A Hybrid Retrieval-Augmented Generation Driven Multi-Agent System for Quality Control in Off-Site Construction

Fanfan Meng[1], Mi Pan[1] and Wei Pan[2]

[1]Department of Civil and Environmental Engineering, University of Macau, Macao SAR, China
[2]Department of Civil Engineering, The University of Hong Kong, Hong Kong, China
yc47936@um.edu.mo, mipan@um.edu.mo, wpan@hku.hk

Abstract –

Quality control is critical for off-site construction, but it still depends heavily on individual expertise and manual inspection. The management and coordination of heterogeneous and fragmented knowledge among the stakeholders involved in quality control of off-site construction remains challenging. To address this knowledge gap, this paper proposes QuosRA, a novel framework for **qu**ality control in **o**ff-**s**ite construction empowered by a hybrid **r**etrieval-augmented generation (RAG)-driven multi-**a**gent system. In particular, QuosRA breaks down multi-disciplinary quality control tasks into role-based subtasks handled by dedicated large language model agents with specific expertise. Each agent is assigned a particular role, expertise, and abilities for automatic quality compliance checking and quality control task guidance. The performance of QuosRA is evaluated using 30 quality inspection cases and 10 task guidance queries, compared against single-agent hybrid RAG and unstructured multi-agent system baselines. The results indicate that structured task coordination in QuosRA achieves better performance in both quality compliance checking (retrieval precision of 82.4% and determination accuracy of 93.3%) and quality control task guidance (retrieval precision of 92.9% and high-quality response score of 80.0%). Overall, QuosRA improves the efficiency and consistency of quality control workflows in off-site construction projects, offering an effective pathway toward scalable and knowledge-driven quality management in the construction industry.

Keywords –

Off-Site Construction; Quality Control; Large Language Model; Multi-Agent System; Retrieval Augmented Generation

1 Introduction

Off-site construction (OSC) is an innovative construction method that refers to the manufacturing and pre-assembly of building elements, components, or modules before their installation at their final locations [1]. It has attracted growing attention with its potential benefits in enhanced productivity, shortened construction periods, improved safety, and minimization of construction waste [2]. Quality control (QC) refers to a set of measures and procedures implemented to ensure that OSC complies with predefined quality standards and satisfies relevant regulatory requirements [3], involving quality compliance checking, quality data collection, and quality issue remediation. In particular, the widespread production of prefabricated components imposes stringent QC requirements to reduce potential quality defects before delivery to the construction site. This highlights the crucial importance of precisely monitoring and controlling the quality of prefabricated components during manufacturing.

Despite its importance, existing research on implementing QC in OSC, spanning design to construction activities, remains inadequate [3], and the current practice is still challenging. QC is highly specialized and complex, wherein quality knowledge is diverse and scattered among different stakeholders and markets [4]. OSC quality inspections highly depend on cross-disciplinary participants with varying knowledge, experiences, and responsibilities. The vast and intricate body of regulations and standards involved across regions further exhibits multi-role practitioners to track, identify, and enforce quality requirements for OSC projects [5]. Besides, the shortage and frequent turnover of skilled QC personnel greatly hinder knowledge sharing and circulation within teams, resulting in limited collective understanding and overreliance on individual capabilities in quality diagnoses [6]. Ultimately, existing quality information systems remain constrained by static rule-based mechanisms and offer limited capacity to interpret diverse QC contexts, integrate fragmented knowledge, or support the interactive decision processes required in complex OSC environments [7].

Recent advancements in LLMs, retrieval-augmented generation (RAG), and multi-agent systems (MAS) provide a promising solution to address these complexities. LLMs offer strong natural language interpretation and context-aware reasoning capabilities for knowledge-intensive tasks in construction, while MAS coordinates multiple intelligent agents to manage complex and distributed workflows [8]. Meanwhile, RAG synergistically merges intrinsic knowledge of LLMs with dynamic and domain-specific information, enhancing the accuracy and credibility of response generation [9]. The integration of LLM-based semantic understanding, RAG-driven knowledge retrieval, and MAS-supported collaborative orchestration could automate highly specialized tasks that demand nuanced comprehension and flexible coordination, such as QC in OSC. This integrated paradigm moves beyond the limitations of traditional agent systems in construction [10], enabling richer interpretation, trustworthy knowledge reuse, and adaptive multi-role coordination.

