the various areas in the BIM knowledge domain. Distinct areas of weaknesses could also be reinforced by providing specific subsets of data and deploying appropriate customization techniques (i.e. RAG). The issue is then determining when and which subareas should be customized. Based on the results of the study, the following recommendations are provided as guidance.

- Recommendation (1). Determine availability of BIM related information

If BIM-related knowledge is openly available on the Internet, ChatGPT could also utilize this as its source, and thus its responses are likely to be accurate without the need for customization. For example, ChatGPT performed well in regards to the KBEE questions related to international standards such as ISO 19650 and BS 1192. These documentations are openly available on their respective websites, and it is presumed that having access allows ChatGPT to respond coherently to their related queries.

- Recommendation (2). Conduct an initial proof-of-concept to inspect quality of responses

Alternatively, an initial investigation of the responses could promptly provide indications on the quality of the answers provided by ChatGPT. A sample of the KBEE problems was initially used to conduct a 'quick and dirty' test to discern its capabilities, which also motivated the need to conduct a more comprehensive study. A similar proof-of-concept could be conducted in-house within organizations to distinguish areas for customization.

- Recommendation (3). Consider customization per type of infrastructure or building

A rough and conservative criterion would be to consider customization per infrastructure or building type. At this level of distinction, the engineering and construction knowledge, as well as BIM implementation procedures and software's used, start to diverge. For example, for infrastructure projects, procedures for BIM implementation, and specialty software vary for Rail/Road/Dams, etc. Moreover, separate public institutions typically exist to govern these individual types, with their own set of policies and regulations regarding BIM.

7. Conclusions

This study evaluated the performance of ChatGPT-3.5, ChatGPT-4, and GPT-4 enhanced with RAG in addressing KBEE problems within the BIM domain. In the first phase, ChatGPT-4 demonstrated statistically significant improvements over ChatGPT-3.5, achieving passing scores across all tested years, and outperforming its predecessor in all KBEE subcategories. Category-wise, the 'Software functions' and the 'BIM guidelines' subcategories provided the highest contributions in improving the overall score of ChatGPT-4. Yet, these subcategories also still had the highest number of incorrect answers, as they lacked access to relevant information for their respective subcategories. Nevertheless, the results provided evidence that ChatGPT-4, albeit within the context of the KBEE, was overall knowledgeable in its understanding of the BIM domain. In the second phase, the integration of RAG with GPT-4 for the 'BIM guidelines' subcategory led to a substantial improvement, increasing its score to 88.6%, a gain of 25.7%. These findings illustrate RAG's effectiveness in complementing GPT-4 by providing it access to specialized, domain-specific information previously beyond its scope. The study verified that this capability was particularly valuable for knowledge categories that varied depending on local and regional contexts.

The study, however, was limited in several aspects. The experiments were conducted using a subsection of the KBEE and thus are not sufficient in assessing ChatGPT's overall knowledge of the BIM domain. Newer versions of ChatGPT which are online, were not tested. RAG was implemented for a single subcategory, and its design alternatives were not exhaustively investigated.

This study contributes by having quantitatively measured the extent of knowledge of ChatGPT for the BIM domain through an objective exam. The tests statistically proved ChatGPT-4's superior performance over its predecessor, while also identifying specific gaps in its knowledge, namely, the 'Software functions', and the 'BIM guidelines' subcategory. The design of an ensemble retriever to improve ChatGPT's responses demonstrated how such weaknesses could be bolstered using customization techniques, but also identified RAG's current limitations. These findings and methodology provide both insights and caveats for future researchers utilizing ChatGPT for their relevant applications. Moreover, the tests confirmed our hypothesis that ChatGPT, and more broadly LLMs, provides a novel knowledge management model, enabling users to use it comprehensively and adaptively for the BIM knowledge domain. Based on this assessment, the study proposed using ChatGPT as a virtual consultant and assistant for BIM procedures and software assistance, information retriever of complex BIM models, and for automating the additional work involved in the BIM delivery process. Finally, the study provided guidance and criteria for determining the deployment of customization techniques for ChatGPT's responses.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Author Contributions

Youngsu Yu: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing—original draft, Writing-review & editing. **Sihyun Kim**: Data curation, Formal analysis, Methodology, Software, Validation. **Wonbok Lee**: Data curation, Software, Visualization. **Bonsang Koo**: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Supervision, Writing-review & editing.

Funding

This work is supported in 2024 by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (grant RS-2021-KA163269).

Data Availability

The data underlying this article were provided by Korean BIM Evaluation Agency under license. Data will be shared on request to the corresponding author with permission of Korean BIM Evaluation Agency.

