ENGINEERING Management 2026, 13(1): 124–145
https://doi.org/10.1007/s42524-026-4237-0

RESEARCH ARTICLE

Yongcheng FU, Tong LYU, Yongqiang CHEN, Ziyun LV

Q&A system for international construction contracts driven by large language model and knowledge graph

Abstract As international construction projects continue to expand, construction enterprises are accumulating vast amounts of contract-related text data, making the effective management and extraction of knowledge from these dense texts essential to mitigate knowledge loss and ensure efficient contract management. The advent of large language models (LLMs) presents a promising avenue for enhancing contract knowledge management through intelligent systems. However, challenges such as hallucination, inflexibility, and lack of interpretability often diminish practitioners' confidence in applying these models to real-world scenarios. This study seeks to develop a knowledge-based question-and-answer (Q&A) system for international construction contracts by integrating both the knowledge graph (KG) and the LLM. Built upon a domain-specific KG derived from the 2022 edition of the Fédération Internationale des Ingénieurs-Conseils (FIDIC) Yellow Book and the NEC4 Conditions of Contract, the system leverages LLM to conduct synergistic reasoning with the KG, enabling it to answer complex queries using both tacit knowledge and external sources. Experimental results demonstrate that the proposed approach markedly enhances the model's performance in Q&A tasks of contract knowledge, achieving an average success rate exceeding 87% in terms of both accuracy and interpretability. This model provides a specialized Q&A system for international construction enterprises, facilitating flexible knowledge acquisition and task-oriented analysis in contract management, while also introducing a novel framework for integrating AI technologies into the management of international construction contracts.

Keywords international construction contract, knowledge graph, large language model, knowledge management

1 Introduction

In international construction projects, collaborating parties mitigate and govern transactional hazards through the design and negotiation of complex international construction contracts (Chen et al., 2018). Effective contract management is pivotal to ensuring the success of a project by mitigating cost overruns, minimizing schedule delays, and reducing the occurrence of claims and disputes (Zheng et al., 2025). During the contract review phase, contract managers analyze extensive contract documents, categorize clauses based on departmental relevance, and conduct related activities such as risk assessment, inconsistency detection, and the identification of semantic ambiguities. These processes enhance the understanding of contractual documents and reduce errors and omissions (Hassan et al., 2021). In the contract execution phase, contract managers monitor events that may lead to claims, conduct investigations, analyze evidence, and execute claim procedures or initiate dispute resolution processes in accordance with the contract terms (Kalogeraki and Antoniou, 2024). These contract management tasks are largely performed manually, and their effectiveness heavily depends on the cognitive abilities of contract managers, requiring significant expertise and specialized knowledge (Candaş and Tokdemir, 2022).

The primary knowledge sources for personnel in international construction enterprises typically consist of colleagues, corporate experience, and internal documentation (Kivrak et al., 2008). However, contract-related internal documents are typically complex and volumi-

Received Dec. 18, 2024; revised Aug. 12, 2025; accepted Sep. 1, 2025

Yongcheng FU, Yongqiang CHEN
College of Management and Economics, Tianjin University, Tianjin 300072, China; Laboratory of Computation and Analytics of Complex Management System, Tianjin University, Tianjin 300072, China

Tong LYU, Ziyun LV (✉)
College of Management and Economics, Tianjin University, Tianjin 300072, China
E-mail: lvziyun_0714@tju.edu.cn

This work was supported by the National Natural Science Foundation of China (Grant Nos. 72031008 and 72101175), and the Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center.

nous. Manually sifting through vast amounts of text to locate specific answers is an exceedingly time-consuming task. International construction projects are characterized by their inherently transient nature and stringent time constraints, which impede effective knowledge sharing among project personnel (Ren et al., 2018). Additionally, consulting external experts is often cost-prohibitive for typical project budgets. While online search capabilities align well with the needs of international construction projects, the reliability and quality of the retrieved information are often questionable, significantly limiting its practical utility.

To address the challenge of difficulty in acquiring contract knowledge, an increasing number of construction organizations are adopting advanced information technologies to optimize the management of contract-related knowledge (Martínez-Rojas et al., 2016). By leveraging technological tools to extract and systematically organize relevant knowledge, these organizations can share knowledge effectively, reduce the recurrence of operational errors, and ultimately improve project performance and outcomes (Banerjee et al., 2023).

Currently, knowledge management within the domain of construction contracts predominantly relies on structured knowledge base. Nevertheless, these knowledge bases encounter several critical challenges, such as data fragmentation, excessive redundancy, and an absence of reliable metrics for assessing the quality of knowledge representation, which impeding the precision of information retrieval in the knowledge bases (Cerovsek, 2011). Furthermore, interoperability remains a critical obstacle (Lu et al., 2015). Existing knowledge management systems frequently lack the capability to support natural language-based question-and-answer (Q&A) interactions, and their limited semantic interoperability compromises the accuracy, adaptability, and overall user experience of knowledge retrieval. Enhancing information retrieval systems—particularly Q&A frameworks—in terms of both precision and interoperability is crucial for advancing the effectiveness of contract knowledge management.

The pursuit of knowledge mastery has long been a fundamental objective in the evolution of artificial intelligence (AI) systems, with pioneering research underscoring the indispensable role of knowledge representation and logical reasoning. Among the most sophisticated AI models, large language models (LLMs) demonstrate exceptional capabilities in semantic comprehension and in-context learning, which enable them to perform complex reasoning tasks and generate coherent textual outputs (Zhao et al., 2023; Yu and Gong, 2024). In particular, the transformative potential of LLMs is increasingly evident in their ability to revolutionize information retrieval systems and enhance human-computer interactions, positioning them as powerful knowledge bases (Fu et al., 2024). These models alleviate the considerable costs and administrative burdens typically associated with the construction and maintenance of specialized knowledge bases, while simultaneously providing more adaptable frameworks for knowledge processing and application through dynamic, interactive Q&A interfaces (Wang et al., 2023).

Within the domain of construction contract management, LLMs present considerable promise as tools for the analysis of contractual data, the acquisition of domain-specific expertise, and the optimization of decision-making processes (Gao et al., 2025). However, the implementation of an LLM-driven Q&A system within this specialized field is not without its challenges. First, LLMs are susceptible to the hallucination phenomenon, wherein they fail to accurately recall or generate verifiable facts, occasionally producing erroneous or misleading information (Bang et al., 2023). Furthermore, LLMs exhibit inherent limitations in terms of knowledge scope, and the resource-intensive nature of retraining these models to integrate new information exacerbates this issue. While fine-tuning can circumvent the need for full model retraining, it introduces the risk of overfitting, which may lead to the detrimental phenomenon of catastrophic forgetting (Chen et al., 2020). Additionally, the interpretability of LLMs remains a topic of significant concern, as these models operate as black-box systems, with the knowledge embedded within their parameters not readily discernible (Danilevsky et al., 2020). This opacity hinders full comprehension and control of LLMs' decision-making processes, thereby limiting their practical application in sensitive domains such as construction contract management.

The knowledge graph (KG) has emerged as a sophisticated, precise, and interpretable framework for knowledge representation. KGs in the contract domain offer the advantage of enhancing the verifiability of contract knowledge through structured modeling. However, their rule-based nature inherently restricts the flexibility of the Q&A experience. Consequently, recognizing that KGs can comprehensively visualize knowledge structures and reasoning paths, research has proposed KG as a promising approach to mitigating the limitations of LLMs (Pan et al., 2024). In efforts to augment the LLMs, research has explored the integration of KGs into the pre-training phase, thereby infusing domain-specific knowledge into these models (Shen et al., 2020). Subsequent advancements have further refined this approach by facilitating the retrieval of relevant entities and relationships from the KG, which in turn enhance the input cues for the LLMs, guiding them toward more informed and contextually accurate outputs (Baek et al., 2023). However, despite these hybrid methodologies, LLMs remain peripheral to the direct reasoning processes anchored in the KG. As a consequence, the step-by-step reasoning capabilities inherent in LLMs are not fully leveraged within these systems, thereby limiting their potential for more sophisticated, knowledge-driven inference.

To address the aforementioned challenges, this paper proposes an advanced knowledge-based Q&A system for international construction contract knowledge, which integrates the KG and the LLM to facilitate collaborative reasoning. The KG serves to encapsulate entities and intricate relationships within international construction contracts by constructing graph-based structures that systematically organize and represent knowledge of contracts. By enhancing the Q&A system with the collaboration of the KG and the LLM, the system effectively mitigates prevalent issues encountered by LLMs in construction domains, including hallucination and inflexibility. Moreover, the multi-hop entity and relationship paths embedded within the KG provide a robust foundation and substantive evidence for question analysis and resolution within the LLM, thereby augmenting the interpretability of the system's logical reasoning processes. This collaborative approach not only furnishes the system with extensive domain-specific knowledge, but also enhances its overall reasoning capacity, thereby enabling the system to effectively address complex, multifaceted questions related to contract. By systematically structuring and effectively storing contract knowledge, the proposed knowledge-based Q&A system empowers contract personnel to retrieve pertinent knowledge with precision and speed, thereby facilitating a deeper understanding of contractual terms. Furthermore, it can provide auxiliary support for critical tasks such as contract analysis, review, and overall management for contract personnel, ultimately assisting organizations in streamlining contract workflows and enhancing decision-making processes in alignment with specific project requirements.

The paper is structured as follows: Section 2 provides an overview of the theoretical foundations of knowledge management in the construction domain, large language models, Q&A systems and knowledge graphs. Section 3 presents the construction of KG of international construction contracts to establish a comprehensive knowledge base. In Section 4, explores contract knowledge through the constructed KG and introduce a knowledge-based Q&A system that incorporates collaborative reasoning between LLM and KG. Section 5 further evaluates the dimensional capabilities of the proposed Q&A system. Finally, Section 6 concludes the paper with a summary of key findings.

2 Related works

2.1 Knowledge management in the construction domain

With the continuous advancement of information technologies and the increasing reliance on data-driven methodologies, knowledge bases have emerged as a highly effective means for the systematic storage and organization of information. Through the extraction of knowledge from diverse sources and its subsequent structuring as data, these knowledge bases facilitate collective archiving, thereby enhancing accessibility, scalability, and shareability (Lacosta and Thomas, 2020). Ontology, serving as the foundational framework for the semantic web and KGs, is recognized for its structural elegance and has become an indispensable tool in the development of knowledge bases for the purposes of knowledge representation and dissemination (Zhang et al., 2015). Its application has been particularly notable in the domain of construction, including risk and safety management (Chen et al., 2025). The construction safety ontology UNOCS, developed in accordance with the LinkedOpenTerms methodology, can be used to share safety knowledge among stakeholders during the construction process (Speiser et al., 2025). The ontology for construction quality assurance can serve as an initial ontology, combined with model data, historical quality data, and potential NCCs, to identify and assess anticipated quality defects (Lünig et al., 2025). Concurrently, research on construction-specific query languages has grown, underscoring that knowledge management extends beyond the mere representation and storage of heterogeneous data, emphasizing instead the critical role of establishing associative relationships that enable advanced functions such as querying, reasoning, and decision-making (Deng et al., 2022).

The construction of knowledge bases predominantly relies on historical knowledge sources, wherein users extract predefined content through query languages tailored to specific data structures, which is inherently inadequate for addressing the dynamic and complex knowledge requirements of international construction projects due to both operational constraints and limitations in content adaptability (Liu et al., 2016). In contrast, transfer learning capitalizes on previously acquired knowledge to address novel yet analogous challenges with greater efficiency, and the integration of knowledge bases with transfer learning techniques facilitates the selection of appropriate knowledge sources and the continuous updating of the knowledge bases, thereby enabling dynamic knowledge transfer and sharing both across and within projects (Pan and Yang, 2010; Xu et al., 2022; Xu et al., 2024). Nevertheless, the effective application of transfer learning necessitates tailored model training that is specifically aligned with the knowledge needs of the target project, which often entails substantial data support, posing potential limitations in terms of applicability. In this context, Q&A systems, an significant area within natural language processing (NLP), are capable of responding to user queries expressed in natural language, offering enhanced efficiency and user-friendliness compared to information retrieval systems (Wen et al., 2022). Leveraging deep learning, a key methodology in the advancement of Q&A systems, these systems substantially improve both

retrieval accuracy and operational efficiency, representing a significant upgrade over conventional database systems (Zhong et al., 2020). Additionally, Q&A systems facilitate natural language understanding and support the intelligent generation of knowledge based on the underlying knowledge base (Tian et al., 2023). Despite advancements, current knowledge-based Q&A systems in the construction sector primarily address factoid questions through simple matching techniques, with non-factoid question requiring extensive domain knowledge, in-deep reasoning and understanding remaining a significant challenge.