However, how the complementary strengths of LLMs, RAG, and MAS can be integrated into a coherent and domain-oriented framework for QC in OSC is still underexplored. Such integration is essential for managing fragmented knowledge, reducing inspection inconsistencies, and coordinating multiple roles across distributed off-site settings. To fill this knowledge gap, this paper aims to propose a novel framework for **qu**ality control in **o**ff-**s**ite construction empowered by a hybrid **R**AG-driven multi-**a**gent system, named QuosRA, that supports dynamic interaction, collaborative analysis, and systematic decision-making of QC-related tasks in OSC. The feasibility and efficiency of QuosRA are also demonstrated by comparing with the single-agent hybrid RAG and native MAS baselines.

2 Literature Review

2.1 Quality Control in Off-site Construction

Quality issues have been identified as one of the most critical factors affecting the implementation and development of OSC. Yin et al. [11] emphasized that OSC projects require increased emphasis on quality control measures to address prevalent quality issues. However, QC in OSC projects predominantly relies on traditional construction management guidelines, with relatively slow progress in knowledge and technological advancements [3].

Studies have explored the applications of advanced informatics technologies for quality management in OSC projects, such as building information modeling [12] and ontology [13]. However, current approaches still face limitations that result in fragmented and labor-intensive workflows, impeding effective domain knowledge capture, retrieval, and reuse. In particular, research rarely provides a systematic workflow for QC in OSC to support efficient decision-making, cross-stakeholder collaboration, and consistent quality assurance.

2.2 LLM-based Multi-agent Applications

LLMs demonstrate great potential for knowledge-intensive tasks without labor-intensive and time-consuming efforts in knowledge representation, which have been increasingly applied in construction, such as construction compliance checking and automated reporting [14]. RAG has been used to further strengthen LLM performance by mitigating hallucinations in domain-specific tasks. In the construction domain, RAG contributes to complex information retrieval and question answering in domain document parsing and regulation inquiry tasks. He et al. [7] designed an improved RAG framework for regulation information extraction in construction quality checks. Wan et al. [15] explored a hybrid RAG strategy by incorporating knowledge graph metadata into vector retrieval, supporting design for additive manufacturing tasks.

However, these RAG methods remain reactive and primarily respond to the given query, offering limited capability for complex task breakdown and multi-step task handling. Recent attention has concentrated on LLM-based MAS, since it allows agents with different roles to integrate diverse inputs, initiate internal reasoning, and coordinate planned, goal-oriented actions rather than passively reacting to user queries [16]. Each agent in the multi-agent system has been assigned a specific responsibility and specializes in a particular task. This highlights a transition from RAG-based reactive retrieval to MAS-enabled proactive orchestration, offering a more autonomous and proactive paradigm for complex and dynamic construction scenarios.

Currently, LLM-based multi-agent systems have gained increasing attention in construction, mainly on automated building design [17], building energy analysis [18], and human-machine collaboration [19]. However, few studies have explored how to employ RAG-driven reliable retrieval and deeper reasoning to enhance LLM-based MAS in handling domain-specific and knowledge-intensive tasks, such as QC in OSC projects. Meanwhile, there is still a lack of a well-designed agentic workflow for QC in OSC that enables multiple stakeholders to achieve more accurate knowledge retrieval, trustworthy compliance checking, coherent guidance generation, and coordinated decision-making.

3 Methodology

This study proposes a novel LLM-based multi-agent system for QC in OSC, named QuosRA, that integrates hybrid RAG with a centralized orchestrator architecture, as shown in Figure 1. Within QuosRA, QC-related tasks are explicitly broken down into specialized subtasks, such as query analysis, knowledge retrieval, and output evaluation, with responsibilities assigned to dedicated agents driven by LLMs tailored to distinct subtasks.