References

Ahmed, S. M., Emam, H. H., & Farrell, P. E. T. E. R. (2014). Barriers to BIM/4D implementation in Qatar *Smart, Sustainable and Healthy Cities*, **533**, 1–818. https://doi.org/10.13140/RG.2.1.4990.4164.

Allu, U., Ahmed, B., & Tripathi, V. (2024). Beyond extraction: Contextualising tabular data for efficient summarisation by language models, *preprint arXiv:2401.02333*. https://doi.org/10.48550/arXiv.2401.02333.

Amer, F., Jung, Y., & Golparvar-Fard, M. (2021). Transformer machine learning language model for auto-alignment of long-term and short-term plans in construction. *Automation in Construction,* **132**, 1–14. https://doi.org/10.1016/j.autcon.2021.103929.

Antaki, F., Touma, S., Milad, D., El-Khoury, J., & Duval, R. (2023). Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmology Science,* **3**, 1–7. https://doi.org/10.1016/j.xops.2023.100324.

Arabzadeh, N., Yan, X., & Clarke, C. L. A. (2021). Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection, *preprint arXiv:2109.10739v1*. https://doi.org/10.48550/arXiv.2109.10739.

Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A survey on RAG with LLMs. *Procedia Computer Science,* **246**, 3781–3790. https://doi.org/10.1016/j.procs.2024.09.178.

Avila, C. V. S., Vidal, V. M., Franco, W., & Casanova, M. A. (2024). Experiments with text-to-SPARQL based on ChatGPT. *In 2024 IEEE 18th International Conference on Semantic Computing (ICSC)*(pp. 277–284). Languna Hills, USA, https://doi.org/10.1109/ICSC59802.2024.00050.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI,* **7**, 52–62. https://doi.org/10.61969/jai.1337500.

Baktash, J. A., & Dawodi, M. (2023). Gpt-4: A review on advancements and opportunities in natural language processing, *preprint arXiv:2305.03195*. https://doi.org/10.48550/arXiv.2305.03195.

Balikas, G. (2023). Comparative analysis of open source and commercial embedding models for question answering. *In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*(pp. 5232–5233). New York, USA. https://doi.org/10.1145/3583780.3615994.

Banfi, F., Barazzetti, L., Previtali, M., & Roncoroni, F. (2017). Historic BIM: A new repository for structural health monitoring. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, **42**, 269–274. https://doi.org/10.5194/isprs-archives-XLII-5-W1-269-2017.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*(pp. 610–623). Virtual Conference. https://doi.org/10.1145/3442188.3445922.

Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature,* **624**, 570–578. https://doi.org/10.1038/s41586-023-06792-0.

Bommarito, M., & Katz, D. M. (2022). GPT takes the bar exam, *preprint arXiv:2212.14402*. https://doi.org/10.48550/arXiv.2212.14402.

Callanan, E., Mbakwe, A., Papadimitriou, A., Pei, Y., Sibue, M., Zhu, X., Ma, Z., Liu, X., & Shah, S. (2023). Can GPT models be financial analysts? An evaluation of ChatGPT and GPT-4 on mock CFA exams, *preprint arXiv:2310.08678*. https://doi.org/10.48550/arXiv.2310.08678.

Cenić, A., Jovanović, M., & Stojković, S. (2024). The Serbian retrieval-augmented generation system based on hybrid search. *In 2024 International Conference Automatics and Informatics (ICAI)* (pp. 420–424). Varna, Bulgaria. https://doi.org/10.1109/ICAI63388.2024.10851665.

Chan, A. (2023). GPT-3 and InstructGPT: Technological dystopianism, utopianism, and "contextual" perspectives in AI ethics and industry. *AI and Ethics*, **3**, 53–64. https://doi.org/10.1007/s43681-022-00148-6.

Chen, B., Liu, Q., Chen, H., Wang, L., Deng, T., Zhang, L., & Wu, X. (2021). Multiobjective optimization of building energy consumption based on BIM-DB and LSSVM-NSGA-II. *Journal of Cleaner Production*, **294**, 420–424. https://doi.org/10.1016/j.jclepro.2021.126153.

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2022). ChatGPT goes to law school. *Journal of Legal Education*, **71**, 21–29. https://doi.org/10.2139/ssrn.4335905.

Choi, J., Koo, B., Yu, Y., Jung, Y., & Ham, N. (2023). Evaluating ChatGPT's competency in BIM-related knowledge via the Korean BIM expertise exam. *KIBIM Magazine*, **13**, 21–29. https://doi.org/10.13161/kibim.2023.13.3.021.

