Research and Implementation of Enterprise-level Private Knowledge Base in the Construction Field Based on GraphRAG Technology

Guojin Li[1], Zhenhao Wu[1], Jiujia Huang[2,*]

[1]Central South Architectural Design Institute Co., Ltd., Wuhan 434400, Hubei
[2]Institute of Artificial Intelligence, Donghua University, Shanghai 201620
**Correspondence Author*

Abstract: ***As one of the important applications of AI, Question Answering (QA) combines NLP and machine learning technologies to achieve efficient and accurate knowledge 问答. General large models such as ChatGPT, Wenxin Yiyan, and Tongyi Qianwen are powerful, but they perform poorly in specific fields such as construction professional knowledge 问答. This paper relies on Microsoft's GraphRAG technology, combines large models with vector databases, learns construction professional knowledge, policies, and regulations, and significantly enhances the learning ability in the construction field. By constructing a visual knowledge graph with Neo4J, flexible search from local to global is achieved, and efficient local storage and sharing of knowledge are realized. Practice shows that the performance of this system in construction field 问答 exceeds that of general large models, and local deployment ensures data privacy and security, providing users with a reassuring and efficient service experience, and providing a new solution for specific field knowledge 问答.***

Keywords: Knowledge ; Artificial Intelligence; GraphRAG; Large Model; Neo4J.

1. INTRODUCTION

With the continuous iterative development of large language models, large language models (LLMs) represented by OpenAI have evolved from the initial GPT-1 to the current GPT-4. The development of LLMs has created new models for enterprise knowledge management and sharing. In practical applications, the problems faced by large models in knowledge Q&A are mainly manifested in the following four aspects: 1) Hallucination problem, that is, the answers generated by large models are inconsistent with or even incorrect in the real world or the user's context, which can be divided into two types: factual hallucination and faithfulness hallucination; 2) Real-time knowledge: Due to reasons such as data collection, collation, and model training, large models cannot update knowledge in real time, and when asking questions, they may provide outdated information; 3) Contradicting facts and logic: It is possible that due to errors in the training data itself, the information generated by large models contradicts known facts or shows logical errors; 4) Knowledge limitations: The data collected by large models is based on publicly available knowledge, targeting the general public, and has not been specifically trained for knowledge in specific fields, resulting in poor performance of large models in knowledge Q&A in specific fields.

In response to the above problems, experts and scholars have proposed two solutions: fine-tuning and retrieval-augmented generation (RAG). The essential difference between them lies in whether a new model needs to be created: fine-tuning forms a new model by enabling a large model to learn domain-specific knowledge, and users ask questions to the new model; RAG vectorizes knowledge, first retrieves new knowledge, and then provides the retrieval results and context to the large model, which comprehensively generates answers. Given the high requirements for the interpretability of knowledge in the construction field and the rapid update of standard data, considering factors such as cost and technical difficulty, RAG technology is more suitable.

Shan et al. (2025) took a service design approach enhanced by multimodal personalization to rethink wine tasting for Chinese consumers [1]. Zhao et al. (2023) proposed smart warehouse track identification based on Res2Net-YOLACT+HSV [2]. Peng et al. (2023) investigated source-free domain adaptive human pose estimation [3], while Peng et al. (2024) introduced a dual-augmentor framework for domain generalization in 3D human pose estimation [4]. Wang et al. (2026) developed QA-ReID, a quality-aware query-adaptive convolution leveraging fused global and structural cues for clothes-changing re-identification [5], and Li et al. (2026) presented MFT, a memory-aware fine-tuning of SAM2 for efficient long-sequence video object segmentation [6]. Yan et al. (2026)