Knowledge bases, information retrieval systems and Q&A systems in domain of construction, with their ability to perform knowledge representation, reuse, and analysis, contribute positively to facilitating information exchange, knowledge sharing, and decision making throughout the construction life cycle. The establishment of construction domain-specific knowledge bases and Q&A systems has become a viable strategy for enhancing the management capabilities of construction enterprises.

2.2 Large language model-based Q&A system

The foundation for accomplishing Q&A tasks lies in the representation learning of natural language, a process through which NLP enables computers to comprehend human language (Chowdhary, 2020). With advancements in NLP, the structure of Q&A systems has evolved from the traditional three-stage pipeline—comprising Question Analysis, Document Retrieval, and Answer Extraction—to the more modern Retriever-Reader model, where the Retriever retrieves relevant documents for a given query, and the Reader infers the final answer from these documents (Zhu et al., 2021). Generative pre-trained language models, leveraging the Transformer architecture, can encode vast amounts of knowledge derived from large-scale textual data within their parameters, enabling them to autonomously generate responses to queries without the need for external information retrieval (Radford et al., 2019). These pre-trained models have demonstrated superior performance across numerous benchmarks, with retrieval-free approaches based on such models ushering in a novel paradigm for Q&A system development (Roberts et al., 2020). By adopting similar Transformer-based architectures and pre-training objectives, current LLMs substantially extend model size, data size, and compute, thereby enhancing their capacity (Kojima et al., 2022). In Q&A tasks, LLMs with in-context learning and prompt engineering markedly outperform other models in both zero-shot and few-shot learning scenarios (Alomari, 2024).

Although LLMs exhibit considerable potential in open-domain Q&A tasks, their direct application to restricted-domain Q&A tasks often presents challenges, primarily due to discrepancies in conversational style, language use, and task objectives across different fields (Wang et al., 2023). A key issue is hallucination, a phenomenon exacerbated by the versatility of LLMs, which has led to the expansion of the term and encompass three types: input-conflicting hallucination, context-conflicting hallucination, and fact-conflicting hallucination (Zhang et al., 2023). Fact-conflicting hallucinations occur when LLMs generate information that contradicts established world knowledge, thereby undermining the reliability of LLMs in producing content and posing significant challenges for their practical deployment in specialized domains particularly compared to other types of hallucinations. LLMs are prone to generating fact-conflicting hallucinations when tasked with domains outside their expertise or when erroneous information has been internalized from the training corpus (Ye et al., 2023). In the context of Q&A tasks, for instance, LLMs tend to provide incomplete or seemingly plausible answers in the absence of relevant knowledge, rather than declining to respond, which exemplifies this issue (Adlakha et al., 2024).

To enhance the adaptability of LLMs to domain-specific applications, fine-tuning through downstream task learning is commonly employed. However, given that LLMs typically consist of billions of parameters, their adaptation to specialized domains necessitates vast amounts of high-quality task-specific data and substantial computational resources, including high-performance GPUs or TPUs (Dettmers et al., 2024). Additionally, the complexity of LLMs architectures render them susceptible to issues such as catastrophic forgetting, where previously learned knowledge is lost, and overfitting, where the model excessively tailors itself to the target domain, which further hinder their effectiveness in domain-specific Q&A tasks (Ling et al., 2023). Moreover, LLMs are often described as “black-box” systems due to the intricate interactions of their nonlinear data processing and vast parameter networks, making it difficult to trace and interpret how specific outputs are generated from inputs (Luo and Specia, 2024). This lack of transparency presents significant challenges in terms of model interpretability, eroding user trust in AI-driven Q&A systems (Zhao et al., 2024). Given these challenges, there is a pressing need to develop domain-specific LLMs that integrate specialized domain knowledge and are optimized to meet specific objectives.

The KG as an ideal knowledge modeling approach is considered to have a complementary relationship with LLMs, where LLMs possess the capability to optimize the construction of KGs, while KGs can use explicit knowledge to guide the training of LLMs and augment their capacity to recall and apply knowledge (Yang et al., 2024). KG augmentation of LLMs can be categorized into three key aspects, including interpretability, pre-training, and inference (Pan et al., 2024). Interpretability of a LLM refers to the ability to understand and explain the inner workings and the decision-making process of the mode. LAMA is one of the pioneering works which

transforms facts from the KG into complete statements through predefined prompt templates and predicts missing entities using the LLM (Petroni et al., 2019). This approach allows knowledge embedded in LLMs to be evaluated. To further enhance LLMs' ability to comprehend actual knowledge, researchers have proposed methods to incorporate KGs during the pre-training phase. Models like ERNIE 3.0 and K-BERT integrate KGs into textual input, allowing the model to access external knowledge during the initial training period and thereby improve the understanding of the factual knowledge (Sun et al., 2021; Liu et al., 2020).

The dynamic nature of real-world knowledge poses a significant challenge, as existing methods cannot update integrated knowledge without retraining of LLMs. Therefore, researchers have proposed separating the knowledge space from the textual space and incorporating knowledge into the reasoning process (McCoy et al., 2019). This is achieved through Retrieval-Augmented Generation (RAG) techniques, which combine external knowledge bases with information retrieval and in-context learning to enhance LLM performance, thereby obviating the need for task-specific retraining (Fan et al., 2024). Module RAG includes a retrieval procedure that directly searches the corpora, such as KGs, tailored to specific scenarios, enriching the system's knowledge context and strengthening its ability to validate knowledge (Gao, et al., 2024). Think-on-Graph presents a tightly coupled KG-enhanced LLM paradigm that enables LLMs to iteratively perform beam search on KGs, identifying the most promising reasoning paths and returning the most probable outcomes (Sun et al., 2024). Similarly, MindMap constructs a prompt pipeline that enables LLMs to comprehend KG inputs and reason by integrating both implicit knowledge and retrieved external knowledge (Wen et al., 2023). Graph RAG utilizes LLMs to construct entity-based KGs and community summaries from groups of closely related entities in a two-phase sequential manner. By retrieving text index based on the graph, it significantly improves the comprehensiveness and diversity of generated answers (Edge et al., 2024).

These advancements have facilitated the development of LLM and KG-driven Q&A systems across a wide range of domains, including law, medicine, finance, and engineering. Notable examples include a knowledge graph-enhanced large language model for Q&A in hydraulic structure safety management, and large language models integrated with domain-specific multimodal knowledge graphs for Q&A in construction project management (Zhang et al., 2025; Zhou et al., 2025).

LLMs have become the preferred model for Readers of domain Q&A systems by virtue of their language comprehension and generation capabilities, and combining domain-specific knowledge beyond the contextual document with the target question is a key enhancement to the Q&A system. Research has been conducted on domain knowledge representation and domain knowledge augmentation for LLMs to provide data from a variety of sources and structures as a complement to the domain-specific knowledge.

2.3 Knowledge graph and knowledge representation of construction contracts

The representation and reasoning of human knowledge is a fundamental area within artificial intelligence (Newell et al., 1959). Knowledge graphs serve as a structured framework to represent entities and their relationships across various real-world domains, facilitating the interconnection of entities and enabling reasoning systems to derive new knowledge (Wang et al., 2017). A KG can be viewed as a graph when considering its graph structure, and when it involves semantics it can be viewed as an organic combination of an ontology and a semantic web, including a schema layer based on the ontology and a data layer that accurately and interpretably represents domain knowledge (Ji et al., 2022). KGs can represent different forms of knowledge, including both factoid and non-factoid knowledge. Factoid knowledge typically encompasses general human knowledge with complete semantics, which can be directly extracted from text in the form of triples (subject, predicate, object) (Ehrlinger and Wöß, 2016). In contrast, non-factoid knowledge pertains to dynamic, event-driven information with complex relationships, often represented through event KGs (Ding et al., 2019). These graphs model events and entities as nodes and capture the complex relationships between them—such as temporal, causal, and conditional connections—through edges, thereby enabling the tracing of event evolution and logic over time (Knez and Žitnik, 2023). A KG of construction safety hazards for large-scale hydropower projects was developed using a BERT-Att-BiLSTM-CRF-based model, encompassing nine categories of entities and eight types of relations associated with safety hazards (Yang et al., 2025). Shaw et al. (2025) constructed an ontology-based prototype KG to support Life Cycle Asset Information Management in the context of the built environment.

Construction contracts are generally structured hierarchically, with contract knowledge embedded in entities, events, and their complex relationships—such as time, conditions, and causality—represented through formal logic, which often lacks scalability for efficient retrieval. KGs address this limitation by extracting and organizing entities, events, and relationships within contracts, based on standardized concepts from the schema layer, thereby enabling more efficient retrieval and reasoning for contract knowledge (Aghaei et al., 2022). In the context of developing KG's schema layers, El-Diraby et al. (2005) proposed a construction knowledge domain ontology that classifies concepts into seven categories: process, product, project, participant, resource, technical

topic, and system. Similarly, Niu and Issa (2015) designed conceptualized and formalized ontology models to represent the knowledge specific to the claims of the AIA A201 contract. At the KG's data layer, Zheng et al. (2023a) utilized the nested structure of KG to construct ontology layers and nested graphs based on NEC contracts, applying these KGs to enhance contract review. Although research has begun to explore the use of KGs for extracting complex relationships within contracts, a formalized method for representing these complex relationships in context remains underdeveloped.

3 KG construction

Knowledge in construction contracts can be effectively captured using KGs, which represent the entities and complex relationships embedded within contracts. This section outlines the process of developing a KG of international construction contracts, which comprises three main steps. First, data related to international construction contracts is collected. This data forms the foundational basis for the subsequent step: the creation of an ontology that defines the key concepts and their relationships within the context of contracts and serves as the schema layer of the KG, ensuring both standardization and structural clarity. Finally, the data layer of the KG is constructed by extracting entities and relationships based on the schema layer and then stored in a Neo4j graph database, facilitating efficient querying.

3.1 Source data: Standard international construction contracts

Data is the basis for developing a high-quality KG, so representative contract texts need to be selected for constructing the KG. To ensure applicability and generalizability across projects, professional institutions have developed standardized templates for international construction contracts, widely used in the industry (Elkhayat and Marzouk, 2022). Project participants can modify parts of these templates to meet specific requirements, while other clauses remain unchanged (Choi et al., 2013). Common templates include the New Engineering Contract (NEC) series by the Institution of Civil Engineers (ICE) and the Fédération International des Ingénieurs-Conseils (FIDIC) contracts series. NEC and FIDIC have different developmental histories and primary business areas. The latest version of the NEC series contracts, NEC4, was released in 2017, providing a legal framework and project management procedures suitable for various forms of construction projects, widely used in the UK and Commonwealth countries. FIDIC contracts are widely accepted globally for their scientific, fair, and rigorous nature, recognized and used by international financial organizations such as the World Bank and the Asian Development Bank. The 2022 edition of the FIDIC Yellow Book is a revision of the 2017 edition, applicable to the Design and Build (DB) contract model. Given the widespread application of NEC4 and the FIDIC Yellow Book in various construction projects worldwide, this study considers these two representative contract templates as the foundation for constructing the ontology and KG for international construction contracts.

3.2 Development of ontology

Ontology construction should follow principles such as clarity, coherence, extendibility, minimal encoding bias, and minimal ontological commitment (Gruber, 1995). Classic development methods include the TOVE method, the skeletal method, the IDEF5 method, the METHONTOLOGY method, and the seven-step method (Uschold and Grüninger, 1996). Based on the METHONTOLOGY method, integrated with the strengths of the skeletal and seven-step methods, a tailored ontology development method specific for international construction contracts has been devised. The detailed process is illustrated in Fig. 1.