Chroma (2024). *Chroma Documentation*. https://docs.trychroma.com/(18 July 2024, date last accessed).

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, **51**, 14–27. https://doi.org/10.3758/s13428-018-1142-4.

Değerli, A. H., & Tatlisu, N. B. (2023). Cooking with ChatGPT and Bard: A study on competencies of AI tools on recipe correction, adaptation, time management, and presentation. *Journal of Tourism & Gastronomy Studies*, **11**, 2658–2673. 10.21325/jotags.2024.1312.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805.

Du, C., Esser, S., Nousias, S., & Borrmann, A. (2024). Text2BIM: Generating building models using a large language model-based multi-agent framework, *preprint arXiv:2408.08054*.https://doi.org/10.48550/arXiv.2408.08054.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Alpacafarm: A simulation framework for methods that learn from human feedback, *preprint arXiv:2305.14387*. https://doi.org/10.48550/arXiv.2305.14387.

Eadie, R., Odeyinka, H., Browne, M., Mahon, C., & Yohanis, M. (2014). Building information modelling adoption: An analysis of the barriers of implementation. *Journal of Engineering and Architecture*, **2**, 77–101. https://pure.ulster.ac.uk/files/85735072/EadieetalBIMFinal.pdf.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: evaluating neural toxic degeneration in language models, *preprint arXiv:2009.11462*. https://doi.org/10.48550/arXiv.2009.11462.

George, A. S., George, A. S. H., & Martin, A. S. G. (2023). A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal*, **1**, 9–23. https://doi.org/10.5281/zenodo.7644359..

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Hassan, S. Z., Shoman, M., Wu, J., Mirjalili, S., & Shah, M. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*, **3**, 1–30. https://doi.org/10.1061/JMENEA.MEENG-6444.

He, C., He, W., Liu, M., Leng, S., & Wei, S. (2025). Enriched construction regulation inquiry responses: A hybrid search approach for large language models. *Journal of Management in Engineering*, **41**, 1–15. https://doi.org/10.1061/JMENEA.MEENG-644.

Hornsteiner, M., Kreussel, M., Steindl, C., Ebner, F., Empl, P., & Schönig, S. (2024). Real-time text-to-cypher query generation with large language models for graph databases. *Future Internet*, **16**, 1–26. https://doi.org/10.3390/fi16120438.

Huang, D., Wei, Z., Yue, A., Zhao, X., Chen, Z., Li, R., Jiang, K., Chang, B., Zhang, Q., Zhang, S., & Zhang, Z. (2023). DSQA-LLM: Domain-specific intelligent question answering based on large language models. *AI-generated content. Communications in Computer and Information Science*, **1946**, 170–180. https://doi.org/10.1007/978-981-99-7587-7_14.

Hwang, J., Song, S. H., Lee, C., Ahn, H., Cho, H., & Kang, K.-I. (2024). BIM and OpenAI-based model for supporting initial construction planning of bridges and tunnels. *Korean Journal of Construction Engineering and Management*, **25**, 24–33. https://doi.org/10.6106/KJCEM.2024.25.6.024.

Jang, S., Lee, G., Oh, J., Lee, J., & Koo, B. (2024). Automated detailing of exterior walls using NADIA: natural-language-based architectural detailing through interaction with AI. *Advanced Engineering Informatics*, **61**, 1–18. https://doi.org/10.1016/j.aei.2024.102532.

Jeong, C. (2024). Fine-tuning and utilization methods of domain-specific LLMs, *preprint arXiv:2401.02981*. https://doi.org/10.48550/arXiv.2401.02981.

Jo, H., Lee, J., Lee, Y., & Choo, S. (2024). Generative artificial intelligence and building design: Early photorealistic render visualization of façades using local identity-trained models. *Journal of Computational Design and Engineering*, **11**, 85–105. https://doi.org/10.1093/jcde/qwae017.

Juvekar, K., & Purwar, A. (2024). Introducing a new hyper-parameter for RAG: context window utilization, *preprint arXiv:2407.19794*. https://doi.org/10.48550/arXiv.2407.19794.

Kang, T. W. (2022). Ontology BIM-based knowledge service framework architecture development. *KIBIM Magazine*, **12**, 52–60. https://doi.org/10.13161/kibim.2022.12.4.052.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering, *preprint arXiv:2004.04906*. https://doi.org/10.48550/arXiv.2004.04906.

Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y., & Radev, D. (2023). Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations, *preprint arXiv:2303.18027*. https://doi.org/10.48550/arXiv.2303.18027.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, **382**, 1–17. https://doi.org/10.1098/rsta.2023.0254.