proposed PRISM, a pipeline for root-cause investigation via specialized multi-agents [7], while Yuan et al. (2026) designed TA-Mem, a tool-augmented autonomous memory retrieval method for large language models in long-term conversational QA [8]. Zhou (2025) studied a digital precision distribution strategy based on collaborative filtering for social media content on private domain platforms in the automotive industry [9]. Shen et al. (2025) applied the whale optimization algorithm to financial payment fraud detection [10]. Sun (2025) explored adaptive interfaces for personalized user experience using a machine learning approach [11]. Wang (2026) examined the application of computer science and technology in the big data context [12]. Ma (2025) proposed a unified framework for congestion diagnosis and dynamic mitigation in complex networks [13]. Jin (2025) optimized order allocation algorithms for industrial internet platforms [14]. Miao (2026) applied big data technologies for enhanced network security analysis [15]. Wang (2025) developed a multi-scale feature-enhanced YOLOv8 for object detection in photovoltaic farm panoramic imagery [16]. Gao (2025) researched intelligent supply chain collaboration and operational efficiency improvement for industrial electrical enterprises [17]. Deng (2025) enhanced neural network performance on tabular data via knowledge distillation and RankGauss transformation [18]. Zhang (2026) built a hybrid LSTM-GARCH model integrating volatility factors from US financial markets [19]. Meng (2023) constructed a neural network-based evaluation system for green cabling of cables [20]. Finally, Junxi et al. (2024) proposed a graph convolutional network based on matrix factorization (GCN-MF) for recommendation systems [21].

Figure 1: Application Modes of 15 RAG Technologies

By leveraging GraphRAG technology and integrating it with Tongyi Qianwen's large language model to establish a private enterprise knowledge base and build a knowledge query platform, through an intuitive and user-friendly human-computer interaction interface, users only need to simply input questions, and the system background intelligently allocates resources, integrates the questions and then submits them to the large model. The platform effectively improves work efficiency and saves query time. Combining relevant technologies, users can obtain accurate and targeted answers, ensuring the accuracy and practicality of information. Adopting a local deployment solution effectively guarantees the privacy and security of enterprise data and avoids the risk of sensitive information leakage.

2. KEY TECHNOLOGIES

2.1 GraphRAG Technology

GraphRAG is a structured, hierarchical Retrieval-Augmented Generation (RAG) framework proposed by Microsoft Research. It breaks through the traditional simple semantic search paradigm based on pure text fragments and integrates the powerful functions of the RAG mechanism and graph database (Graph DB). This method dynamically creates a knowledge graph by combining the input extensive corpus with a large language model (LLM). The graph not only contains detailed entity relationships but also combines community summaries and advanced outputs of graph machine learning, which jointly act on optimizing the prompt generation process during querying. Based on the comprehensive strategy of graph database and large language model, GraphRAG shows a significant improvement in efficiency when answering comprehensive and global questions. Its intelligent performance and knowledge mastery have achieved a qualitative leap compared to other advanced methods applied to private datasets in the past. Its implementation steps include six steps: constructing text units, extracting graph elements, enhancing the graph, community summarization, document processing, and network graph visualization process, as shown in Figure 2.

Figure 2: GraphRAG Workflow

2.2 Large Language Model

This system uses Tongyi Qianwen-7B (Qwen2-7B) developed by Alibaba Cloud. As a member of the Tongyi Qianwen large language model series, it has a powerful scale of 7 billion parameters. Qwen2-7B is built based on the advanced Transformer architecture, and its pre-training data is rich and diverse, covering various types such as web text, professional books, code, etc., ensuring the comprehensiveness and depth of the model.

Qwen2 outperforms many previous open-source models including Qwen1.5 in terms of performance. Even in comparison with closed-source models, it has demonstrated excellent performance on multiple benchmarks such as language understanding, generation, multilingual processing, code generation, mathematical operations, and reasoning. Its remarkable features are as follows: 1) Large-scale high-quality training corpus, which guarantees the accuracy of knowledge; 2) It has an advanced model architecture and shows powerful capabilities in logical reasoning, context understanding, etc.; 3) According to the description, Qwen-7B's performance significantly exceeds existing open-source models of similar scale. Even in some metrics, it is not inferior to larger-sized models, showing strong competitiveness; 4) The model weights of Qwen 2 have been open-sourced on the Hugging Face and ModelScope platforms, and the relevant sample code has also been made public on Github, providing great convenience for users.

