Research on Intelligent Question Answering Technology for Indicator Procurement Based on Large Language Model

Longxue Qiao[1], Shengchun Xiao[1], Fuxian Dou[2], Zhicheng Liu[3], Shuai Li[3,*]
[1]Department of Bidding Procurement, Medical Supplies Center of PLA General Hospital, Beijing, China
[2]Western Medical Branch of PLA General Hospital, Beijing, China
[3]Department of Telemedicine, Medical Supplies Center of PLA General Hospital, Beijing, China
**Corresponding Author*

Abstract*:* Bidding Q&A and technical consultation are crucial for ensuring fair competition and enhancing transparency in procurement. However, traditional manual and search-engine-based methods struggle with the growing volume of public bidding information. While large language models (LLMs) offer potential for intelligent question-answering systems, their direct application is limited by high fine-tuning costs, tendency to generate hallucinations, and lack of domain-specific accuracy. To address this, this study proposes a hybrid approach integrating Knowledge Graph (KG) and Retrieval-Augmented Generation (RAG) to optimize generative models for intelligent bidding consultation. First, a domain-specific knowledge base is constructed using a KG and jointly learned with an LLM to enhance its professional knowledge representation. Second, the RAG framework dynamically retrieves relevant information from this knowledge base to ground the LLM's responses, thereby improving inference for complex queries. Tests demonstrate that, compared to traditional manual consultation and search engine retrieval, this proposed scheme significantly improves both the accuracy ($p<0.05$) and response efficiency ($p<0.05$) of bidding Q&A and technical consultation. The study provides a valuable reference for developing effective intelligent consultation systems in the bidding domain.

Keywords: Big Language Model; Question Answering System; Bidding; Knowledge Graph; RAG

1. Introduction

In bidding activities, bidding Q&A and technical consultation can provide decision-making support information such as market analysis, supplier evaluation, bidding process, regulations, product information, etc. for both bidding and tendering parties, which is of great significance in ensuring bidding efficiency, compliance, quality, and fairness. In recent years, the scale of global bidding has developed rapidly, and with the development of Internet technology, the public bidding information has grown exponentially, further increasing the difficulty of information search. The traditional manual consultation and search engine retrieval methods have been unable to meet the actual needs of the bidding process [1]. How to provide efficient and high-quality information consultation for both bidding parties has become an important problem that needs to be solved urgently.

With the development of big language models and text mining technologies, intelligent question answering systems that can parse user questions and generate targeted responses have gradually emerged, creating opportunities for efficient information consulting in the field of bidding and tendering. Accurately generating replies by analyzing user intentions can not only meet user needs, but also improve the efficiency of information consultation between bidding parties, thereby ensuring the openness and transparency of bidding information. In recent years, Generative Large Language Models (GLLM), represented by GPT, have attracted widespread attention from researchers due to their good universality, context understanding ability, and natural language text generation ability. However, traditional GLLM models have problems such as high secondary training costs, model hallucinations, and lack of accuracy in professional fields, which limit their application in intelligent question answering systems. To address the above issues,

the following methods are commonly used to improve the performance of generative models: firstly, a joint learning approach combining knowledge networks and large language models is adopted, which involves using knowledge graphs to induce domain knowledge entities, obtaining questioning intentions through problem semantic analysis, retrieving and inferring knowledge entity information related to the problem based on the knowledge graph, and inputting it into GLLM to enhance its domain knowledge expression ability [2]; Secondly, adopting Retrieval Augmented Generation (RAG) technology to improve GLLM can enhance the implementation of complex answer reasoning by introducing and retrieving information from external knowledge bases [3]. This technology has significant advantages in improving the quality of text generation, increasing answer diversity, reducing erroneous information, improving efficiency, and reducing costs; Thirdly, the generative ability of GLLM in specific domain Q&A can be enhanced through model fine-tuning techniques. For example, the LoRA (Low Rank Adaptation) method can simulate full parameter fine-tuning by adding a bypass matrix based on the inherent low rank characteristics of large models [4]. By modifying a small number of model parameters, it can not only improve the performance of generative models in specific domains, but also significantly reduce the secondary training cost of GLLM.