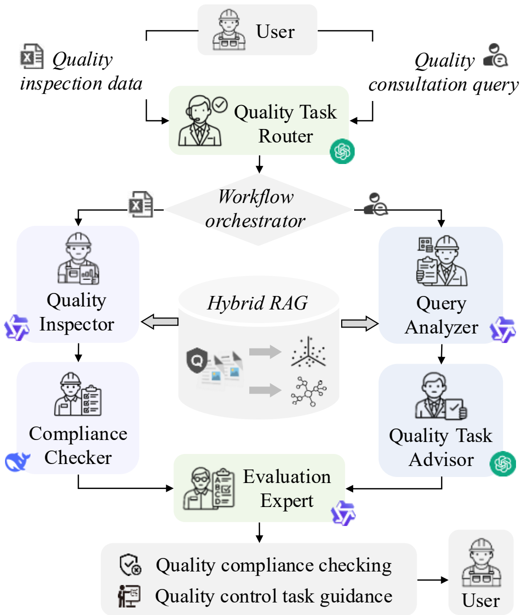

Figure 1. Conceptual framework of QuosRA

QuosRA is organized around a central workflow orchestrator that governs the overall interaction workflow. When a user query is input, Quality Task Router determines and dispatches it to the corresponding processing workflow. Quality inspection data recorded in Excel files are directed to the left-side workflow, where Quality Inspector processes data and searches requirements for Compliance Checker to check compliance. Task guidance queries are routed to the right-side workflow, where Query Analyzer parses user intent and retrieves interpretations for the downstream Quality Task Advisor to generate context-aware guidance. Quality Inspector and Query Analyzer are supported by a shared domain-specific knowledge base empowered by a hybrid RAG [20]. Evaluation Expert further reviews the synthesized results for both workflows to provide coherent and reliable QC assistants.

This centralized-orchestrated architecture of QuosRA adopts the hybridization of hierarchical and decentralized mechanisms, which has been identified as a promising approach for achieving scalability and robustness in MAS [21]. The central orchestrator ensures stable workflow sequencing, whereas the domain-specific agents perform specialized subtasks with local autonomy. It enables a balance between centralized oversight and decentralized adaptability for complex QC-related tasks.

3.1 Hybrid RAG for Knowledge Retrieval and Reasoning

Domain-specific knowledge of QC tasks in QuosRA is provided by the hybrid RAG [20], which integrates vector-based search for semantic matching with graph-based retrieval for explicit relational reasoning, aiming to enhance retrieval accuracy, trustworthiness, and reliability. The hybrid RAG pipeline is illustrated in Figure 2 and briefly elaborated below. More details can be found in previous work [20].

Quality-related documents are preprocessed and segmented into text chunks. Entities and relationships are extracted from text chunks and stored as node and edge chunks. Based on this representation, subgraphs are constructed and subsequently merged into a unified knowledge graph. Simultaneously, both node and edge chunks are encoded into vector embeddings to capture semantic features. The knowledge graph is further consolidated by assigning importance weights to nodes and edges, where edge weights are determined using frequency-based scores ($Weight_{edge}$) and node weights are computed using PageRank scores (PR_{node}).

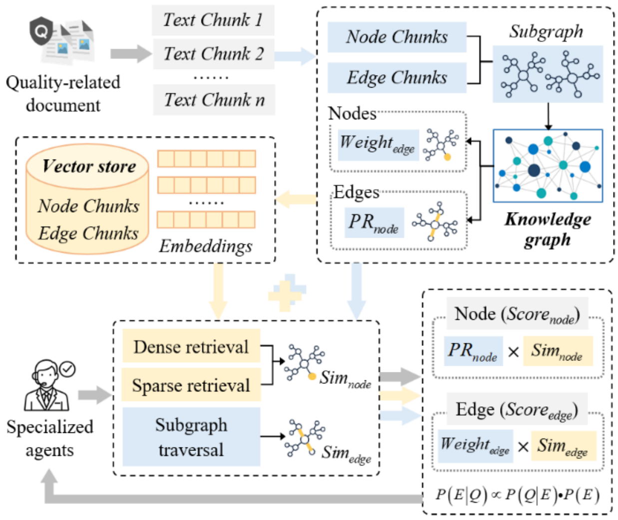

Figure 2. Procedural pipeline of the hybrid RAG

Following the role constraints encoded in the system message, Quality Inspector and Query Analyzer perform external knowledge retrieval and subsequent reasoning by invoking the hybrid RAG. The retrieval process integrates dense and sparse retrieval, as well as subgraph traversal. Specifically, dense and sparse retrieval are employed to support embedding-level and keyword-level search for entity matching. Dense retrieval measures semantic similarity using cosine similarity ($Socre_D$), while sparse retrieval adopts a Robertson-Sparck Jones Inverse Document Frequency (RSJ-IDF)-based weighting scheme without term frequency, with named entity recognition (NER) and part-of-speech (POS) tagging information to prioritize related entities ($Socre_S$). The similarity score of nodes is integrated based on the relative score fusion (RSF) (see Equation 1). In addition, subgraph traversal is employed to enable multi-hop reasoning over the knowledge graph (see Equation 2), where h denotes the hop distance.