Kim, A., Muhn, M., & Nikolaev, V. (2024). Financial statement analysis with large language models, *preprint arXiv:2407.17866*. https://doi.org/10.48550/arXiv.2407.17866.

Koo, B., Jung, R., Yu, Y., & Kim, I. (2021). A geometric deep learning approach for checking element-to-entity mappings in infrastructure building information models. *Journal of Computational Design and Engineering*, **8**, 239–250. https://doi.org/10.1093/jcde/qwaa075.

Korea Expressway Corporation (2020). *Expressway Smart Design Guidelines*. Korea Expressway Corporation. https://smartconstruction.kr/storage/app/images/823d0c2b36aecd6746d7533035cc6806.pdf. (12 December 2024, date last accessed).

Korea Expressway Corporation (2021). *Expressway BIM Design Evaluation Criteria*. Korea Expressway Corporation, Korea.

https://www.bimkorea.or.kr/main/download.jspid=491&ek=559cb990c9dffd8675f6bc2186971dc2. (12 December 2024, date last accessed).

Korean Institute of BIM (KIBIM) (2021). *BIM Educational Resources*. KIBIM. https://www.bimkorea.or.kr/main/download.jsp?id=283&ek=0f49c89d1e7298bb9930789c8ed59d48.

Koubaa, A. (2023). GPT-4 vs. GPT-3.5: a concise showdown. *TechRxiv*. https://doi.org/10.36227/techrxiv.22312330.v2. (7 October 2024, date last accessed).

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health* **2**, 1–17. https://doi.org/10.1371/journal.pdig.0000198.

Lam, H. Y. I., Ong, X. E., & Mutwil, M. (2024). Large language models in plant biology. *Trends in Plant Science*, **29**, 1145–1155. https://doi.org/10.1016/j.tplants.2024.04.013.

LangChain (2024). *Build a Retrieval Augmented Generation (RAG) App*. https://python.langchain.com/docs/tutorials/rag/. (19 July 2024, date last accessed).

Lecler, A., Duron, L., & Soyer, P. (2023). Revolutionizing radiology with GPT-based models: Current applications, future possibilities, and limitations of ChatGPT. *Diagnostic and Interventional Imaging*, **104**, 269–274. https://doi.org/10.1016/j.diii.2023.02.003.

Lee, J., Yoo, Y., & Cha, S. (2024). Generative early architectural visualizations: Incorporating architect's style-trained models. *Journal of Computational Design and Engineering*, **11**, 40–59. https://doi.org/10.1093/jcde/qwae065.

Lee, W., Kim, S., Yu, Y., & Koo, B. (2022). Development of graph-based deep learning methods for enhancing the semantic integrity of spaces in BIM models. *Korean Journal of Construction Engineering and Management*, **23**, 45–55. https://doi.org/10.6106/KJCEM.2022.23.3.045.

Lemaire, C., Rivest, L., Boton, C., Danjou, C., Braesch, C., & Nyffenegger, F. (2019). Analyzing BIM topics and clusters through ten years of scientific publications. *Journal of Information Technology in Construction*, **24**, 273–298. https://www.itcon.org/2019/15.

Levchenko, O., & Kashchenko, T. (2017). Expert systems in the BIM environment. *INTERSECTII/INTERSECTIONS*, **14**, 85–94. http://www.intersections.tuiasi.ro/index.php/JIIR.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W. T., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks, *preprint arXiv:2005.11401v4*. https://doi.org/10.48550/arXiv.2005.11401.

Liao, L., Teo, E. A. L., Li, L., Zhao, X., & Wu, G. (2021). Reducing non-value-adding BIM implementation activities for building projects in Singapore: Leading causes. *Journal of Management in Engineering*, **37**, 1–17. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000900.

Liu, X., Li, Z., & Jiang, S. (2016). Ontology-based representation and reasoning in building construction cost estimation in China. *Future Internet*, **8**, 1–22. https://doi.org/10.3390/fi8030039.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, **1**, 1–22. https://doi.org/10.1016/j.metrad.2023.100017.

Mandikal, P., & Mooney, R. (2024). Sparse meets dense: a hybrid approach to enhance scientific document retrieval, *preprint arXiv:2401.04055*. https://doi.org/10.48550/arXiv.2401.04055.

Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, **1**, 1–24. https://doi.org/10.1007/s10506-024-09396-9.