2. Methods

The research proposes the integration of knowledge graph and RAG strategies to enhance the application effectiveness of generative models in intelligent Q&A applications for bidding, including the following process: firstly, classifying user question types (such as regulatory queries, process guidance, bid generation, etc.) and using deep learning models to parse question semantics; Secondly, based on traditional bidding rules, a bidding knowledge graph and entity information retrieval function are constructed to provide LLM with domain rules (such as bidding domain rules, enterprise competition network structure, qualification review, and document compliance); Thirdly, using GLLM as the basic framework for question answering and enhancing the large language model based on RAG technology; Fourthly, based on the context information aggregation mechanism of the big language model, achieve the semantic fusion of knowledge graph and RAG enhanced LLM output; Fifth, conduct testing experiments on the performance of the intelligent question answering system based on expert evaluation, and compare it with traditional generative models. As shown in Figure 1, the main functional modules include the following:

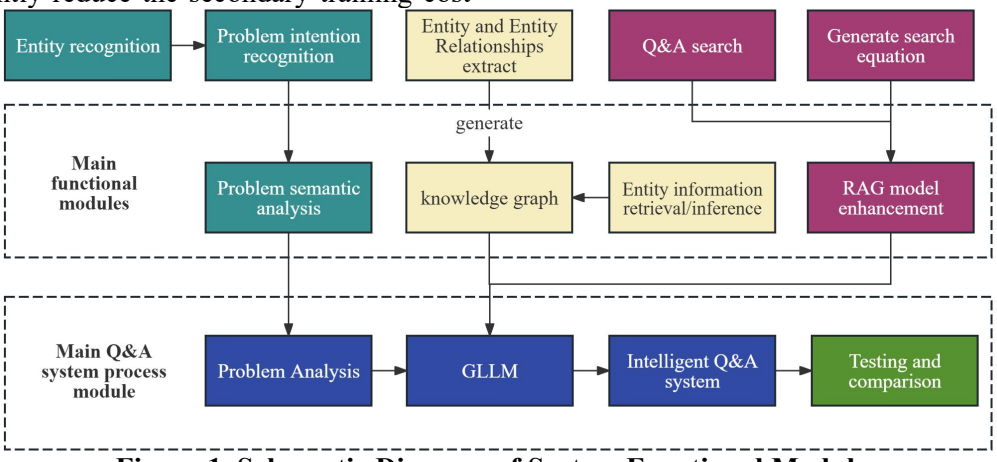

Figure 1. Schematic Diagram of System Functional Modules

2.1 Problem Semantic Analysis Module

Question Semantic Parsing accurately captures user intent by transforming their questions into a structured form that can be understood by computers. The semantic analysis of problems adopts two steps: entity recognition and question intention recognition. Firstly, Named Entity Recognition (NER) will be used to extract entities from the problem, and then the question intention will be determined from the entities, providing classification support for generating answer decisions [5].

Entity recognition is implemented using a

RoBERTa+BiGRU+Multi Attention+CRF composite model. The RoBERTa model is an improved design based on the BERT model. During the pre training process, the model enhances its generalization and segmentation ability by establishing dynamic MASK+whole word MASK (whole word mask) process, Byte Air Encoding (BPE) mixed representation, and other strategies [6], generating a word vector matrix from the original problem statement. The BiGRU model (Bidirectional Gated Recurrent Unit) is designed based on a recurrent neural network (RNN) structure, which includes bidirectional GRU processing sequences. It can capture long-distance dependencies and temporal relationships between words from the input word vector matrix, and combine with Multi Attention to achieve contextual relationship extraction. Finally, CRF (Conditional Random Field) further supplements the constraint relationships between words based on word vectors and contextual features, and forms entity labels and entity relationship information.

The problem intention recognition is implemented using the RoBERTa+DPCNN (Deep Pyramid Convolutional Neural Networks) model. RoBERTa converts the input problem text into an embedded vector containing semantic features and contextual relationships. The DPCNN model receives the RoBERTa output vector and extracts and integrates multi-level semantic features through convolutional and pooling layers, ultimately forming an intention probability distribution [7]. This model structure not only captures local phrase features in the question, but also has the ability to extract global features of long question documents, effectively improving the accuracy and robustness of intent recognition.