$$Sim_{node} = 0.7Score_D \times 0.3Score_S \quad (1)$$

$$Sim_{edge} = \frac{Sim_{node}}{h+2} \quad (2)$$

To align semantic similarity with graph structural importance, a fusion scoring strategy is designed inspired by Bayes' theorem (see Equations 3-4). Driven by hybrid RAG, agents retrieve and prioritize QC domain knowledge under task-specific contexts to support the execution of their assigned tasks.

$$Score_{node} = Sim_{node} \times PR_{node} \quad (3)$$

$$Score_{edge} = Sim_{edge} \times Weight_{edge} \quad (4)$$

3.2 Agent Responsibilities Analysis

QuosRA is a distributed intelligence mechanism where specialized agents handle distinct QC subtasks in OSC, coordinated through a central router. QuosRA consists of six key agents, each with one particular role, expertise, and specialized abilities:

- ***Quality Task Router*** serves as the initial interface with users and processes both Excel and textual inputs. This agent analyzes queries, classifies them into compliance checking or guidance tasks, and routes them to the appropriate workflow handled by specialized agents.
- ***Quality Inspector*** processes structured quality inspection data recorded in Excel files and extracts inspection items. This agent is allowed to retrieve the corresponding quality requirements associated with each inspection item through the hybrid RAG. The responsibility is limited to focusing on mapping inspection items to relevant regulatory requirements.
- ***Compliance Checker*** performs quality compliance determinations based on processed inspection data and quality requirements provided by the upstream Quality Inspector. To reduce the spread of potential uncertainties, this agent only conducts compliance checking reasoning without extra retrieval behaviors.
- ***Query Analyzer*** rewrites user queries of task guidance and extracts domain-specific entities as keywords for retrieval. Based on query analyses, the agent invokes the hybrid RAG to retrieve supporting quality knowledge for query intent interpretation, which is passed downstream to guidance generation.
- ***Quality Task Advisor*** receives retrieved knowledge from Quality Analyzer to further generate coherent task guidance, which can include the clarification of quality issues, preliminary cause analysis, and recommended remediation measures, synthesized with explicit consideration of applicable standards, engineering practices, and project-specific contexts.
- ***Evaluation Expert*** consolidates and reviews the outputs produced by Quality Task Advisor and compliance checker agent to ensure accuracy and consistency of the response. Based on the final verification, this agent sends validated compliance results and QC task guidance to the user.

Through coordinated interactions among specialized agents, QuosRA adopts clear responsibility divisions and constrained interactions, which help mitigate the spread of potential uncertainties across subtasks, enabling more reliable and consistent QC in OSC.

3.3 Model Selection

In this study, agents within QuosRA are powered by three open-source medium-sized LLMs, selected based on their complementary strengths. Each model is chosen to align with subtask requirements of specialized agents, such as knowledge retrieval, reasoning, or structure output generation. Table 1 summarizes the mapping between selected open-source LLMs and agents, highlighting capability alignment.

Table 1 Model selection for agents in QuosRA

Open-source LLM	Agent	Capability	References
GPT-OSS-20b	Quality Task Router	Task abstraction capability; Instruction-following robustness	[20, 22]
	Quality Task Advisor	High-quality synthesis; Logically long-form coherence; Well-structured expression	
Qwen3-14b	Quality Inspector Quality Analyzer Evaluation Expert	Information retrieval stability Conservative yet rigorous style; Low hallucination tendency	[20, 23]
Deepseekr1-14b	Compliance Checker	Strong logical reasoning capability	[20, 24]

GPT-OSS-20B is chosen to drive Quality Task Router and Quality Task Advisor. Previous studies have shown that GPT-OSS exhibits strong logical coherence, concise expression, and clear structural organization in output generation [20, 22], which is suitable for task orchestration and guidance synthesis grounded in reliably retrieved evidence. Qwen3-14B powers Quality Inspector, Quality Analyzer, and Evaluation Expert, leveraging its reliable and stable knowledge retrieval capabilities. Qwen3-14B has been proven to show a cautious yet rigorous and robust style in knowledge retrieval and reasoning performance [20], making it particularly effective for retrieval-oriented tasks [23]. Moreover, the low hallucination tendency and balanced performance of this model ensure reliability and consistency, which are critical for comprehensive evaluation tasks. Compliance Checker is powered by DeepSeek-R1-14B, chosen for its robust logical reasoning capability [24]. Although DeepSeek-R1 excels in structured reasoning tasks, it exhibits a higher risk of hallucination compared to more conservative models [20]. To reduce the risk of uncertainty spread and improve overall trustworthiness, Compliance Checker is strictly restricted from any knowledge retrieval behaviors, relying exclusively on inspection data and pre-retrieved requirements provided by Quality Inspector.