2.3 Vector Database

In the field of artificial intelligence, the Text Embedding technology is the key to understanding and processing natural language. With the development of technology, various models have emerged. Among them, the "text-embedding-v2" model has attracted wide attention in the industry with its unique technical features and multi-scenario applications. The "text-embedding-v2" model is a deep learning-based text embedding model that converts text data into fixed-length vectors, enabling machines to understand and process natural language. This model is an upgraded version of "text-embedding-v1", aiming to improve the quality and efficiency of text representation.

3. SYSTEM ARCHITECTURE DESIGN

The construction of the system covers three major modules: the ordinary user application, the model trainer application, and the model management platform application, effectively separating the application functions, model training, and model management work. The system architecture is shown in Figure 3. Among them, both the ordinary user application and the model training application use Gradio technology to build intuitive graphical interfaces. As an efficient tool, Gradio can quickly build an interactive interface for artificial intelligence application scenarios, provide rich input and output components, and support the interaction between various types of data such as text, sound, image, and video and the application. This feature enables developers to focus more on optimizing AI business scenarios and processes without being overly concerned about building the user interface, so it has been widely used in the field of machine learning.

In the process of system implementation, Langchain is used as the development framework for large language models. The Langchain framework integrates the API interfaces of multiple large models, provides a unified access point, simplifies the process of using and integrating models, and makes it possible to freely build LLM applications. The Langchain framework mainly consists of six core modules, including the input and output module, the data connection module, the chain module, the memory module, the agent module, and the callback module. In addition, the Langchain development community has rich resources, and the official website and relevant materials are detailed, providing strong support for developers.

The system also integrates the PyTorch deep learning framework, which was developed by Facebook and officially open-sourced on the GitHub platform in 2017. PyTorch is known for its high flexibility, fast development speed, and dynamic computing, and has now been iterated to version 2.0.

By comprehensively applying the above technologies, a comprehensive platform including two front-end applications and a back-end management system has been successfully built.

4. KNOWLEDGE LEARNING

With the booming development of the construction industry, knowledge such as construction standards, specifications, and relevant policies has also been updated rapidly. In this context, large models need to continuously absorb new knowledge. Continuous learning and updating have become the key to ensuring the effectiveness and accuracy of large models, enabling them to effectively cope with the rapidly iterating industry knowledge environment. When GraphRAG conducts knowledge learning, it mainly goes through three steps: knowledge indexing, knowledge updating, and knowledge retrieval. The specific steps are as follows.

4.1 Knowledge Indexing and Dynamic Update Mechanism

Knowledge indexing is the primary step for the GraphRAG system to conduct knowledge learning. This process makes full use of the processing capabilities of large language models to deeply process the collected vertical domain knowledge documents. The specific steps include document loading, content slicing to adapt to the model processing unit, format conversion to ensure data compatibility, and vectorization processing to capture text semantic features. Through knowledge text extraction, the system can form a triple structure including entities, attributes, and relationships, and the triples are then stored in a vector database for subsequent rapid retrieval and analysis. The GraphRAG system also has a built-in knowledge dynamic update mechanism to ensure that as new knowledge emerges, the old indexes can be updated in a timely manner, thus maintaining the timeliness and accuracy of the knowledge base.

Figure 3: Enterprise-level Private Knowledge Base System Framework in the Construction Field