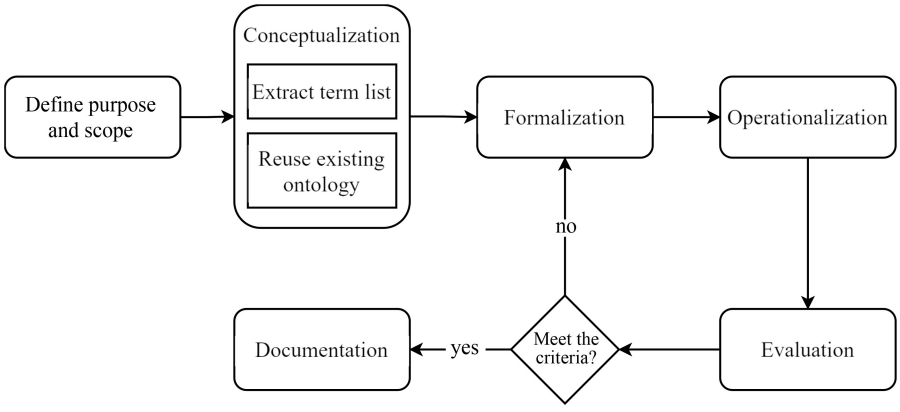


Fig. 1 Ontology development process.

In collaborative with contract experts, the competency questions the ontology should address were clarified, and knowledge sources were identified, including standardized terminology and unstructured international construction contract text data. Extracting the relevant terminology from the data source and reusing the existing builds the belonging list, ensuring the quality and efficiency of the construction process (Niu and Issa, 2015). The relationships between concepts were clarified, including constructing class hierarchies and identifying relationships between concepts. Fig A1 in Appendix presents the hierarchical relationships between major and minor categories of the international construction contract ontology, such as "environment", "promise", "resource", "actor", "process", "action", and "product".

Protégé was chosen as the ontology development environment, establishing the initial ontology hierarchy and inputting key terms from the terminology list as instances into Protégé. To convert domain knowledge into formal representations that can be processed by computer systems. The practicality and accuracy of the ontology were validated through foundational assessments and iterative improvements. The HermiT reasoner in Protégé was used to verify the consistency of the ontology, ensuring no contradictions in its intrinsic properties and assertions. Following the KG construction in Section 3.3, the ontology framework was validated by filling it at the instance level. The iteratively developed ontology framework for the international construction contract domain is shown in Table A1 in Appendix.

3.3 Construction of KG

A KG is a large semantic network, with its core in the representation of relationships between entities. Constructing a KG is essentially about building triples, which are structures composed of a subject, predicate, and object. Subjects and objects are entities in the KG, while predicates describe the relationships or attributes between subjects and objects. Contract knowledge mainly involves non-factoid knowledge that includes various complex relationships, such as conditional relationships, temporal relationships, and causal relationships. Fig. 2 shows a sentence from Clause 9 of the NEC4 contract conditions, which contains an "If-then" relationship connecting <Project Manager, Issue, Termination Certificate> and <Parties, Implement, Termination Procedure>. Without modeling the "If-Then" relationship, this conditional relationship cannot be expressed using discrete triples.

Although recent studies have emerged on modeling non-factoid knowledge, such as the Event KG and the concept of "triple-as-node," a formal method for representing such knowledge in contract contexts is still lacking (Guan et al., 2022). Therefore, before extracting triples, a method capable of expressing complex contractual relationships needs to be constructed. A systematic framework of contract analysis methods was established, and the methods of extracting triples of contract clauses are summarized in Table 1.

Especially, to capture the complex relationships embedded in contracts, we defined a set of relational attributes that retain the intricate interaction logic, as presented in Table 2. These attributes are stored as edge properties which can be directly filtered or matched in Cypher queries. This approach enables queries that go beyond "what" an action is, allowing retrieval of "under what condition, within what time, with what exceptions, and for what purpose" it applies.

Due to the complexity of the contract, the results of extracting triples manually may differ. Implementation rules were developed to regulate operations during extraction, showing in Table A2 in Appendix. As illustrated in Fig A2 in Appendix, the specific process of forming triples through entity recognition and relation extraction based on contract clauses is demonstrated using the NEC termination clause: "If either Party wishes to terminate the Contractor's obligation to Provide the Works, it notifies the Project Manager and the other Party, providing details of the reason for termination." The process begins with the clause's application as an example. Subsequently, the manually extracted triples underwent data format refinement, which included replacing Chinese punctuation with English equivalents, capitalizing the first letter of relation

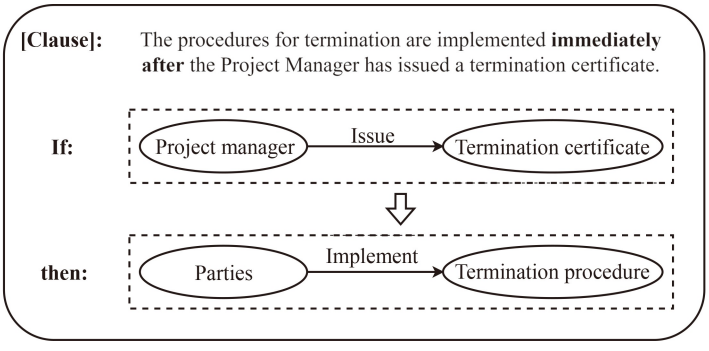


Fig. 2 Example of "If-then" relationship in NEC4 contract conditions.

Table 1 Triple extraction method for international construction contract KG

Extraction step	Description
Syntactic analysis of contract sentences	Analyze the contract clauses to identify their grammatical components, including conditional adverbial clauses, subjects, predicates, objects, and other key elements.
Extraction of key triples	Extract the most critical triples, usually the subject-predicate-object structure of the main clause. Ensure the extracted triples can express the main actions and involved objects.
Supplementing related Information	Starting from the key triples, analyze the sentence's logical structure to find other related triples. Supplement relevant details such as prerequisites until all information is fully expressed.
Checking and explaining New Nodes	Check whether the extracted triples introduce new concepts or objects, determine if further explanation or clarification of newly introduced nodes is needed to ensure the logical consistency.
Ontological classification of Nodes	Classify all extracted nodes according to their nature and relationships to ensure information conforms to the structured ontology, facilitating system organization and retrieval.

Table 2 Relationship attribute of the international contraction contract KG

Relationship attribute	Description
Time	Temporal constraints or deadlines
Condition	Triggering events or situational requirements
Exception	Exceptions or alternative cases
Object	The entity or artefact that an action is performed upon
Purpose	Intended outcomes or goals
Meaning	Definitional clauses linking terms to their conceptual meaning

names, enclosing attribute labels in double quotes, and verifying the correct placement of node attributes.

Eventually, the NEC4 contract conditions formed 1,559 triples, and the 2022 edition of the FIDIC Yellow Book formed 2,234 triples. These triples serve as the foundational data for the KG of international construction contracts, providing support for subsequent ontology refinement and KG applications.

The basis for the utilization of knowledge represented as ternary structured entity-relationship paths through the KG lies in the ability to be stored and retrieved graphically and efficiently. Neo4j was selected as the storage engine for the international construction contract KG. During the initial manual extraction process, results were presented in Excel format, and then batch imported into the Neo4j database using Python. Neo4j Browser, a visualization tool for graph databases, provides users with an intuitive interface for exploring and analyzing data. Fig. 3(a) shows the visualization of the international construction contract KG, including 3,150 nodes and 3,812 relationships. Different colors of the nodes represent different types of entities. For example, green nodes represent entities of type "Actor" and blue nodes represent entities of type "Behavior" The edges between nodes represent relationships between entities, and the density of edges reflects the complexity of the relationships.

By visually analyzing the KG, the connections and roles between entities in the KG can be deeply explored. Core entities are usually located at the center of the graph, such as "Contractor", "Engineer", and "Employer", reflecting their importance and complex relationships, as shown in Fig. 3(b). Peripheral nodes often involve explanations and supplementary information for contract clauses, as shown in Fig. 3(c). Additionally, Neo4j's Cypher query language was used to perform simple queries and various combined queries on the built KG, ensuring flexible retrieval and analysis of data in the graph database.

4 Q&A system building

In this study, a novel approach using KG to enhance LLM is proposed so that the two work together in the Q&A system for international construction contract. The Q&A system utilizes LLM as a Reader to overcome the template dependency and low interoperability problems of traditional knowledge-based Q&A systems by virtue of its powerful semantic understanding and in-contex learning capabilities, while mitigating the hallucination, inflexibility and black-box problems of the LLM based on domain knowledge provided by the KG, which enhances the in-depth reasoning capabilities and interpretability of the Q&A system. The Q&A system proposed in this study can be divided into three parts:

1) Triple vector database development. This component utilizes word embedding techniques to embed the entities and relationships from the KG into a vector database, facilitating the matching of user query with the entities and relationships in the KG.

2) Contract knowledge exploration. This process performs entity recognition and relation extraction on the user query and locates the initial entities through cosine similarity comparison. Following the confirming of the initial entities, the knowledge exploration process is initiated to uncover problem-specific reasoning paths within the KG, thereby constructing the contract subgraphs.

3) Answer generation. This process employs hybrid prompt templates designed to enable the LLM to comprehend and utilize both the user query and the contract subgraphs for reasoning, ultimately generating the final answer.

The system's architecture design, module functionalities, and their collaborative workflow are presented in Fig. 4.

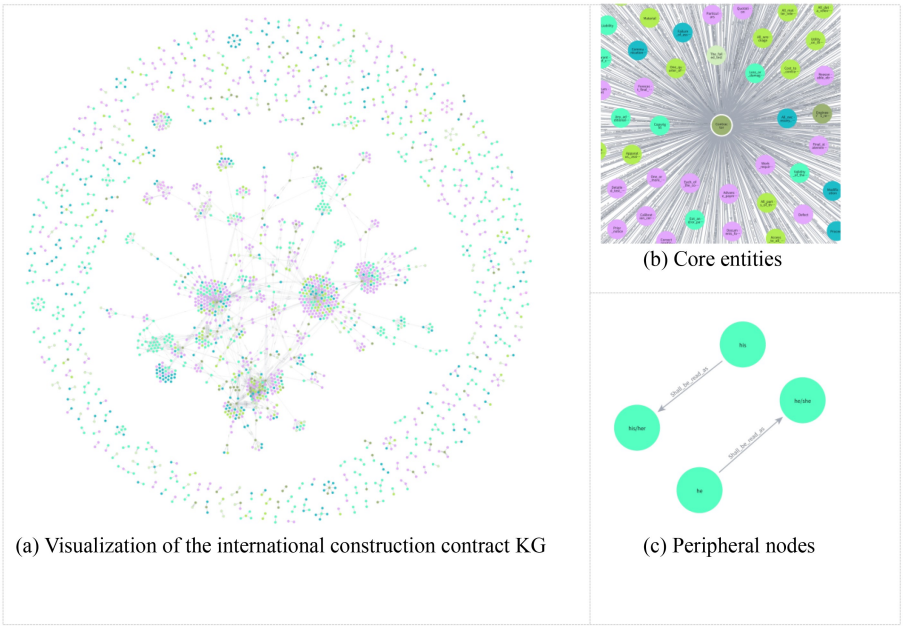


Fig. 3 Visualization of the international construction contract KG and some node examples.