4 Experiments

The proposed QuosRA is validated considering quality compliance checking and quality control task guidance. QuosRA builds upon the capabilities of the hybrid RAG that have been shown in prior work to provide reliable retrieval and reasoning performance [20]. Accordingly, the validation focuses on assessing the efficiency of the structured workflow design, as well as the feasibility of explicit responsibility division and constrained agent interactions in QC of OSC, rather than pursuing further improvements in domain knowledge retrieval or reasoning performance.

4.1 Experimental Design

The evaluation is conducted based on an OSC quality reference integrating both regulatory documents and practical field knowledge, including twelve regulations and standards issued by mainland China, Hong Kong SAR, and Macao SAR, and five industry guidelines and confidential project documents from industry practitioners to capture procedural knowledge and field practices in precast concrete production.

Baseline configurations are designed to isolate the impact of structured subtask coordination in QuosRA. Specifically, two baselines are applied to validate the overall performance of QuosRA. The first is a single-agent hybrid RAG powered by Qwen3-14B, which reactively responds to the given query for all QC tasks without agent decomposition. The second baseline is a native MAS with an unstructured workflow, in which both the hybrid RAG and all agents are powered by Qwen3-14B. The agent roles set in this baseline are identical to those defined in QuosRA. However, role-specific responsibilities are not enforced through explicit task decomposition or coordination mechanisms. All agents are allowed to directly access the hybrid RAG and independently process the complete QC-related task in parallel, without interaction constraints or subtask assignments. This baseline is designed to examine the impact of structured subtask decomposition and constrained coordination in QuosRA under equivalent agent and retrieval configurations. Furthermore, the LLM configurations of specialized agents in QuosRA are illustrated in Table 1, where the hybrid RAG setting is identical to baseline methods. This experimental design ensures consistent retrieval behavior across all settings, enabling a focused examination of the impact of role specialization and structured coordination.

4.2 Evaluation Task and Metrics

To fully capture real-world tasks faced by quality inspectors and decision makers in OSC projects, two types of QC-related tasks are designed.

Quality compliance checking involves determining compliance with regulatory requirements. For instance, a measured deviation of 7 mm indicates that the longitudinal reinforcement spacing in precast concrete column satisfies the ±10 mm allowable limit defined in clause 4.3.8 of DBJT 15 171-2019 and is therefore determined as compliant. The outputs are required to both extract relevant normative clauses and reasonably judge correctly. This evaluates the accuracy of compliance determinations and the relevance of the supporting references to the ground-truth, which were manually annotated to include specific normative clauses and guideline provisions. An Excel file recording 30 inspection items was provided as input, covering inspection objects, inspection points, measured data, inspectors, and inspection time. For each inspection item, QuosRA and tested baselines are asked to provide the corresponding regulatory requirements and explicit compliance determination. Two metrics are adopted to reflect the reliability of supporting evidence and the correctness of compliance judgment. ***Retrieval precision (RP)*** measures the proportion of given references that are actually correct and relevant among N inspection items, which is defined in Equation (5).

$$Retrieval\ precision = \frac{1}{N}\sum_{i=1}^{N}\frac{|S_i \cap R_i|}{S_i} \tag{5}$$

where S_i denotes the set of references retrieved by the tested method for inspection item i, and R_i refers to the ground-truth set. ***Determination accuracy (DA)*** measures the proportion of inspection items for which the correct compliance determination is given.

Quality control task guidance, which represents complex queries requiring procedural or planning advice, such as "*what is the definition of cellular defects, and how to deal with them*?" Such task involves analyzing query intention and rewriting the retrieved knowledge according to the indications made by the user query. This task emphasizes professional relevance, response coherence, and content completeness. A total of 10 representative QC task guidance queries are designed based on common quality concerns in OSC projects [20], such as defect clarification, regulatory interpretation, and measure suggestions. Two metrics are used. ***Retrieval precision (RP)*** is adopted as the same metric used in quality compliance checking to explore the reliability of supporting references. ***High-quality response score (HQS)*** evaluates the overall quality of the generated guidance, emphasizing how effectively retrieved knowledge is synthesized into actionable and coherent guidance. Specifically, this metric jointly assesses the response coherence and content completeness of outputs based on five criteria: (1) whether the quality issue is explicitly identified; (2) whether at least one relevant regulatory clause is clearly referenced; (3) whether clear and actionable measures are provided; (4) whether the guidance is logically consistent without obvious contradictions; and (5) whether the response sufficiently addresses the key aspects of the query without major omissions. A response is considered high-quality if it satisfies at least four criteria. Over a set of R responses, HQS is computed as:

$$HQS = \frac{1}{R}\sum_{r=1}^{R} 1\left(\sum_{j=1}^{5} c_{r,j} \geq 4\right) \quad c_{r,j} \epsilon \{0,1\} \qquad (6)$$

where r indexes the responses and $c_{r,j}$ indicates whether the j-th criterion is satisfied for response r (1 if satisfied, 0 otherwise).

5 Results and Analyses

For quality compliance checking, Table 2 illustrates the results of the retrieval precision and determination accuracy across three tested methods. Experimental results demonstrate that QuosRA exhibits a modest yet systematic improvement over two baseline methods, underscoring the effectiveness of its role-specialized architecture and constrained agent interactions. Specifically, compared with the hybrid RAG baseline (RP 80.6%, DA 90.0%), QuosRA achieves a moderate increase in RP (82.4%) and a more noticeable improvement in DA (93.3%). The limited enhancement in RP suggests that the primary contributions of QuosRA are not centered on greatly improving retrieval precision, but on stabilizing the downstream determination process once relevant regulatory knowledge has been retrieved. This observation aligns with the design intent of QuosRA and highlights its ability to promote greater stability and reliability in decision-making, particularly in quality compliance scenarios that require precise interpretation and integration of regulatory knowledge. For example, when retrieving the allowable deviation of the centerline position of reserved holes in concrete prefabricated slabs, Quality Inspector passes validated outputs (no more than 5 mm) to Compliance Checker, which performs consistent and reliable determinations under the structured workflow. The explicit separation between retrieval and compliance checking stages facilitates a direct and dependable translation of accurate retrieval results into compliance decisions, highlighting the scaling principle that coordination architecture can stabilize model capabilities in complex logical tasks [25]. In addition, the comparatively lower performance of the native MAS (RP 76.32%, DA 86.67%) provides empirical evidence of the negative returns from unstructured coordination [25], which indicates that multi-agent interaction without explicit role separation and coordination constraints often introduces logical interference and inconsistent interpretation of retrieved references, thereby undermining overall compliance accuracy. In contrast, QuosRA exhibits a closer alignment between RP and DA, mitigating the influence of irrelevant or conflicting references and supporting more consistent determinations of retrieved requirements.

Table 2 Experimental results of compliance checking

Method	Description	RP	DA
Hybrid RAG	Single-agent hybrid RAG	80.6%	90.0%
Native MAS	Multi-agent hybrid RAG without role separation	76.3%	86.7%
QuosRA	Role-specialized Multi-agent hybrid RAG	82.4%	93.3%

For quality control task guidance, responses produced by QuosRA show stronger reliability and higher quality than the other two baselines. The results are illustrated in Table 3. For RP, QuosRA attains the highest score (92.9%), outperforming both the hybrid RAG (88.2%) and the native MAS (83.3%). This improvement indicates that the role-specialized workflow in QuosRA contributes to more accurate alignment between query intent and retrieved regulatory references. By comparison, the native MAS shows a modest degradation in RP compared to the hybrid RAG, suggesting that solely introducing multiple agents without structured workflow design and explicit role constraints may lead to coordination noise and overlapping knowledge retrieval. Regarding HQS, both the hybrid RAG and the native MAS achieve the same HQS of 60.0%. This indicates that without structured design in agentic workflows, baseline methods may struggle with fragmented and incoherent guidance outputs. QuosRA improves the HQS to 80.0%, reflecting a higher proportion of responses that provide clear problem identification, explicit regulatory grounding, actionable measures, and logically consistent guidance. Besides, a cross-metric comparison further reveals that although the hybrid RAG achieves relatively strong retrieval precision, its overall quality of generated guidance remains limited, highlighting a clear gap between accurate retrieval and effective task-level synthesis. This observation further corroborates our previous findings in [20], which showed that Qwen3-14b exhibit more reliable and conservative retrieval behavior, yet tend to underperform in generating well-structured and coherent task guidance. By decoupling knowledge retrieval and guidance generation and assigning specialized agents powered by different LLMs, QuosRA demonstrates that a well-architected MAS can transform raw model capacity into superior task-oriented performance through strategic labor division [25]