McAuley, B., & Carroll, P. (2017). Establishing the key pillars of innovation required to execute a successful BIM strategy within a construction SME in Ireland. *In Proceedings of the 3rd CitA BIM Gathering in Dublin*. Dublin, Ireland. https://doi.org/10.21427/r42p-3×25.

Meadati, P., & Irizarry, J. (2010). BIM–a knowledge repository. *In Proceedings of the 46th Annual International Conference of the Associated Schools of Construction*, Vol. **12**, Boston, USA.

Ministry of Land, Infrastructure and Transport (MOLIT) (2016). *Report of Enhancement and Substantiation Plan for Core Technology of Open BIM-based Architectural Design*. MOLIT. https://www.kaia.re.kr/portal/landmark/readTskFinalView.dotskId=101186&yearCnt=1&menuNo=200100. (19 November 2024, date last accessed).

Ministry of Land, Infrastructure and Transport (MOLIT) (2017). *Appendix of Open BIM-based Building Design Standards and Infrastructure Construction*. MOLIT. https://dl.nanet.go.kr/SearchDetailView.do?cn=MONO1201717586. (19 November 2024, date last accessed).

Ministry of Land, Infrastructure and Transport (MOLIT) (2018). *BIM Roadmap and Activation Strategies for Public SOC Projects*. MOLIT. https://www.codil.or.kr/viewDtlConRpt.do?gubun=rpt&pMetaCode=OTKCRK190549. (19 November 2024, date last accessed).

Ministry of Land, Infrastructure and Transport (MOLIT) (2020a). *Architect's Scope of Work and Consideration for Public Works Projects*. MOLIT. https://www.molit.go.kr/USR/I0204/m_45/dtl.jspgubun=&search=&search_dept_id=&search_dept_nm=&old_search_dept_nm=&psize=10&search_regdate_s=&search_regdate_e=&srch_usr_nm=&srch_usr_num=&srch_usr_year=&srch_usr_titl=&srch_usr_ctnt=&lcmspage=1196&idx=4652. (19 November 2024, date last accessed).

Ministry of Land, Infrastructure and Transport (MOLIT) (2020b). 2030 *architecture BIM Activation Roadmap*. MOLIT. https://www.molit.go.kr/USR/NEWS/m_71/dtl.jsp?lcmspage=1&id=95084979. (19 November 2024, date last accessed).

Ministry of Land, Infrastructure and Transport (MOLIT) (2020c). *Basic Guidelines for BIM in the Construction Industry*. MOLIT. https://www.molit.go.kr/USR/policyData/m_34681/dtl.jsp?search=&srch_dept_nm=&srch_dept_id=&srch_usr_nm=&srch_usr_titl=Y&srch_usr_ctnt=&search_regdate_s=&search_regdate_e=&psize=10&s_category=&p_category=&lcmspage=1&id=4516. (19 November 2024, date last accessed).

Ministry of Land, Infrastructure and Transport (MOLIT) (2022a). *Construction Industry BIM Stipulation Guidelines (owner, contractor, designer)*. MOLIT. https://www.molit.go.kr/USR/policyData/m_34681/dtl.jsp?search=&srch_dept_nm=&srch_dept_id=&srch_usr_nm=&srch_usr_titl=Y&srch_usr_ctnt=&search_regdate_s=&search_regdate_e=&psize=10&s_category=&p_category=&lcmspage=1&id=4634. (19 November 2024, date last accessed).

Ministry of Land, Infrastructure and Transport (MOLIT) (2022b). *S-Construction 2030*. Korea. MOLIT, https://www.molit.go.kr/USR/NEWS/m_71/dtl.jsp?lcmspage=7&id=95086971. (19 November 2024, date last accessed).

Omrani, P., Hosseini, A., Hooshanfar, K., Ebrahimian, Z., Toosi, R., & Akhaee, M. A. (2024). Hybrid retrieval-augmented generation approach for LLMs query response enhancement. *In Proceedings of the 10th International Conference on Web Research*, pp. 22–26. Tehran, Iran. https://doi.org/10.1109/ICWR61162.2024.10533345.

OpenAI (2022). New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model/. (3 July 2024, date last accessed).

OpenAI (2023). GPT-4 technical report, *arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774.

Oztermeli, A. D., & Oztermeli, A. (2023). ChatGPT performance in the medical specialty exam: an observational study. *Medicine*, **102**, 1–5. https://doi.org/10.1097/MD.0000000000034673.

Pan, F., Canim, M., Glass, M., Gliozzo, A., & Hendler, J. (2022). End-to-end table question answering via retrieval-augmented generation, *preprint arXiv:2203.16714*. https://doi.org/10.48550/arXiv.2203.16714.