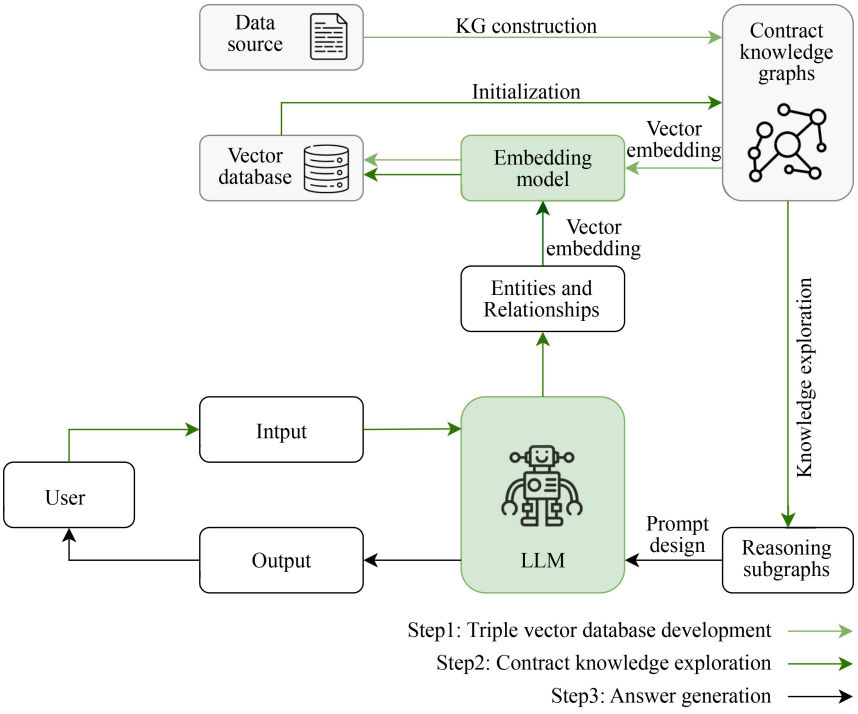


Fig. 4 Overall construction process.

4.1 Triple vector database

Word embedding is a technique that maps words or phrases from the vocabulary into real-number vectors (Mikolov et al., 2013a). By mapping the vocabulary into a lower-dimensional continuous vector space, each word or phrase is represented as a vector over the real-number field. Word embedding is a crucial foundation for downstream tasks in NLP, embedding the semantic information of natural language into numerical vectors, enabling computers to better understand and process text data. Many revolutionary outcomes in NLP have stemmed from the development of word embedding, such as Word2Vec, GloVe and BERT (Mikolov et al., 2013b; Pennington et al., 2014; Devlin et al., 2019).

Word embedding technology was used to embed entities

and relationships from the KG into a vector database, facilitating the matching of user query questions with the entities and relationships in the KG to provide accurate and relevant answers to users.

In this context, this study particularly focuses on the performance of cosine similarity of the word embedding model, which is an important measure of the similarity between two semantic vectors. The model selected for word embedding is the multi-Q&A-mpnet-base-dot-v1 model. This model is a sentence-transformers model based on the Transformer architecture, mapping sentences and paragraphs into a 768-dimensional dense vector space. It is designed specifically for semantic search tasks and has been trained on multiple data sources containing 215 million (question, answer) pairs. The model's advantage lies in its ability to capture contextual information in text, effectively extracting the semantic information of the input text through a multi-layer self-attention mechanism and fully connected network, generating corresponding semantic vectors. Moreover, the multi-Q&A-mpnet-base-dot-v1 model supports various loss functions and optimization methods, making it adaptable to different tasks and data characteristics, improving the model's performance and generalization capability. The model's robust semantic representation capability and extensive corpus training provide reliable foundational support for constructing the vector database.

4.2 Contract knowledge exploration and reasoning subgraph construction

Before embarking on knowledge exploration, it is essential to identify the core entities and initial entities that will guide the exploration process—these are the initial nodes of the reasoning paths. Given a query, the LLM is prompted to perform entity recognition and relation extraction to extract core entities, relationships, attributes, and the linking of the query entities, relationships, and attributes to the entities, relationships and attributes of the KG is performed through vector embedding to locate the initial entity set on the KG. Specifically, the query entities, relationships, and attributes are encoded using the multi-Q&A-mpnet-base-dot-v1 model to obtain the query embedding vectors. Next, the cosine similarity between each query entities, relationships, attributes and each entities, relationships, attributes in the triple vector database is computed. By comparing these cosine similarities, the entity in the triple vector database with the highest cosine similarity is selected as the nearest neighbor for each query entity to form the initial entity set for knowledge exploration and reasoning subgraphs construction.

Following the confirming of the initial entities, the knowledge exploration process is initiated to uncover problem-specific reasoning paths within the KG, thereby constructing the contract knowledge subgraphs. This exploration employs an entity keyword-based information retrieval process within KGs, encompassing both path-based exploration (Wen et al., 2023), which utilizes start and end nodes connected via multi-hop relational paths, and neighbor-based retrieval (Soman et al., 2024), which focuses on directly adjacent relationships and entities. Throughout the process, pruning and ranking strategies are applied at various stages to enhance efficiency and relevance. Knowledge exploration of this study consists of these two primary approaches: path-based exploration and neighbor-based exploration. Path-based exploration mainly focuses on the traversal of extended reasoning chains composed of multiple entities and relationships to provide in-depth reasoning logic for the contract subgraphs, which thereby augment the overall reasoning capacity and interpretability of the Q&A system. In contrast, neighbor-based exploration focuses on the expansion of the reasoning network by clustering neighbor entities that closely associated with the initial nodes, which enriches contract subgraphs by integrating dens domain-specific knowledge thereby improving the Q&A system's accuracy, as well as its capacity for summarization.

Path-based exploration is a relationship-driven approach that explore the associative connections between initial entities within the KG, with the aim of constructing in-depth contract subgraphs for question answering. The process begins by selecting one initial entity from the initial entity set as the first initial node for exploration, while the remaining initial entities are used as candidate entities to form the candidate entity set. Each entity in the candidate entiy set is sequentially seleted as the target node, and entity-relationship exploration is initiated outward from the initial node to identify all potential paths connecting the initial node to the target node within a predefined number of hops, denoted as K, and store these paths. If no paths are found between the initial node and target node, then search for the next target node. Theoretically, this procedure continues until all paths, with the initial entity as the initial node and each candidate entity as the target node, are extracted, thereby completing the knowledge exploration for the first initial entity. Once an initial entity has been used as the initial node, it is excluded from the candidate entity set for subsequent exploration, preventing the generation of redundant paths to optimize exploration efficiency. The above exploration process is then executed sequentially for each initial entities, generating the reasoning subgraphs that connect any two initial entities. Through this two-phase knowledge exploration, all potential associative paths between the initial entities in the KG are comprehensively extracted.

Neighbor-based exploration is an entity-driven approach that explores the neighboring relationships of the initial entities within the KG, aiming to construct a breadth-oriented contract subgraphs for question

answering. In this process, each entity in the initial entity set conducts a one-hop exploration, identifying neighboring entities that are directly connected to the initialentity through a single relational link and yielding all paths connecting the initial entity and neighboring entities. This exploration fetches all the context triples associated with initial entities, generating the reasoning subgraphs that captures the relationships of length one for each entity. Nevertheless, certain core nodes—such as "Contractor", "Engineer" and "Customer"—tend to have a large number of neighboring nodes, and directly selecting them as the initial nodes can lead to a reduced relevance to the query, causing the knowledge exploration inefficient and of limited significance. Therefore, core nodes with extensive relational connections are specifically excluded from the knowledge exploration to enhance the overall efficiency of the neighbor-based exploration and ensure that the generated contract subgraphs remain relevant to the query.

Due to the vast volume of triples in the vector database, the contract subgraphs generated through knowledge exploration may contain non-essential or semantically repetitive entities and relationships as well as redundant paths, which can lead to contract subgraphs exceeding the maximum token limit of the LLM, diminishing exploration efficiency and increasing computational cost. To address this, pruning of the contract subgraphs is necessary. The pruning process begins by identifying initial node entities within the KGs whose associated triples exceed a predefined threshold, and then these triples are grouped into a candidate set for pruning. Next, a cosine similarity comparison is conducted within each set of triples associated with the same initial entity, and triples with cosine similarity above a specified threshold are discarded. By merging the remaining subgraphs, the final contract subgraphs are obtained, ensuring they are both comprehensive and concise, thereby providing a solid foundation for subsequent processing.

The knowledge exploration strategy and the subsequent refinement of contract subgraphs outlined above offer crucial support for the later stages of reasoning and generation, fostering a deeper understanding and more effective application of contract knowledge. The strengths of the path-based and neighbor-based exploration methods lie in their ability to effectively integrate key information from user queries, constructing relevant contract subgraphs that balance both depth and breadth. This approach ensures that the Q&A system receives more precise and relevant inputs, ultimately enhancing system performance and overall application effectiveness.

4.3 Answer generation

After knowledge exploration, the contract subgraphs is used to prompt the LLMs for collaborative reasoning, generating the final output. A structured prompt template is created, consisting of system messages, user messages, and AI messages. These messages are integrated using LangChain technology to form a comprehensive prompt. System messages provide essential background information and guidance, while user messages present the queries, guiding the reasoning process. AI messages combine the user's queries with the reasoning subgraphs to supply the necessary reasoning material. The AI messages are then passed to the LLM, which processes them to produce detailed, accurate reasoning results and deliver the final answer to the user.

4.4 Practical application

Termination clauses are of critical significance within contracts. Clause 9, Termination, of the NEC4 contract outlines 22 categories of termination reasons, 4 types of termination procedures, 4 categories of amounts due, and their corresponding relationships. The following section uses a contract termination scenario to comprehensively demonstrate the reasoning process of the integrated KGs and LLM, contrasting its output with that of a standalone LLM. For this demonstration, GPT-4 is employed as the LLM. The specific workflow is illustrated in Fig. 5.

User Query is "In the NEC4 contract, if the Client terminates the contract due to the contractor does Corrupt Act, what are the amounts the Client should pay upon termination?" Information Extraction by GPT-4 from the user query include "if the Client terminates the contract due to Corrupt Act" and "amounts the Client should pay upon termination". The localized attributes and entities in the KGs are obtained through vector database matching, including "if_the_Contractor_does_a_Corrupt_Act,...", "Reason22", "Amount_due", and "Client". The KGs are subsequently explored, where solid lines represent path-based knowledge exploration and dashed lines denote neighbor-based knowledge exploration. A prompt template then enables GPT-4 to generate the final answer through joint reasoning over the user query and the retrieved contract subgraphs.

For the GPT-4 model without KG enhancement, responses typically begin by enumerating general considerations and providing their explanations. These usually cover aspects such as payment for completed work, expense reimbursement, prohibition of payments for corrupt practices, damages, penalties or fines, withholding of payments, and dispute resolution. Additionally, GPT-4 explicitly highlights the potential legal consequences of corrupt practices—including criminal charges, civil penalties, and debarment from future contracts—and advises owners to consult legal counsel to ensure compliance with applicable laws and regulations. However, it should be noted that while GPT-4's responses demonstrate comprehensiveness, their broad scope may make them overly generalized for contract novices or managers.