Table 3 Experimental results of QC task guidance

Method	Description	RP	HQS
Hybrid RAG	Single-agent hybrid RAG	88.2%	60.0%
Native MAS	Multi-agent hybrid RAG without role separation	83.3%	60.0%
QuosRA	Role-specialized Multi-agent hybrid RAG	92.9%	80.0%

Regarding time efficiency, experimental results under both standalone (all-on-server) and distributed (laptop-server coordination) deployment modes indicate that QuosRA introduces a 2.7- to 3.2-fold increase in latency compared with the hybrid RAG baseline. Compared with the native MAS baseline, QuosRA achieves a 41%-49% reduction in execution time through explicit role separation and structured task decomposition. Considering that QC in OSC does not usually require strict real-time responses, the increased latency could be an acceptable trade-off for the improved reasoning reliability and structured decision support.

6 Discussion

This study proposes a centralized-orchestrated architecture multi-agent system tailored for QC in OSC. The evaluation results across QC-related tasks, involving quality compliance checking and quality control task guidance, demonstrate the effectiveness of QuosRA in supporting complex QC scenarios in OSC. Specifically, the clear responsibility divisions and constrained interactions in QuosRA mitigate interference and inconsistency in unconstrained multi-agent systems, leading to more accurate compliance checking and coherent, actionable QC guidance outputs. This design echoes the necessity of structured knowledge sharing among stakeholders in OSC [26] and offers a strategic solution to the pervasive challenges of labor shortages and expert scarcity. By mimicking a professional labor division where expert-level agents supervise technique-specific agents, QuosRA also provides a resilient management template for hierarchical workforce division, expert-level oversight, and task decomposition for QC in OSC. The cross-metric analysis further suggests that baseline methods are often limited in simultaneously achieving retrieval rigor, reasoning reliability, and generating structured responses. This finding further verifies that the lack of communication and cooperation may struggles in vital information sharing, leading to inefficiencies, conflicts, and knowledge misuse [21]. QuosRA exhibits promising potential to facilitate the transformation of retrieved knowledge into decision-relevant outputs, which demonstrates the strength of LLM-based multi-agent applications in cross-coordination and evaluation for more accurate, reliable, and robust solutions [17].

Practically, QuosRA enables non-experts to conduct QC-related tasks through natural language interactions without requiring expertise in regulatory coding or rule formalization, allowing QC to be implemented in an automated and consistent manner. Natural language can be used to obtain domain knowledge, check compliance, and access expert-level guidance in OSC. Although these interactions can also be implemented through RAG pipelines [6, 15, 20], such approaches remain reactive and limited in capability for proactive task breakdown and subtask coordination. In this regard, this paper also contributes to a scalable paradigm to expand the application of RAG and LLM-based MAS into QC coordination in OSC, and can be generalized to other complex engineering scenarios.

7 Conclusions

This study proposes QuosRA, a novel framework for QC in OSC driven by a hybrid RAG-driven multi-agent system, and validates its feasibility and efficiency through quality compliance checking and quality control task guidance. Evaluation results demonstrate that QuosRA improves the reliability and logical consistency of outputs for relevant queries. This research contributes an agentic workflow that explicitly decomposes QC-related tasks into coordinated subtasks, supported by a centralized orchestration mechanism. QuosRA provides an effective solution to address the challenges in QC of OSC projects, containing fragmented knowledge utilization, high manual effort, and limited decision support. Overall, this study establishes a solid foundation for applying LLM-based MAS for knowledge-driven management tasks in construction. Future work will extend QuosRA toward a multimodal assistant by integrating vision-language models for QC scene-aware perception. Furthermore, comprehensive ablation studies and resource efficiency analysis will be conducted to evaluate the contributions of different agent roles and further optimize system performance across diverse deployment settings.

8 Acknowledgements

This work was funded by The Science and Technology Development Fund, Macau SAR (File no. 0101/2024/RIB2) and University of Macau (File no. SRG2023-00006-FST).

References

[1] C. Goodier, and A. Gibb. Future opportunities for off-site in the UK. *Construction Management and Economics*, 25(6):585-595, 2007.

[2] M. Pan, Y. Yang, Z. Zheng, and W. Pan. Artificial intelligence and robotics for prefabricated and modular construction: A systematic literature review. *Journal of Construction Engineering and Management*, 148(9):03122004, 2022.