Park, K. S., & Kim, K. P. (2017). BIM application and adoption in the UK housing sector,*Integrated Building Information Modelling*. pp. 46–81. Bentham eBooks, https://doi.org/10.2174/97816810845721170101.

Prieto, S. A., Mengiste, E. T., & García de Soto, B. (2023). Investigating the use of ChatGPT for the scheduling of construction projects. *Buildings*, **13**, 1–16. https://doi.org/10.3390/buildings13040857.

Public Procurement Service (PPS) (2022). *Guidelines for Applying BIM for Facilities Projects*. PPS. https://www.pps.go.kr/kor/bbs/view.do?key=00324&bbsSn=2212280034. (19 November 2024, last date accessed).

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics,* **11**, 1316–1331. https://doi.org/10.1162/tacl_a_00605.

Rane, N., Choudhary, S., & Rane, J. (2023). Integrating Building Information Modelling (BIM) with ChatGPT, Bard, and similar generative artificial intelligence in the architecture, engineering, and construction industry: applications, a novel framework, challenges, and future scope. *SSRN*. https://doi.org/10.2139/ssrn.4645601.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval,* **3**, 333–389. https://doi.org/10.1561/1500000019.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications, *preprint arXiv:2402.07927*. https://doi.org/10.48550/arXiv.2402.07927.

Saka, A. B., & Chan, D. W. (2020). Knowledge, skills and functionalities requirements for quantity surveyors in building information modelling (BIM) work environment: an international Delphi study. *Architectural Engineering and Design Management*, **16**, 227–246. https://doi.org/10.1080/17452007.2019.1651247.

Seoul Institute of Technology (2022). *A Study on the Guideline and Roadmap for Applying BIM to Seoul*. Seoul Institute of Technology. https://www.sit.re.kr/upload/result/20221019193734898706.pdf. (18 November 2024, date last accessed).

Shin, S., Kim, J., Lee, E., Yhee, Y., & Koo, C. (2025). ChatGPT for trip planning: The effect of narrowing down options. *Journal of Travel Research,* **64**, 247–266. https://doi.org/10.1177/00472875231214196.

Shojaei, R. S., Oti-Sarpong, K., & Burgess, G. (2023). Leading UK construction companies' strategies to tackle BIM training and skills challenges. *International Journal of Construction Education and Research,* **19**, 383–404. https://doi.org/10.1080/15578771.2022.2123071.

Singh, A. K., Pal, A., Kumar, P., Lin, J. J., & Hsieh, S. H. (2023). Prospects of integrating BIM and NLP for automatic construction schedule management. *In ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, Vol. **40**, pp. 238–245. Chennai, India, https://doi.org/10.22260/ISARC2023/0034.

Strong, E., DiGiammarino, A., Weng, Y., Basaviah, P., Hosamani, P., Kumar, A., Nevins, A., Kugler, J., Hom, J., & Chen, J. H. (2023). Performance of ChatGPT on free-response, clinical reasoning exams. *MedRxiv*, 1–14. https://doi.org/10.1101/2023.03.24.23287731. medRxiv.

Takagi, S., Watari, T., Erabi, A., & Sakaguchi, K. (2023). Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: Comparison study. *JMIR Medical Education*, **9**, 1–6. https://doi.org/10.2196/48002.

Tang, Y., & Yang, Y. (2024). Do we need domain-specific embedding models? An empirical investigation, *preprint arXiv:2409.18511*. https://doi.org/10.48550/arXiv.2409.18511.

Ullah, K., Lill, I., & Witt, E. (2019). An overview of BIM adoption in the construction industry: Benefits and barriers. *In* Lill I., & Witt E. (Eds.), *10th Nordic Conference on Construction Economics and Organization (Emerald Reach Proceedings Series*, Vol. **2**. pp. 297–303. Emerald Publishing Limited, Tallinn, Estonia. https://doi.org/10.1108/S2516-285320190000002052.

Wang, C., Long, Q., Meng, X., Cai, X., Wu, C., Meng, Z., Wang, X., & Zhou, Y. (2024). BioRAG: a RAG-LLM framework for biological question reasoning, *preprint arXiv:2408.01107*. https://doi.org/10.48550/arXiv.2408.01107.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models, *preprint arXiv:2211.01910*. https://doi.org/10.48550/arXiv.2201.11903.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2023). Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data,* **18**, 1–32. https://doi.org/10.48550/arXiv.2304.13712.

Yepes, A. J., You, Y., Milczek, J., Laverde, S., & Li, R. (2024). Financial report chunking for effective retrieval-augmented generation, *preprint arXiv:2402.05131*. https://doi.org/10.48550/arXiv.2402.05131.