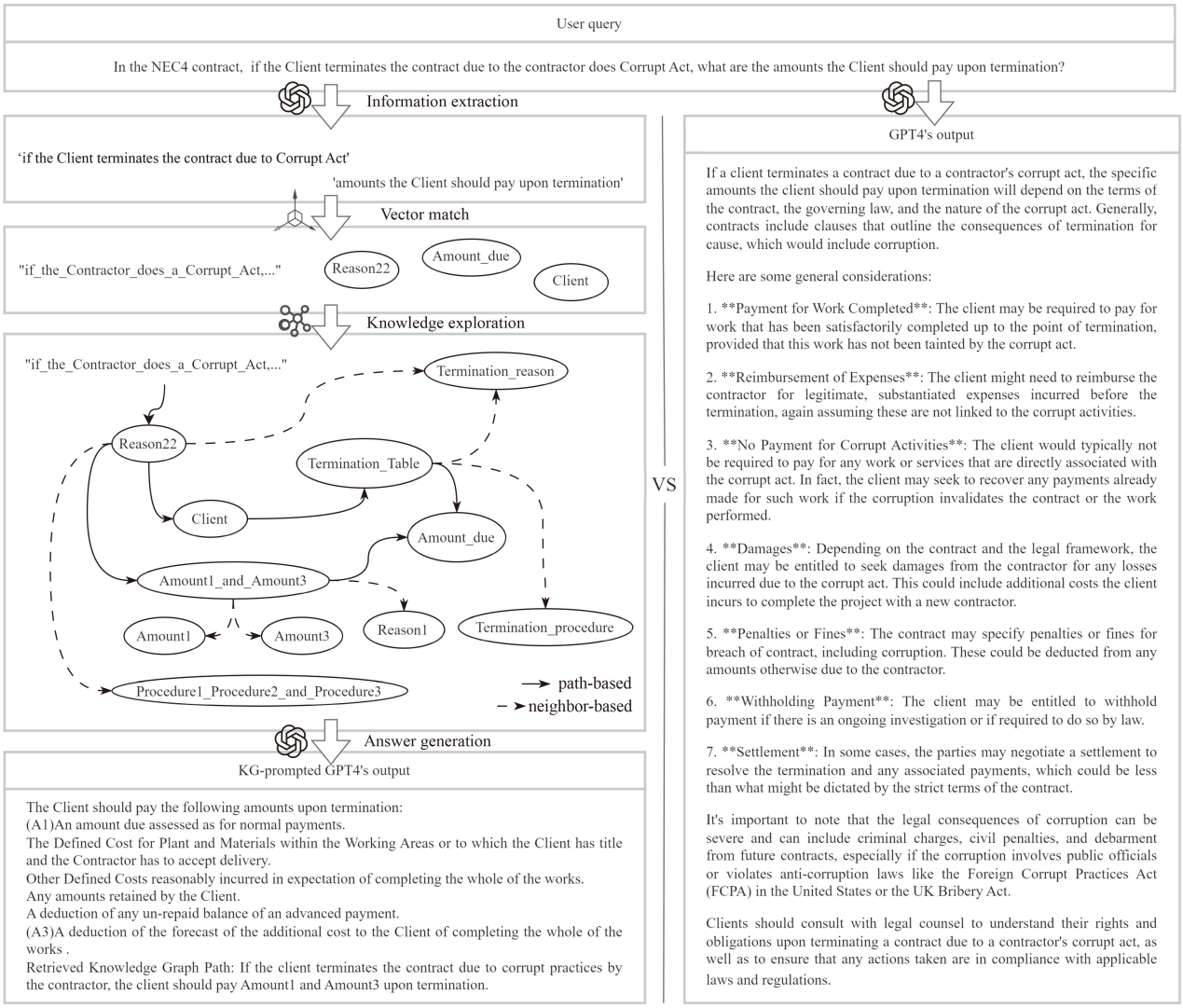


Fig. 5 Illustration of the application of KG-prompted GPT4 and unprompted GPT-4 answers.

Such responses often lack specificity and fail to address concrete practical issues. In contrast, responses generated by the KG-augmented LLM exhibit greater precision. By successfully retrieving specific contractual clauses regarding "amounts due" after termination due to contractor corruption and the associated "termination procedures," the model accurately addresses the query regarding payable amounts. Crucially, it avoids including irrelevant content beyond the scope of the question.

The accuracy and interpretability of LLMs in domain-specific Q&A tasks are often limited. In contrast, the KG-enhanced LLM approach we propose improve performance by encouraging LLMs to integrate their own knowledge with the retrieved reasoning subgraphs. This collaborative reasoning process provides well-defined and comprehensive reasoning paths for answering questions, thereby improving the accuracy, comprehensiveness, and interpretability of the Q&A system.

5 Experiments and evaluation

The performance of the Q&A system for international construction contract was evaluated on a series of tasks requiring complex reasoning and domain knowledge, comparing it with baseline models.

5.1 Dataset

In Q&A dataset, questions are commonly categorized based on the type of expected response, typically distinguishing between factoid and non-factoid questions (Cortes et al., 2022). Factoid questions are straightforward and fact-based, often requiring brief answers in the form of short phrases or sentences. In contrast, non-factoid questions are open-ended requiring complex and elaborate responses, like hypothetical type questions, causal questions, confirmation questions and summarization questions

(Mishra and Jain, 2016). Non-factoid questions generally present greater complexity and necessitate advanced natural language understanding and reasoning, making them a key focus of evaluation. While several public datasets for non-factoid Q&A systems exist, such as SQuAD, MS MARCO, WikiQ&A, and the Natural Questions corpus, these are predominantly designed for open-domain Q&A tasks and are based on answering questions related to text passages (Kwiatkowski et al., 2019). None of these datasets were suitable for our experiments. Consequently, we created our own dataset, specifically tailored for Q&A on the topic of international construction contract knowledge.

This study established its evaluation dataset using the 2017 FIDIC Yellow Book and NEC4 contract conditions as primary data sources, manually curating a set of tasks that combine multiple question types. Specifically, an objective Q&A task was designed using single-select multiple-choice questions, requiring the model to identify the single most appropriate answer from four options. This task aimed to evaluate the model's mastery of fundamental elements within international construction contracts—including the subject matter, objects, and contextual factors—covering all core clauses of NEC4 and all Level 1 clauses of the FIDIC Yellow Book 2022. Complementing this, a subjective Q&A task employed open-ended questions that demanded accurate, professional, and comprehensive responses to simulate authentic contract management scenarios. This task assessed the model's comprehensive command of international construction contract knowledge, covering NEC4 core clauses, option clauses, and cost components, alongside all Level 1 clauses of the FIDIC Yellow Book 2022. This dataset was designed to reflect both clause-level completeness and cognitive-level diversity. Each Level 1 clause of the FIDIC Yellow Book 2017 and each core clause of NEC4 is represented by at least two questions, ensuring that no clause is evaluated in isolation. For clauses addressing critical issues such as claims and variations, approximately five distinct questions were developed to capture different dimensions of interpretation and application. By combining objective and subjective question formats, the dataset effectively assesses both foundational knowledge and advanced reasoning skills, supporting a well-rounded evaluation of model performance in practical contract scenarios.

All questions were formulated by two Master's students in Engineering Management, both possessing academic training in international construction contracts, and strictly adhered to the source documents. Each question underwent rigorous assessment based on three criteria—clarity (unambiguity), relevance (connection to specific clauses), and practicality (reflection of genuine contract management needs)—rated on a 1-5 scale. Only questions scoring above 4 across all criteria were selected. The final curated test set comprises 158 questions, including 58 objective multiple-choice questions and 100 subjective open-ended questions.

5.2 Evaluation metrics

Previous Q&A systems in the construction domain have primarily emphasized correctness as the key evaluation criterion, with metrics such as accuracy, precision, recall, and F1 score being commonly used. For evaluation tasks with definitive correct answers, such as multiple-choice questions and fill-in-the-blank exercises, accuracy serves as the quantitative metric enabling clear performance comparisons between models; consequently, this study adopts accuracy as the evaluation measure for the objective question-answering task. For subjective Q&A tasks, a single correctness metric is inadequate to fully evaluate its performance of in-depth reasoning, interpretability, and summarization in Q&A tasks. Therefore, this study proposes a set of multidimensional evaluation metrics that address various aspects of output quality, focusing on Accuracy, Conciseness, Comprehensiveness, and Interpretability, providing a more comprehensive assessment of the model's performance. The definitions and criteria for the four evaluation metrics are as follows:

1) Accuracy: The model's responses should be consistent with international construction contract terms and industry standards, without factual errors or misleading content.

2) Conciseness: The model's responses should address the issue with the minimum necessary information, reducing redundant details.

3) Comprehensiveness: The model's responses should cover all core aspects of the question, with no critical information omitted.

4) Interpretability: The model's responses should present a clear logical chain and supporting evidence, making it easier for the user to understand the reasoning process.

To refine these metrics and mitigate evaluation bias, this study further specifies detailed assessment standards for each dimension (as outlined in Table 3), ensuring consistency throughout the evaluation procedure.

5.3 Evaluation method

There are two primary evaluation methods for LLMs in Q&A tasks: automatic evaluation and human evaluation (Chang et al., 2024). Automatic evaluation typically relies on standard metrics and tools, such as ROUGE and BLEU, to assess the model's performance by comparing its output with reference answers. However, since most of the questions in Q&A tasks of international construction contract are open-ended and lack fixed reference answers, automatic evaluation is not fully suitable for this context. In situations where automatic evaluation is insufficient,

Table 3 Details of the evaluation metrics

Dimension	Explanation
Accuracy	alignment with the question's core requirements
	complete and correct addressing of key elements
	use of precise terminology avoiding ambiguity
Conciseness	employment of concise syntax for direct response delivery
	exclusion of irrelevant content
	elimination of redundant information
Comprehensiveness	coverage of all pertinent contractual clauses
	inclusion of typical applications and edge cases
Interpretability	provision of evidential support and logical reasoning
	structured step-by-step explanation

human evaluation—where evaluators provide feedback on the model's responses—more closely reflects real-world application scenarios. Nevertheless, manual evaluation can be resource-intensive, costly, and time-consuming. It is also prone to inconsistencies due to cultural and personal differences among evaluators, requiring their specialized expertise.

As LLMs continue to improve, they show potential to outperform human annotation in many tasks (Gilardi et al., 2023). Some studies have proposed the LLM-as-a-Judge approach as a promising alternative to traditional human evaluation, reducing the need for human involvement and providing explanations (Zheng et al., 2023b). Despite the increasing use of LLM-as-a-Judge approach, concerns about the reliability of LLMs remain due to the presence of various biases, such as position bias, verbosity bias, and sentiment bias (Ye et al., 2024). Position bias refers to the tendency of LLMs to favor responses in certain positions over others, with this bias becoming more pronounced as the number of potential answers increases. The evaluation in this study involves only pairwise comparison of the responses, which is less affected by position bias. Verbosity bias refers to the fact that some models show a positive correlation between preference and answer length. The model evaluation metrics in this study include conciseness and comprehensive, which are able to mitigate the effect of answer length on the evaluation results to some extent. Sentiment bias refers to the tendency of LLMs to select content without sentiment elements. The Q&A system and model in this study is less involved in questions based on sentiment analysis, sentiment bias has minimal influence. Chen et al. (2024) explored the multiple biases of human judges and LLM judges and found that all judges display significant biases, but diverge in their specific inclinations.

Given these considerations, this study adopts the LLM-as-a-Judge approach for evaluation of subjective tasks. GPT-4, widely recognized in both academia and industry, is commonly used to assess the quality of model-generated responses due to its ability to provide accurate evaluations that closely align with manual assessments (Huang et al., 2024). In this study, we specifically employ a pairwise comparison method. For each contract-related question, both the baseline model and the KG-enhanced model generate responses, which are then evaluated by GPT-4O based on details of the evaluation metrics, and the win, draw, and loss rates for the KG-enhanced model are calculated. This comprehensive, multidimensional evaluation provides valuable insights into the performance of the Q&A system for international construction contracts and offers critical guidance for both model refinement and practical application.

5.4 Baseline models

To comprehensively and systematically evaluate the Q&A system for international construction contract, a series of LLMs were selected as baseline models, including GPT-4, GPT-3.5, and GLM-4. The baseline models and the models with KG enhancing were separately tasked with answering the test questions, and the results were compared to validate the effectiveness of the model (OpenAI et al., 2024).

1) GPT-4 was released by OpenAI in 2023. Its main applications include natural language generation, translation, text summarization, etc. GPT-4's innovation lies in its deeper understanding and generation capabilities, excelling in multiple NLP tasks.

2) GPT-3.5 is an improved version of GPT-3, released by OpenAI in 2022. GPT-3.5 performs well in handling complex natural language tasks, particularly in dialogue systems and text generation tasks.

3) GLM-4 is a new generation foundational model launched by Zhipu AI in 2024. GLM-4 performs excellently in various tasks, including language understanding, text generation, and Q&A systems.

5.5 Evaluation results

Table 4 shows the performance of the three baseline models and the corresponding KG augmented models in the objectivity task.

In the comparative analysis of baseline models, GPT-4 demonstrated the strongest performance on the objective task, achieving an accuracy of 81.03%. Furthermore, when contrasting baseline models with their KG-

Table 4 Experimental evaluation result of objective tasks

Model	Original LLM	LLM with KG
	Accuracy (%)	
GPT-3.5	67.24	72.41
GPT-4	81.03	86.21
GLM-4	65.52	70.69

augmented counterparts, the KG-enhanced models consistently exhibited improvements in objective task performance, yielding an approximate 5% increase in accuracy across the evaluated baselines.

The overall evaluation results of subjective tasks, shown in Fig. 6, indicate that the Q&A system, leveraging the KG to enhance the LLM, demonstrate strong performance across most competence dimensions, significantly surpassing the baseline model in both accuracy and interpretability.