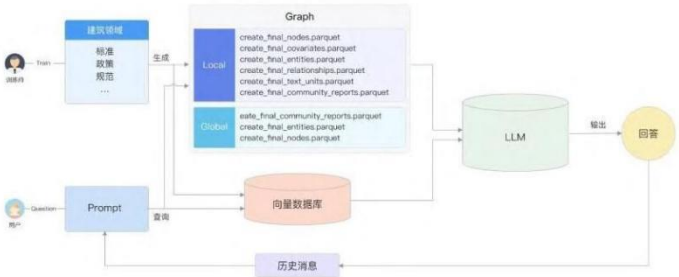

Figure 4: Schematic Diagram of Knowledge Retrieval

4.2 Precise Knowledge Retrieval Strategy

In the knowledge retrieval stage, GraphRAG utilizes the LLM (Large Language Model) service to optimize the query process. The retrieval process is shown in Figure 4. First, the system uses the LLM to perform intelligent extraction and generalization on the user's query keywords, which includes identifying capitalization variants, aliases, synonyms, etc. of the keywords, thereby broadening the retrieval scope and improving the comprehensiveness of the retrieval. Subsequently, based on these keywords, the system adopts the depth-first search (DFS) or breadth-first search (BFS) strategy to traverse the subgraph in the knowledge graph and search

for local subgraphs within a specified number of hops (N hops) to quickly locate the knowledge fragments most relevant to the user's query.

4.3 Knowledge Generation and Query Optimization

The knowledge generation process is a crucial step where GraphRAG transforms the retrieved local subgraph data into readable text. First, the system formats this subgraph data into a text form suitable for large model processing, and then submits it, along with the user's specific question, as context information to the large model for processing. In GraphRAG, there are two main query methods: Local Query and Global Query.

Local Query: Suitable for questions targeting specific entities or specific goals, it can precisely return local knowledge fragments related to the query, providing users with detailed and targeted answers.

Global Query: More suitable for questions that extract the main idea of knowledge or perform summary extraction. From a global perspective, the system can integrate and present the overall knowledge framework related to the query, helping users quickly grasp the core points of the knowledge.

5. CONCLUSIONS AND OUTLOOK

This paper focuses on the construction of a private knowledge base in the construction field. Innovatively combining the QWen2 large model and GraphRAG technology, and comprehensively applying the Langchain language processing framework and the PyTorch artificial intelligence framework, a knowledge 问答 processing system integrating model training, knowledge 问答, and local deployment has been successfully built. This system has preliminarily verified the feasibility of applying GraphRAG technology in the construction vertical field, demonstrated the broad application prospects of private knowledge base deployment, and also provided valuable practical experience for the application exploration of large language models in other vertical fields and diverse scenarios.

In the initial application process of the system, through fine prompt engineering and the integration of historical context information, the system has demonstrated excellent comprehension and 问答 performance. However, in the in-depth exploration of knowledge training and system construction, researchers have also encountered some challenges: 1) Stability and efficiency issues in the knowledge learning and training stage: During the training of the local large language model based on QWen2, communication errors occur with a small probability, revealing insufficient training stability, and the long training time has also become an urgent problem to be solved. Future research must focus on improving the stability and efficiency of model training; 2) Constructing a comprehensive evaluation mechanism: In order to more accurately evaluate the performance of this knowledge base 问答 system, a complete evaluation mechanism covering multiple dimensions such as accuracy, precision, and recall rate must be constructed. Establish an expert evaluation and Lang-Smith third-party evaluation mechanism, and strengthen the debugging, testing, evaluation, and monitoring of the system to ensure the continuous optimization and improvement of system performance; 3) Prospect of promotion and application: As an important part of the new quality productivity of enterprises, this system should not only be widely promoted and applied within the enterprise, but also actively respond to the general market demand for AI knowledge and expand to a wider range of fields. By actively expanding the construction of knowledge bases in other fields, further 挖掘 the application potential of the system and promoting the in-depth application and integration of AI technology in more fields.

REFERENCES

[1] Shan, X., Xu, Y., Xia, T., & Lin, Y. S. (2025, October). Rethinking Wine Tasting for Chinese Consumers: A Service Design Approach Enhanced by Multimodal Personalization. In 2025 International Conference on Content-Based Multimedia Indexing (CBMI) (pp. 1-5). IEEE.