Yin, M., Tang, L., Webster, C., Xu, S., Li, X., & Ying, H. (2023). An ontology-aided, natural language-based approach for multi-constraint BIM model querying. *Journal of Building Engineering,* **76**, 1–23. https://doi.org/10.1016/j.jobe.2023.107066.

Yu, Y., Ha, D., Lee, K., Choi, J., & Koo, B. (2022). ArchShapesNet: a novel dataset for benchmarking architectural building information modeling element classification algorithms. *Journal of Computational Design and Engineering*, **9**, 1449–1466. https://doi.org/10.1093/jcde/qwac064.

Yu, Y., Kim, S., Lee, W., & Koo, B. (2023). Ensemble-based deep learning approach for performance improvement of BIM element classification. *KSCE Journal of Civil Engineering,* **27**, 1898–1915. https://doi.org/10.1007/s12205-023-2331-y.

Zhang, B., Yang, H., Zhou, T., Babar, A., & Liu, X. Y. (2023a). Enhancing financial sentiment analysis via retrieval-augmented large language models. *In Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 349–356. New York, USA. https://doi.org/10.1145/3604237.3626866.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhang, E., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023b). Siren's song in the AI ocean: a survey on hallucination in large language models, *preprint arXiv:2309.01219*. https://doi.org/10.48550/arXiv.2309.01219.

Zhao, S., & Qian, Q. (2017). Ontology-based heterogeneous materials database integration and semantic query. *AIP Advances*, **7**, 1–18. https://doi.org/10.1063/1.4999209.

Zheng, J., & Fischer, M. (2023). BIM-GPT: a prompt-based virtual assistant framework for BIM information retrieval, *preprint arXiv:2304.09333*. https://doi.org/10.48550/arXiv.2304.09333.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers, *preprint arXiv:2211.01910*. https://doi.org/10.48550/arXiv.2211.01910.

Zhu, J., Nisbet, N., Yin, M., Wei, R., & Brilakis, I. (2024). Cypher4BIM: Releasing the power of graph for building knowledge discovery, *preprint arXiv:2405.16345*. https://doi.org/10.48550/arXiv.2405.16345.

Appendix

A. The incorrect problems of ChatGPT-4 in the 'bim guidelines' subcategory

The 'BIM guidelines' subcategory showed a higher correct rate with ChatGPT-4 compared to ChatGPT-3.5, yet it still had the lowest correct rate (62.90%). Table A1 lists the problems that ChatGPT-4 answered incorrectly in this subcategory.

Table A1: Incorrectly answered problems by ChatGPT-4 in the 'BIM guidelines' subcategory.

No.	Problems	Correct answer
1	Q: What is the BIL for the planning and design phase according to the 'Basic Guideline for Applying BIM for Facility Projects v2.0' published by Public Procurement Service (PPS)? 1) BIL10 2) BIL15 3) BIL20 4) BIL30	3
2	Q: Which of the following is incorrectly linked to the BIM standards or guidelines of major public institutions in Korea? 1) Public Procurement Service (PPS): Basic Guidelines for Applying BIM in Facility Projects 2) MOLIT: Basic Guidelines for BIM in the Construction Industry 3) Korea Expressway Corporation: Smart Design Guidelines for Highways 4) Korea National Railway: Civil-BIM Work Guidelines	4
3	Q: What is the required level of precision for BIM terrain data during the detailed design phase in the 'Expressway Smart Design Guidelines'? 1) 1:5000 2) 1:2000 3) 1:1000 4) 1:500	3
4	Q: Which of the following is not an appropriate method of security management when conducting a project based on BIM in public works? 1) Network separation 2) Access control 3) Digital signatures and approvals 4) Model security	1
5	Q: Which of the following is incorrectly matched regarding the LOD and its application stages and content according to the 'Basic Guidelines for BIM in the Construction Industry'? 1) Level of Development 100: Representation of forms required at the basic (planning) design stage 2) Level of Development 300: Representation of the existence of all elements required at the detailed (low) design stage 3) Level of Development 400: Representation of all elements available at the construction stage 4) Level of Development 500: Information available for use in maintenance and management stages	1
6	Q: Which of the following is the least appropriate description regarding the BIL according to the 'Basic Guideline for Applying BIM for Facility Projects v2.0' published by Public Procurement Service (PPS)? 1) BIL20: Representation of shapes required at the planning design level 2) BIL30: Representation of all elements required at the basic design level 3) BIL40: Representation of all elements required at the detailed design level 4) BIL50: The level of detail varies according to client requirements, such as maintenance	4
7	Q: Which of the following is not related to information management in BIM collaboration as specified in the 'Basic Guidelines for BIM in the Construction Industry'? 1) Model management 2) History management 3) Schedule management 4) Cost management	2