5.6 Discussion

The following are the principal observations and conclusions drawn from the experimental results of subjective tasks.

KG Improves Answer Accuracy Through Clause-Level Knowledge and Structured Reasoning. Our KG provides fine-grained, domain-specific knowledge tailored to international construction contracts, enabling the LLMs to reference explicit clauses' content, structured relationships, and cross-document dependencies when generating answers. This structure allows the model to identify more precise answer sources and avoid factual errors, which contributes to the observed improvement in accuracy. For instance, GPT-3.5 with KG enhancement achieved a 97% win rate in accuracy against its non-KG version, a result attributed to the KG's ability to supplement the base model's limited domain coverage and provide reasoning information chains containing complex relationships. The varying effects among different LLMs may be due to differences in their base model training data and architecture. GPT-3.5 and GLM4 have relatively smaller base training data volumes, so their improvement after integrating the KG is more pronounced, whereas GPT-4 already has a high level of knowledge coverage, so the improvement is relatively smaller. Unlike general expectations about KG benefits, this study provides quantifiable evidence in a highly specialized domain, demonstrating how structured clause-level knowledge can directly support correct answers, especially for models with smaller training corpora.

Model Responses are More Concise, but Coverage Improvement is Limited. For example, the conciseness win rate for KG-enhanced GPT-4 reached 96%, outperforming other KG-enhanced models. However, its win rate in comprehensiveness dropped to 32%, indicating a reduction compared to the original GPT-4 without KG support. This trade-off likely reflects GPT-4's already strong baseline coverage; the KG helps streamline responses by filtering out redundant content, but may inadvertently omit some relevant details. In contrast, GPT-3.5 and GLM-4 benefit more uniformly from KG augmentation, as their weaker domain knowledge leaves greater room for both conciseness and content enrichment. These findings highlight a general tension between

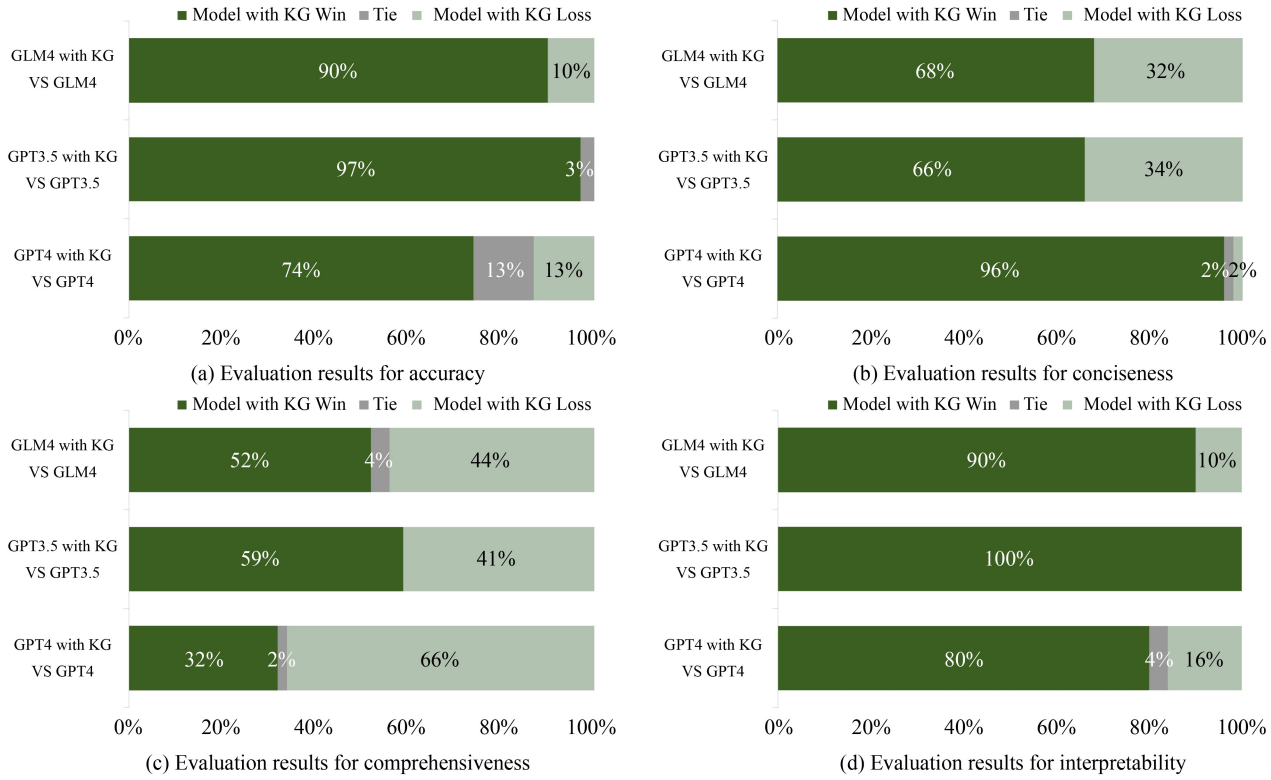


Fig. 6 Experimental evaluation result of subjective tasks.

conciseness and comprehensiveness in LLM outputs. Our KG design prioritizes practical user needs—favoring clarity and brevity to support quick understanding in applied contract scenarios. While this focus may slightly reduce informational completeness, it improves overall usability and response efficiency in real-world applications.

KG Significantly Enhances model Interpretability. While prompt engineering can influence output clarity, it lacks the capacity to embed latent domain logic or support clause traceability. In contrast, our KG-enhanced system significantly improves interpretability by structuring contract knowledge into explicit, logical pathways. For instance, KG-enhanced GPT-3.5 achieved a 100% win rate in interpretability, with similar gains observed across all evaluated models. This improvement stems from the KG's ability to provide clear and systematic logical relationships and multi-hot knowledge reasoning paths. The KG organizes and presents international construction contract knowledge in a structured way, allowing the LLMs to follow clear reasoning paths, identify relevant provisions, and understand relationships between contract elements when answering questions. This structured information not only helps the model more accurately locate relevant knowledge points but also clearly displays the relationships and logical links between various knowledge points. Moreover, the KG supports direct referencing of specific clauses and associated documents during answer generation, providing traceability that enhances both transparency and user trust.

6 Conclusions

This study presents the development of an intelligent Q&A system for international construction contracts through the integration of two emerging technologies KGs and LLMs, which is designed to facilitate contract knowledge management and support the learning of contract-related knowledge by international construction personals. An ontology framework for international construction contracts is proposed, and an ontology-based KG is constructed using over 3,000 triple groups extracted from the NEC4 and FIDIC Yellow Book 2022 contracts. The KG incorporates complex conditional, temporal, and causal relationships, enabling it to capture the full semantic context of the contract text. This structure, in conjunction with insights from the contract ontology, ensures a comprehensive representation of contract knowledge. The Q&A system, powered by the knowledge graph and LLM, supports natural language interactions, enhancing the understanding of the queries and generating relevant, context-specific answers.

This study presents three key advantages over existing Q&A systems. First, by employing a LLM as the Reader of the Q&A system, the system leverages its advanced semantic understanding and text generation capabilities, underpinned by a vast parameterized knowledge base. This enables the system to accurately comprehend complex queries, perform reasoning, and generate fluent, contextually appropriate answers. Second, the integration of reasoning subgraphs derived from the KG offering a robust foundation of intensive domain-specific knowledge and intricate chains of logical inference, substantially augments the system's capacity to accurately and comprehensively answer complex questions reducing hallucinations. These explicit reasoning paths not only enhance the system's cognitive abilities but also provide critical transparency into its underlying decision-making process, mitigating the opacity typically associated with black-box models. Consequently, this feature ensures the system's practical applicability in real-world contracts management contexts, fostering greater interpretability and reliability in its outputs. Finally, the system's flexibility is another notable advantage. The contract knowledge can be updated efficiently and at low cost through the KG, which can be easily integrated with various types of LLMs. Evaluation results demonstrate that the proposed Q&A system outperforms the baseline model in four key dimensions: accuracy, conciseness, comprehensiveness and interpretability.

Building upon the insights derived from this study, the proposed system establishes a comprehensive framework for the creation of advanced intelligent tools that empower contract personnel of international construction to efficiently retrieve and assimilate contract knowledge through AI-driven technologies. This system facilitates dynamic, flexible contract-related inquiries and learning processes, while also providing task-assisted analytical support for contract management, tailored to the specific requirements of projects and aligned with organizational objectives. By aiding contract personnel in critical activities such as risk assessment, requirements analysis, and compliance verification during contract review, the system significantly reduces both time and effort, thereby enhancing the efficiency, precision, and overall quality of the review process. Moreover, it contributes to improved management practices and project outcomes by optimizing the utilization of contract knowledge. This approach not only streamlines the management of contract-related information within international construction contract departments but also ensures the high-quality extraction, representation, and storage of contract knowledge, fostering the reuse and sharing of knowledge across the organization. Ultimately, it plays a pivotal role in advancing a robust, adaptive knowledge management system within international construction enterprises.

Despite the progress made in this study within the domain of international construction contracts, several limitations persist. Firstly, a major limitation of this study

lies in the constrained scope of data sources and evaluation design. Specifically, the knowledge graphs and evaluation dataset were constructed exclusively from the NEC4 contract and the FIDIC 2017 Yellow Book, restricting coverage to a narrow set of contract types and regions. This results in a limited coverage of contract types and fails to incorporate agreements governing a wider range of project types and regions, such as the AIA or JCT contracts. The dataset predominantly focuses on the core elements of construction contract documentation—specifically, the general conditions—while excluding supplementary documents like bills of quantities, change orders, and, crucially, real-world contract files. Future research will aim to address these limitations by expanding the scope of contract data in two key dimensions. Horizontally, the inclusion of diverse contract types—such as EPC Turnkey, AIA, and JCT contracts—will be considered, along with regional standard forms, such as China's Model Text for Construction Project Contracts. Vertically, the research will enhance clause granularity by incorporating actual project contract conditions and associated documents—including contract data sheets, contract agreements, and other ancillary files. This will serve to strengthen the model's reasoning capabilities, particularly for non-standard clauses. Moreover, the current study does not include a direct comparison with Retrieval-Augmented Generation (RAG) baselines. This is an intentional design choice to isolate and evaluate the effect of structured knowledge integration via KG alone, without the confounding influence of traditional text retrieval methods. However, a comprehensive comparison with RAG pipelines—especially those using unstructured retrieval over the same corpus—is essential to further validate the advantages of our approach. Therefore, future work will include benchmarking against representative RAG systems to evaluate performance across accuracy, conciseness, comprehensiveness, and interpretability.

Secondly, KGs in this study were constructed through manual triple extraction, a process that is both time-intensive and laborious. Moreover, this approach critically lacks support for dynamic clause modifications that may occur during contract execution. To address these limitations, future research will explore the integration of natural language processing (NLP) techniques—such as entity recognition and relation extraction—for automated KG construction. Additionally, we propose developing a structured extraction module that leverages LLMs for contract amendment identification, generating structured proposals for KG updates such as node additions and relation changes. Looking further ahead, we envision advancing beyond static representations by evolving the KG framework into a context-sensitive structure using graph neural networks (GNNs). This would potentially enable LLMs to interact with and manipulate knowledge structures dynamically. However, this direction remains outside the scope of the present work and is proposed as a subject of future investigation.

Thirdly, the current research focuses exclusively on English-language contract texts for the construction of KGs and the design of the Q&A system, utilizing the English-trained embedding model *multi-Q&A-mpnet-base-dot-v1*. While this approach has validated the methodology's effectiveness in a monolingual context, the system exhibits significant linguistic limitations when confronted with the multilingual realities of international construction projects—such as contracts in French, Arabic, or other languages, as well as multilingual user queries. Its effectiveness in processing non-English contract texts or user queries cannot be guaranteed and currently relies on translation tools or adaptation to multilingual models. To address this limitation, future work will replace the monolingual embedding model with a multilingual alternative, enabling the mapping of contract texts across languages into a unified semantic space to achieve cross-lingual entity alignment. Furthermore, entities and relations within the KG will be annotated with multilingual labels to ensure translation consistency.