[2] Zhao, X., Zhang, L., & Hu, Z. (2023). Smart warehouse track identification based on Res2Net-YOLACT+HSV. Innovation & Technology Advances, 1(1), 7–11. https://doi.org/10.61187/ita.v1i1.2

[3] Peng, Qucheng, Ce Zheng, and Chen Chen. "Source-free domain adaptive human pose estimation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[4] Peng, Qucheng, Ce Zheng, and Chen Chen. "A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[5] Wang, Y., Jiang, K., Zhang, T., Tian, K., & Jiang, G. (2026). QA-ReID: Quality-Aware Query-Adaptive Convolution Leveraging Fused Global and Structural Cues for Clothes-Changing ReID. arXiv preprint arXiv:2601.19133.
[6] Li, G., Yuan, H., Chen, S., Hu, Q., Wang, J., & Jiang, K. (2026). MFT: Memory-Aware Fine-Tuning of SAM2 for Efficient Long-Sequence Video Object Segmentation. IEEE Signal Processing Letters.
[7] Yan, W., Wu, Y., Liang, P., Yuan, M., Liu, J., Yang, J., & Li, X. (2026, March). PRISM: Pipeline for Root-cause Investigation via Specialized Multi-agents. In 2026 International Conference on Generative Artificial Intelligence and Information Security (GAIIS) (pp. 709-712). IEEE.
[8] YUAN, Mengwei, et al. TA-Mem: Tool-Augmented Autonomous Memory Retrieval for LLM in Long-Term Conversational QA. In: 2026 9th International Conference on Advanced Algorithms and Control Engineering (ICAACE). IEEE, 2026. p. 2684-2688.
[9] Zhou, Z. (2025, November). Digital precision distribution strategy for social media content on private domain platforms in the automotive industry: a collaborative filtering model based on user behavior. In Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (pp. 516-521).
[10] Shen, Zepeng, et al. "Research on Application of Whale Optimization Algorithm in Financial Payment Fraud Detection." 2025 4th International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID). IEEE, 2025.
[11] Sun, L. (2025, November). Adaptive Interfaces for Personalized User Experience: A Machine Learning Approach. In Proceedings of the 2025 International Conference on Artificial Intelligence and Sustainable Development (pp. 457-462).
[12] Wang, J. (2026). Research on the Application of Computer Science and Technology in the Context of Big Data. International Journal of Advance in Applied Science Research, 5(1), 72-77.
[13] Ma, J. (2025). A Unified Framework for Congestion Diagnosis and Dynamic Mitigation in Complex Networks. International Journal of Advance in Applied Science Research, 4(11), 36-41.
[14] Jin, L. (2025). Optimization of Order Allocation Algorithms for Industrial Internet Platforms. International Journal of Advance in Applied Science Research, 4(12), 44-48.
[15] Miao, J. (2026). Big Data Technologies for Enhanced Network Security Analysis: Applications and Approaches. International Journal of Advance in Applied Science Research, 5(4), 16-20.
[16] Wang, J. (2025). Multi-Scale Feature-Enhanced YOLOv8 for Object Detection in Photovoltaic Farm Panoramic Imagery. International Journal of Advance in Applied Science Research, 4(10), 7-11.
[17] Gao, W. (2025, November). Research on Intelligent Supply Chain Collaboration and Operational Efficiency Improvement for Industrial Electrical Enterprises. In Proceedings of the 2025 International Conference on Artificial Intelligence and Sustainable Development (pp. 163-170).
[18] Deng, Xiaoxiao. "Enhancing Neural Network Performance on Tabular Data via Knowledge Distillation and RankGauss Transformation." 2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). IEEE, 2025.
[19] Zhang, X. (2026, March). A Hybrid LSTM-GARCH Model Integrating Volatility Factors from the US Financial Markets. In Proceedings of the 2026 International Conference on AI Decision-Making and Management (pp. 183-189).
[20] Meng , L. (2023). Research on the Evaluation System of Green Cabling of Cables Based on Neural Network. Innovation & Technology Advances, 1(2), 25–31. https://doi.org/10.61187/ita.v1i2.37
[21] Junxi, Y., Wang, Z., & Chen, C. (2024). GCN-MF: A graph convolutional network based on matrix factorization for recommendation. Innovation & Technology Advances, 2(1), 14–26. https://doi.org/10.61187/ita.v2i1.30