[3] Y. Liu, F. Yao, Y. Ji, et al. Quality control for offsite construction: Review and future directions. *Journal of Construction Engineering and Management*, 148(8), 2022.

[4] Z. Ma, S. Cai, N. Mao, et al. Construction quality management based on a collaborative system using BIM and indoor positioning. *Automation in Construction*, 92:35-45, 2018.

[5] Y. Zhang, M. Pan, W. Pan, Y. Yang, and J. Wu. Enhancing modular construction supply chain: Drivers, opportunities, constraints, concerns, strategies, and measures. *Developments in the Built Environment*, 18:100408, 2024.

[6] R. Rachel, and K. Han. Automated model-based 3D scan planning for prefabricated building components. *Journal of Computing in Civil Engineering*, 37(2):04022058, 2023.

[7] C. He, W. He, M. Liu, et al. Enriched construction regulation inquiry responses: A hybrid search approach for large language models. *Journal of Management in Engineering*, 41:04025001, 2025.

[8] T. Guo, X. Chen, Y. Wang, R. et al. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1-10, Jeju, Korea, 2024.

[9] Y. Gao, Y. Xiong, X. Gao, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint*, 2(1):2312.10997, 2023.

[10] L. Xiang, Y. Tan, G. Shen, and X. Jin. Applications of multi-agent systems from the perspective of construction management: A literature review. *Engineering, Construction and Architectural Management*, 29(9): 3288-3310, 2022.

[11] X. Yin, H. Liu, Y. Chen, and M. Al-Hussein. Building information modelling for off-site construction: Review and future directions. *Automation in Construction*, 101:72-91, 2019.

[12] Y. Zhou, Z. Zheng, J. Lin, X. Lu. Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking. *Computers in Industry*,142:103746, 2022.

[13] F. Meng, and X. Cao. An intelligent framework for rework risk identification in prefabricated construction processes based on compliance checking. *Engineering, Construction and Architectural Management*, 32(8):5066-5098, 2025.

[14] H. Pu, X. Yang, J. Li, and R. Guo. AutoRepo: A general framework for multimodal LLM-based automated construction reporting. *Expert Systems with Applications*, 255:124601, 2024.

[15] Y. Wan, Z. Chen, Y. Liu, et al. Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. *Advanced Engineering Informatics*, 65:103212, 2025.

[16] S. Chen, Y. Liu, W. Han, W. Zhang, and T. Liu. A survey on LLM-based multi-agent system: Recent advances and new frontiers in application. *arXiv preprint*, 2412.17481, 2024.

[17] Y. Dong, Z. Zhan, Y. Hu, et al. AI BIM coordinator for non-expert interaction in building design using LLM-driven multi-agent systems. *Automation in Construction*, 180:106563, 2025.

[18] L. Zhang, V. Ford, Z. Chen, and J. Chen. Automatic building energy model development and debugging using large language models agentic workflow. *Energy and Buildings*, 327:115116, 2025.

[19] J. Liu, H. Li, C. Chai, K. Chen, and D. Wang. A LLM-informed multi-agent AI system for drone-based visual inspection for infrastructure. *Advanced Engineering Informatics*, 68:103643, 2025.

[20] F. Meng, and M. Pan. Advancing quality control in off-site construction with large language models enhanced by hybrid retrieval-augmented generation. *Advanced Engineering Informatics*, 71:104381, 2026.

[21] L. Sun, Y. Yang, Q. Duan, Shi, Y., et al. Multi-agent coordination across diverse applications: A survey. *arXiv preprint*, 2502.14743, 2025.

[22] S. Agarwal, L. Ahmad, J. Ai, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint*, 2508. 10925, 2025.

[23] A. Yang, A. Li, B. Yang, B. et al. Qwen3 technical report. *arXiv preprint*, 2505.09388, 2025.

[24] J. Chua and O. Evans. Are DeepSeek R1 And Other Reasoning Models More Faithful? *arXiv preprint*, 2501.08156, 2025.

[25] Y. Kim, K. Gu, et al. Towards a Science of Scaling Agent Systems. *arXiv preprint*, 2411.11504, 2024.

[26] S. Ma, Z. Li, L. Li, et al. Exploring the effect of stakeholder relationship quality on technological innovation in off-site construction: the mediating role of the knowledge sharing. *Journal of Civil Engineering and Management*, 29(1):77-92, 2023.