Finally, while the Q&A model developed in this study primarily focuses on architectural design for knowledge Q&A systems—providing contract knowledge learning services for professionals—it has not yet been extended to concrete contract management tasks, such as contract risk identification or compliance review, due to the lack of validation through Q&A testing on actual executed contracts. Future research can enhance the practical applicability of the proposed technology by developing specialized models for risk assessment tasks. Such systems would leverage KGs constructed from real-world contracts, combined with LLMs, to build task-specific Q&A models capable of: identifying contractual risks by querying risk review checkpoints; assessing risk levels against enterprise standards; and providing actionable negotiation recommendations and implementation measures. This approach would significantly improve the utility of the technology in practical contract management.

Appendixes

Figure A1 presents the conceptual classification for the international construction contract domain, and Table A1 details the relationships between major categories under this classification. Table A2 details the rules for triple extraction for international construction contract KG.

Figure A2 presents the specific process of forming triples through entity recognition and relation extraction based on contract clauses.

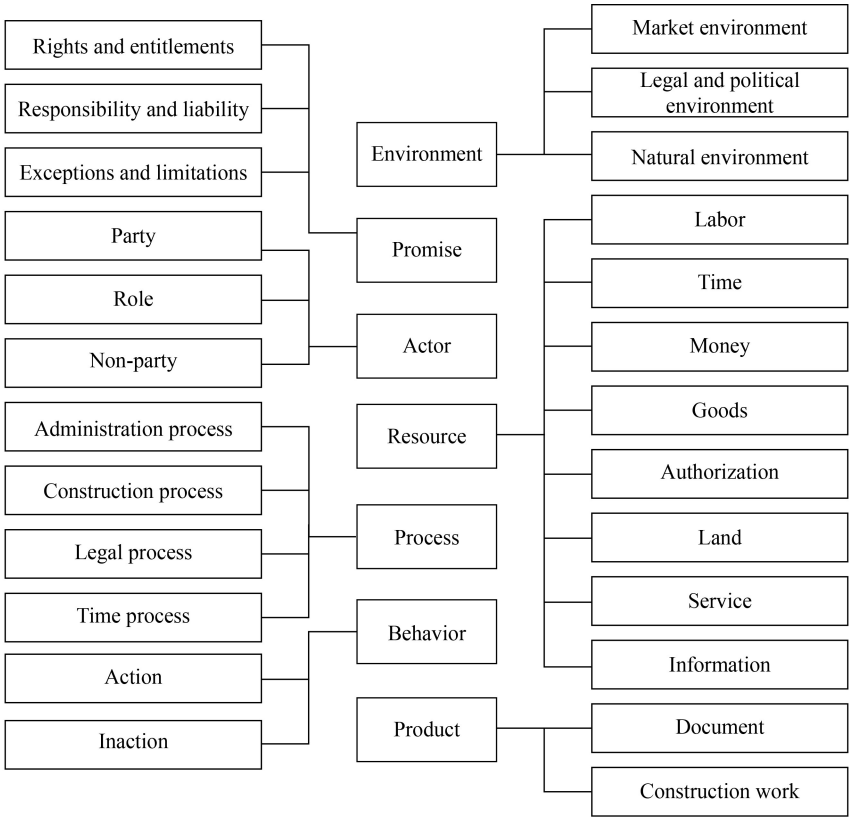


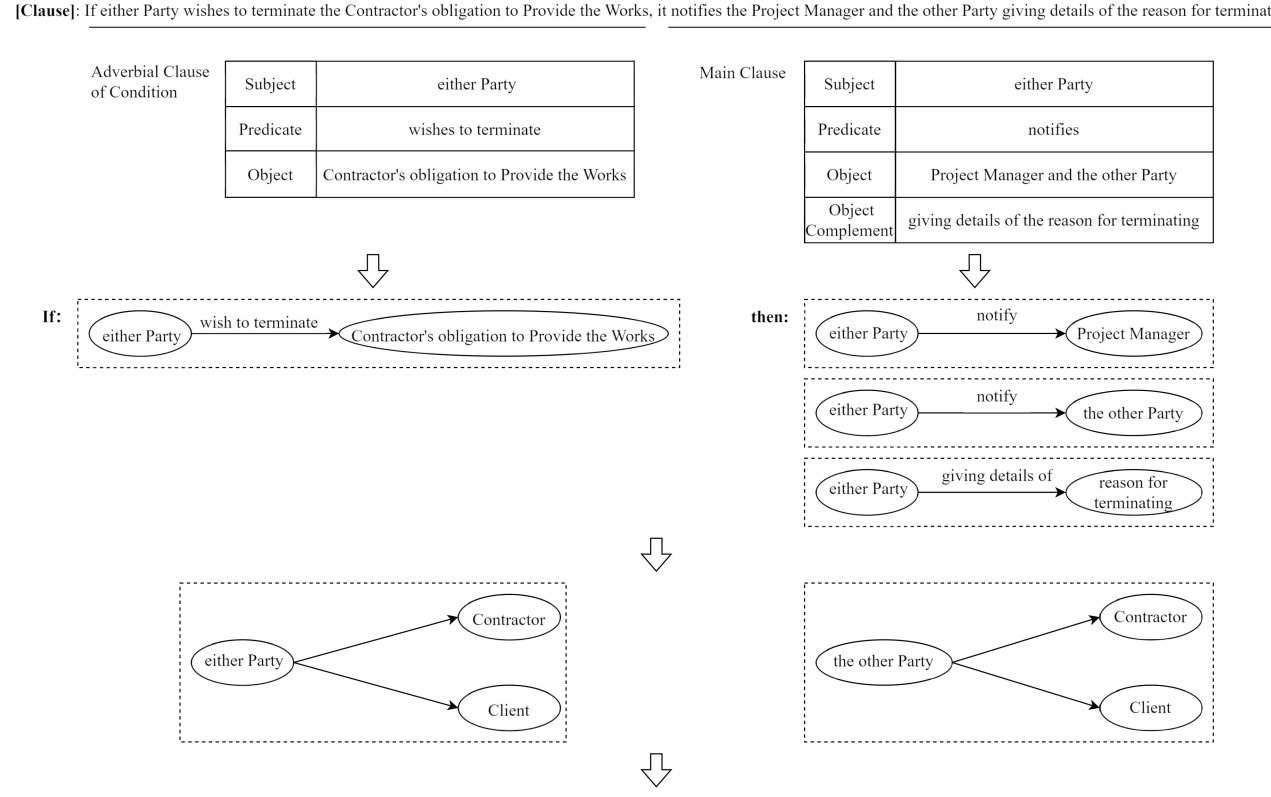
Fig. A1 Conceptual classification for international construction contract domain.

Table A1 Relationships between major categories under the classification scheme

OS	Actor	Behavior	Environment	Process	Product	Promise	Resource
Actor	Negotiate	Conduct	Respond to	Execute	Produce	Fulfill	Utilize
Behavior	Influence	Influence	Respond to	Obey	Produce	Obey	Utilize
Environment	Influence	Influence	\	Influence	Influence	Determine	Influence
Process	Guide	Regulate	Respond to	Facilitate	Serve For	Ensure	Utilize
Product	State	Reflect	State	State	\	Fulfill	Consume
Promise	Involve	Regulate	Adapt to	Guide	Fulfill	\	Require
Resource	Support	Support	\	Support	Support	Fulfill	\

Table A2 Triple extraction rules for international construction contract KG

Rule	Description
Phrase extraction	Treat noun phrases as single entities and verb phrases (including negative modifiers) as single relationships to accurately capture entities and relationships in contract text, enhancing triple completeness.
Standard writing format	Connect words in phrases for entities and relationships with underscores. Use the base form for verbs in relationships, avoid plural forms, and omit meaningless articles (such as “the” and “a”).
Splitting conjunctive relationships	In clauses with conjunctive relationships in subject/predicate/object components, these components should be split into complete <subject, predicate, object> triples.
Passive voice conversion	In sentences with predicates in passive voice, swap the subject and object to convert the passive voice into active voice, aiming to unify the expression of identical predicates without altering the contract’s meaning.
Completion and replacement of subject/object	When there is a lack of clear subject/object in the description of the action, or the subject/object is unclear in the use of pronouns, the semantics of the triplet are clarified by assigning the action to a specific contract participant.
Adjustment and completion of attributes	Enhance the expression by adding attributive modifiers to ensure accurate expression and standardizing the expression of attributes to avoid confusion.
Pattern-based extraction	Set pattern-based extraction methods for some common complex sentence structures to ensure the completeness of semantic expression and formal consistency.



Node 1	Node 1 Type	Node 1 Property	Relationship	Relationship Property	Node 2	Node 2 Type	Node 2 Property
Either_party	Actor	{}	Include	{}	Contractor	Actor	{}
Either_party	Actor	{}	Include	{}	Client	Actor	{}
The_other_party	Actor	{}	Include	{}	Client	Actor	{}
The_other_party	Actor	{}	Include	{}	Contractor	Actor	{}
Either_party	Actor	{}	Give_details_of	{"Condition":"If_either_Party_wishes_to_terminate_the_Contractor's_obligation_to_Provide_the_Works"}	The_reason_for_termination	Product	{}
Either_party	Actor	{}	Notify	{"Condition":"If_either_Party_wishes_to_terminate_th_Contractor's_obligation_to_Provide_the_Works", "Object":"Termination_reason}	Project_manager	Actor	{}
Either_party	Actor	{}	Notify	{"Condition":"If_either_Party_wishes_to_terminate_th_Contractor's_obligation_to_Provide_the_Works", "Object":"Termination_reason}	The_other_party	Actor	{}

Fig. A2 Schematic diagram of triplet extraction process.

Competing Interests The authors declare that they have no competing interests.

References

Adlakha V, BehnamGhader P, Lu X H, Meade N, Reddy S (2024). Evaluating correctness and faithfulness of instruction-following models for question answering. Transactions of the Association for Computational Linguistics, 12: 681–699

Aghaei S, Masoudi S, Chhetri T R, Fensel A (2022). Question answering over knowledge graphs: A graph-driven approach. In: IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 296–302

Alomari E A (2024). Unlocking the potential: A comprehensive systematic review of ChatGPT in natural language processing tasks. Computer Modeling in Engineering and Sciences, 141(1): 43–85

Banerjee S, Potts C M, Jhala A H, Jaselskis E J (2023). Developing a construction domain–specific artificial intelligence language model for NCDOT's CLEAR program to promote organizational innovation and institutional knowledge. Journal of Computing in Civil Engineering, 37(3): 04023007

Baek J, Aji A F, Saffari A (2023). Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. arXiv:2306.04136

Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, Lovenia H, Ji Z, Yu T, Chung W, Do Q V, Xu Y, Fung P (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning,

hallucination, and interactivity. arXiv:2302.04023

Candaş A B, Tokdemir O B (2022). Automating Coordination Efforts for Reviewing Construction Contracts with Multilabel Text Classification. Journal of Construction Engineering and Management, 148(6): 04022027

Cerovsek T (2011). A review and outlook for a 'Building Information Model'(BIM): A multi-standpoint framework for technological development. Advanced Engineering Informatics, 25(2): 224–244

Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y, Ye W, Zhang Y, Chang Y, Yu P S, Yang Q, Xie X (2024). A Survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3): 1–45

Chen G H, Chen S, Liu Z, Jiang F, Wang B (2024). Humans or LLMs as the judge? A study on judgement Biases. arXiv:2402.10669

Chen S, Hou Y, Cui Y, Che W, Liu T, Yu X (2020). Recall and learn: Fine-tuning deep pretrained language models with less forgetting. arXiv:2004.12651

Chen Y, Liang B, Hu H (2025). Research on ontology-based construction risk knowledge base development in deep foundation pit excavation. Journal of Asian Architecture and Building Engineering, 24(3): 1640–1658

Chen Y, Wang W, Zhang S, You J (2018). Understanding the multiple functions of construction contracts: The anatomy of FIDIC model contracts. Construction Management and Economics, 36(8): 472–485

Choi S J, Gulati M, Posner E A (2013). The dynamics of contract evolution. New York University Law Review, 88(1): 1–50