Table A1: Continued

No.	Problems	Correct answer
8	Q: Which level corresponds to the application guidelines that, by sector, each contracting authority prepares tailored to the type of project or characteristics of the contracting authority, according to the basic BIM guidelines and subguidelines in the construction industry? 1) Level 1–1 2) Level 1–2 3) Level 2–1 4) Level 2–2	3
9	Q: Which of the following statements is incorrect regarding the MOLIT's plan for mandatory BIM adoption throughout the entire lifecycle of public construction projects? 1) BIM adoption will begin in the road sector for projects over 100 billion KRW, starting in late 2022 2) Sequential adoption will follow in railways and buildings (2023), and in rivers and ports (2024) 3) In 2026, BIM requirements will be expanded to projects over 50 billion KRW 4) In 2028, BIM requirements will be expanded to projects over 10 billion KRW	4
10	Q: Which of the following connections is incorrect? 1) Singapore's e-Submission System – CORENET 2) UK's Provisional International Standard – PAS 3) Japan's Strategy for Responding to the Fourth Industrial Revolution in the Construction Industry – e-Construction 4) Korea Expressway Corporation's BIM brand – EX-BIM	3
11	Q: Which of the following is NOT an appropriate description for the basic highway design in the 'Expressway BIM Design Evaluation Criteria'? 1) Implementation of BIM-based design 2) BIM mode'lling of route planning 3) Review of alternative routes 4) Smart design	2
12	Q: Which of the following is NOT a major revision in the Public Procurement Service's 'Basic guidelines for applying BIM for facilities projects' as of December 2022? 1) Excluding quantity data preparation for architectural and structural elements in the planning phase 2) Expanding the scope of modelling for MEP (Mechanical, Electrical, Plumbing) fields 3) Expanding the scope of BIM utilization during the construction phase 4) Establishment of procedures for the owner agency to present design focal points reflecting project characteristics	2
13	Q: Which of the following is NOT an appropriate description of the policies and support plans related to the adoption of BIM by major institutions? 1) Korea Expressway Corporation: EX BIM Roadmap (2017) 2) Korea Land and Housing Corporation: LH Civil-BIM Roadmap (2018) 3) Korea Water Resources Corporation: BIM Infrastructure Maintenance Master Plan (2012) 4) Korea National Railway: Railway BIM Roadmap 2030 (2018)	3

*All questions and multiple-choices in the table are originally in Korean and have been translated into English by the authors for readability.

B. The incorrect problems of ChatGPT-4 with rag in the 'bim guidelines' subcategory

ChatGPT-4 significantly improved its accuracy with the application of RAG, but it still produced incorrect answers for certain questions. Table B1 presents the questions that ChatGPT-4 with RAG continued to answer incorrectly.

Table B1: Incorrectly answered problems by ChatGPT-4 with RAG in the 'BIM guidelines' subcategory.

No.	Question	Correct answer	Answer of GPT-4 with RAG
1	Q: Which of the following is NOT an appropriate description for the basic highway design in the 'Expressway BIM Design Evaluation Criteria'? 1) Implementation of BIM-based design 2) BIM modelling of route planning 3) Review of alternative routes 4) Smart design	2	4
2	Q: Which of the following is NOT a major revision in the Public Procurement Service's 'Basic guidelines for applying BIM for facilities projects' as of December 2022? 1) Excluding quantity data preparation for architectural and structural elements in the planning phase 2) Expanding the scope of modelling for MEP (Mechanical, Electrical, Plumbing) fields 3) Expanding the scope of BIM utilization during the construction phase 4) Establishment of procedures for the owner agency to present design focal points reflecting project characteristics	2	No answer
3	Q: Which of the following is NOT an appropriate description of the policies and support plans related to the adoption of BIM by major institutions? 1) Korea Expressway Corporation: EX BIM Roadmap (2017) 2) Korea Land and Housing Corporation: LH Civil-BIM Roadmap (2018) 3) Korea Water Resources Corporation: BIM Infrastructure Maintenance Master Plan (2012) 4) Korea National Railway: Railway BIM Roadmap 2030 (2018)	3	2

*All questions and multiple-choices in the table are originally in Korean and have been translated into English by the authors for readability.

Received: December 16, 2024. **Revised:** March 10, 2025. **Accepted:** March 11, 2025