Cortes E G, Woloszyn V, Barone D, Möller S, Vieira R (2022). A systematic review of question answering systems for non-factoid questions. Journal of Intelligent Information Systems, 58(3): 453–480

Chowdhary K R (2020). Natural language processing. Fundamentals of Artificial Intelligence, 603–649

Danilevsky M, Qian K, Aharonov R, Katsis Y, Kawas B, Sen P (2020). A survey of the state of explainable AI for natural language processing. arXiv:2010.00711

Deng H, Xu Y, Deng Y, Lin J (2022). Transforming knowledge management in the construction industry through information and communications technology: A 15-year review. Automation in Construction, 142: 104530

Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L (2024). Qlora: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36: 10088–10115

Devlin J, Chang M-W, Lee K, Toutanova K (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805

Ding X, Li Z, Liu T, Liao K (2019). ELG: An event logic graph. arXiv:1907.08015

Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, Truitt S, Larson J (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv:2404.16130

Ehrlinger L, Wöß W (2016). Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS), 48(1–4): 2

El-Diraby T A, Lima C, Feis B (2005). Domain taxonomy for construction concepts: Toward a formal ontology for construction knowledge. Journal of Computing in Civil Engineering, 19(4): 394–406

Elkhayat Y, Marzouk M (2022). Selecting feasible standard form of construction contracts using text analysis. Advanced Engineering Informatics, 52: 101569

Fan W, Ding Y, Ning L, Wang S, Li H, Yin D, Chua T S, Li Q (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 6491–6501

Fu T, Liu S, Li P (2024). Intelligent smelting process management system: Efficient and intelligent management strategy by incorporating large language model. Frontiers of Engineering Management, 11(3): 396–412

Gao Y, Gan Y, Chen Y, Chen Y (2025). Application of large language models to intelligently analyze long construction contract texts. Construction Management and Economics, 43(3): 226–242

Gao Y, Xiong Y, Wang M, Wang H (2024). Modular RAG: transforming RAG systems into LEGO-like reconfigurable frameworks. arXiv:2407.21059

Gilardi F, Alizadeh M, Kubli M (2023). ChatGPT outperforms crowd workers for text-annotation tasks. In: Proceedings of the National Academy of Sciences of the United States of America, 120(30): e2305016120

Gruber T R (1995). Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies, 43(5–6): 907–928

Guan S, Cheng X, Bai L, Zhang F, Li Z, Zeng Y, Jin X, Guo J (2022). What is event knowledge graph: A Survey. IEEE Transactions on Knowledge and Data Engineering, 35(7): 1–21

Hassan F U, Le T, Lv X (2021). Addressing legal and contractual matters in construction using natural language processing: A critical review. Journal of Construction Engineering and Management, 147(9): 03121004

Huang H, Qu Y, Bu X, Zhou H, Liu J, Yang M, Xu B, Zhao T (2024). An empirical study of LLM-as-a-Judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. arXiv:2403.02839

Ji S, Pan S, Cambria E, Marttinen P, Yu P S (2022). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems, 33(2): 494–514

Kalogeraki M, Antoniou F (2024). Claim management and dispute resolution in the construction industry: Current research trends using novel technologies. Buildings, 14(4): 967

Kivrak S, Arslan G, Dikmen I, Birgonul M T (2008). Capturing knowledge in construction projects: Knowledge platform for contractors. Journal of Management Engineering, 24(2): 87–95

Knez T, Žitnik S (2023). Event-centric tmporal knowledge graph construction: A survey. Mathematics, 11(23): 4852

Kojima T, Gu S S, Reid M, Matsuo Y, Iwasawa Y (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35: 22199–22213

Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh A, Alberti C, Epstein D, Polosukhin I, Devlin J, Lee K, Toutanova K, Jones L, Kelcey M, Chang M-W, Dai A M, Uszkoreit J, Le Q, Petrov S

(2019). Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7: 453–466

Lacosta A, Thomas C (2020). Enterprise social networking knowledge management tools and knowledge dynamics. In: 2020 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), 1–8

Ling C, Zhao X, Lu J, Deng C, Zheng C, Wang J, Chowdhury T, Li Y, Cui H, Zhang X, Zhao T, Panalkar A, Cheng W, Wang H, Liu Y, Chen Z, Chen H, White C, Gu Q, Zhao L (2023). Domain specialization as the key to make large language models disruptive: A comprehensive survey. arXiv:2305.18703

Liu H, Lu M, Al-Hussein M (2016). Ontology-based semantic approach for construction-oriented quantity take-off from BIM models in the light-frame building industry. Advanced Engineering Informatics, 30(2): 190–207

Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, Deng H, Wang P (2020). K-bert: Enabling language representation with knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, 34(03), 2901–2908

Lu Y, Li Y, Skibniewski M, Wu Z, Wang R, Le Y (2015). Information and communication technology applications in architecture, engineering, and construction prganizations: A 15-year review. Journal of Management Engineering, 31(1): A4014010

Lünig J N, Seiß S, Schwerdtner P, Melzner J (2025). Reducing construction quality costs through ontology-based inspection planning. Advanced Engineering Informatics, 68: 103650

Luo H, Specia L (2024). From Understanding to Utilization: A survey on explainability for large language models. arXiv:2401.12874

Martínez-Rojas M, Marín N, Vila M A (2016). The role of information technologies to address data handling in construction project management. Journal of Computing in Civil Engineering, 30(4): 04015064

McCoy R T, Pavlick E, Linzen T (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. arXiv:1902.01007

Mikolov T, Chen K, Corrado G, Dean J (2013a). Efficient estimation of word representations in vector space. arXiv:1301.3781

Mikolov T, Sutskever I, Chen K, Corrado G S, Dean J (2013b). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26: 3111–3119

Mishra A, Jain S K (2016). A survey on question answering systems with classification. Journal of King Saud University - Computer and Information Sciences, 28(3): 345–361

Newell A, Shaw J C, Simon H A (1959). Report on a general problem solving program. In: IFIP Congress, 256, 64

Niu J, Issa R R A (2015). Developing taxonomy for the domain ontology of construction contractual semantics: A case study on the AIA A201 document. Advanced Engineering Informatics, 29(3): 472–482

OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, et al. (2024). GPT-4 technical report. arXiv:2303.08774

Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X (2024). Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7): 3580–3599

Pan S J, Yang Q (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10): 1345–1359

Pennington J, Socher R, Manning C D (2014). Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543

Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y, Miller A (2019). Language models as knowledge bases? arXiv:1909.01066

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8): 9

Ren X, Deng X, Liang L (2018). Knowledge transfer between projects within project-based organizations: The project nature perspective. Journal of Knowledge Management, 22(5): 1082–1103

Roberts A, Raffel C, Shazeer N (2020). How much knowledge can you pack ito the prameters of a language model? arXiv:2002.08910

Shaw C, De Andrade Pereira F, De Riet M, Hoare C, Farghaly K, O' Donnell J (2025). Knowledge graph for policy- and practice-aligned life cycle analysis and reporting. Automation in Construction, 176: 106282

Shen T, Mao Y, He P, Long G, Trischler A, Chen W (2020). Exploiting structured knowledge in text via graph-guided representation learning. arXiv:2004.14224

Soman K, Rose P W, Morris J H, Akbas R E, Smith B, Peetoom B, Villouta-Reyes C, Cerono G, Shi Y, Rizk-Jackson A, Israni S, Nelson C A, Huang S, Baranzini S E (2024). Biomedical knowledge graph-optimized prompt generation for large language models. Bioinformatics, 40(9): btae560

Speiser K, Seiß S, Boukamp F, Melzner J, Teizer J (2025). From fragmented data to unified construction safety knowledge: A process-based ontology framework for safer work. Automation in Construction, 176: 106293

Sun J, Xu C, Tang L, Wang S, Lin C, Gong Y, Ni L M, Shum H Y, Guo J (2024). Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. arXiv:2307.07697

Sun Y, Wang S, Feng S, Ding S, Pang C, Shang J, Liu J, Chen X, Zhao Y, Lu Y, Liu W, Wu Z, Gong W, Liang J, Shang Z, Sun P, Liu W, Ouyang X, Yu D, et al. (2021). ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv:2107.02137

Tian D, Li M, Ren Q, Zhang X, Han S, Shen Y (2023). Intelligent question answering method for construction safety hazard knowledge based on deep semantic mining. Automation in Construction, 145: 104670

Uschold M, Grüninger M (1996). Ontologies: Principles, methods and applications. Knowledge Engineering Review, 11(2): 93–136

Wang C, Liu X, Yue Y, Tang X, Zhang T, Jiayang C, Yao Y, Gao W, Hu X, Qi Z, Wang Y, Yang L, Wang J, Xie X, Zhang Z, Zhang Y (2023). Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv:2310.07521

Wang Q, Mao Z, Wang B, Guo L (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12): 2724–2743

Wen Y, Wang Z, Sun J (2023). MindMap: Knowledge graph prompting sparks graph of thoughts in large language models. arXiv: 2308.09729

Wen Y, Zhu X, Zhang L (2022). CQACD: A concept question-answering system for intelligent tutoring using a domain ontology with rich semantics. IEEE Access: Practical Innovations, Open Solutions, 10: 67247–67261

Xu J, Bu J, Li J (2024). A knowledge transfer framework based on deep-reinforcement learning for multistage construction projects. IEEE Transactions on Engineering Management, 71: 11361–11374

Xu J, He M, Jiang Y (2022). A novel framework of knowledge transfer system for construction projects based on knowledge graph and transfer learning. Expert Systems with Applications, 199: 116964

Yang L, Chen H, Li Z, Ding X, Wu X (2024). Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. IEEE Transactions on Knowledge and Data Engineering, 36(7): 3091–3110

Yang Y, Xiang P, Wang D (2025). Knowledge graph and mitigation measures recommendation for safety hazards in large-scale hydropower projects using diverse heterogeneous inspection data. Automation in Construction, 178: 106419

Ye H, Liu T, Zhang A, Hua W, Jia W (2023). Cognitive Mirage: A review of hallucinations in large language models. arXiv:2309.06794

Ye J, Wang Y, Huang Y, Chen D, Zhang Q, Moniz N, Gao T, Geyer W, Huang C, Chen P Y, Chawla N V, Zhang X (2024). Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. arXiv:2410.02736

Yu Z, Gong Y (2024). ChatGPT, AI-generated content, and engineering management. Frontiers of Engineering Management, 11(1): 159–166

Zhang D, Ma G, Qu T, Wang X, Zhou W, Wang X (2025). A knowledge graph-enhanced large language model for question answering of hydraulic structure safety management. Advanced Engineering Informatics, 66: 103468

Zhang S, Boukamp F, Teizer J (2015). Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for job hazard analysis (JHA). Automation in Construction, 52: 29–41

Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, Wang L, Luu A T, Bi W, Shi F, Shi S (2023). Siren's Song in the AI ocean: A survey on Hallucination in large language models. arXiv:2309.01219

Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D, Du M (2024). Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology, 15(2): 1–38

Zhao W X, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, Du Y, Yang C, Chen Y, Chen Z, Jiang J, Ren R, Li Y, Tang X, Liu Z, Liu P, Nie J, Wen J (2023). A survey of large language models. arXiv:2303.18223

Zheng C, Wong S, Su X, Tang Y (2023a). A knowledge representation approach for construction contract knowledge modeling. arXiv: 2309.12132

Zheng C, Wong S, Su X, Tang Y, Nawaz A, Kassem M (2025). Automating construction contract review using knowledge graph-enhanced large language models. Automation in Construction, 175: 106179

Zheng L, Chiang W L, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing E P, Zhang H, Gonzalez J E, Stoica I (2023b). Judging LLM-as-a-Judge with MT-bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36: 46595–46623

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

Zhou S, Liu K, Li D, Fu C, Ning Y, Ji W, Liu X, Xiao B, Wei R (2025). Augmenting general-purpose large-language models with domain-specific multimodal knowledge graph for question-answering in construction project management. Advanced Engineering Informatics, 65: 103142

Zhu F, Lei W, Wang C, Zheng J, Poria S, Chua T S (2021). Retrieving and reading: A comprehensive survey on open-domain question answering. arXiv:2101.00774