2.2 Knowledge Graph and Retrieval Module

To improve the performance of generative models, knowledge in the bidding field is identified as triplets and combined into a network knowledge graph. This structure can accept problem entity sequences and obtain entity relationships through entity information retrieval and inference processes, thereby enhancing knowledge graph compatibility [8]. A composite document retriever is embedded on the basis of traditional graph retrievers to be compatible with document knowledge retrieval functions. By converting the relationships between entities into embeddings, domain knowledge information is provided for generative models [9]. The composite document retriever can associate external documents with corresponding entities through entity linking. That is, after the entity recognition function identifies the entity in the problem, based on the problem intention recognition, entity recognition, and entity relationship information, it uses knowledge graph entity query and document similarity comparison functions to obtain the text block information corresponding to the entity. It binds the entity link to the text block information and fills the entity link object into the extension slot, providing external document extension information for the knowledge graph.

2.3 Model Enhancement Module

The study adopts RAG as a method for enhancing large language models, and selects appropriate strategies to enhance generative models by comparing the enhancement effects. The specific process includes: firstly, selecting relevant documents from the list shown in Table 1 as domain knowledge document data sources, which can cover bidding processes, bidding regulations and other domain documents, ensuring the representativeness of the data sources; Secondly, document preprocessing is used to decompose the document into text blocks according to paragraph structure, and questions are designed based on the content of the text blocks to generate "question answer" pairs; Thirdly, identify entity types based on knowledge types to form keywords, and integrate them with entity keywords in the knowledge graph to establish a search equation. Download the full text from the website shown in Table 1 through the search equation, and obtain 147 documents through data filtering and preprocessing; Fourth, use entity analysis models combined with manual review to annotate the corpus.

Table 1. Classification of Q&A Samples

Problem type	Document quantity	Paragraph block	Problem sample quantity	Entity
bidding process	4	875	76	254
tendering regulations	23	2001	120	237
bidding documents	11	253	54	63

project management	17	2227	84	118
evaluation related	8	248	43	68
feedback and appeal	13	702	21	64
industry standard	34	3264	55	125
bidding Case	37	629	47	94

3. Experiments

3.1 Sample Preparation

3.1.1 Preparation of domain documents

As shown in Table 1, the samples are divided into 8 categories according to the bidding process, bidding regulations, and other topics, respectively from Global Tenders, Global Public Procurement Database, Tendersinfo, and Jus Mundi database, including 147 documents, divided into 10199 text blocks according to natural paragraphs.

3.1.2 Q&A on sample preparation

32 documents were randomly selected based on topic classification. Firstly, 8 documents were manually annotated using YEDDA tool, and the remaining 24 documents were pre annotated using the intelligent algorithm BiLSTM CRF. After manual verification, 1023 entity sample information was formed. Based on the topic document and combined with TenderBoard Asia, Bidding Source, and GLOMACS online Q&A content, 500 Q&A pairs were developed by domain experts. The Q&A pairs were in the form of triplets, including questions, search context, and answers. Each question was independently annotated and verified by two experts to ensure its correctness.

3.2 Entity Recognition Testing

As shown in Figure 2, the RoBERTa+ BiGRU+MultiAttention+CRF model is defined. The model development language and machine learning framework are Python 3.7 and TensorFlow 1.14, respectively. The training is based on the Ubuntu 14.04 system platform, with AMD Ryzen 7 3700X processor and NVIDIA GeForce RTX 3090 GPU. Using the initialization model parameters shown in Table 2, the training set, test set, and validation set were set at 7:2:1, respectively. The training sample problems were input into the RoBERTa+BiGRU+Multi Attention+CRF model, and word labels were formed through processes such as word vector generation, contextual semantic extraction, and semantic feature supplementation. Through iterative training and validation optimization, the optimal model was achieved.

Table 2. Entity Recognition Model Initialization Parameters

Parameter Name	Parameter Value
optimizer	Adagrad
learning rate	1e-5
BiGRU hidden layer	128
epochs	30
batch_size	20

The main title (on the first page) should begin from the top edge of the page, centered, and in Times New Roman 16-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Please initially capitalize only the first word in other titles, including section titles and first, second-order headings (for example, “Titles and headings” — as in these guidelines). Leave two blank lines after the title.

After completing the training of the model, CRF and BERT BiGRU models were selected as the control groups, and the test sets were input separately. Five senior (over 5 years old) tenderers were selected for double-blind evaluation of the test results, and accuracy, recall, and F1 were selected to evaluate the model results.

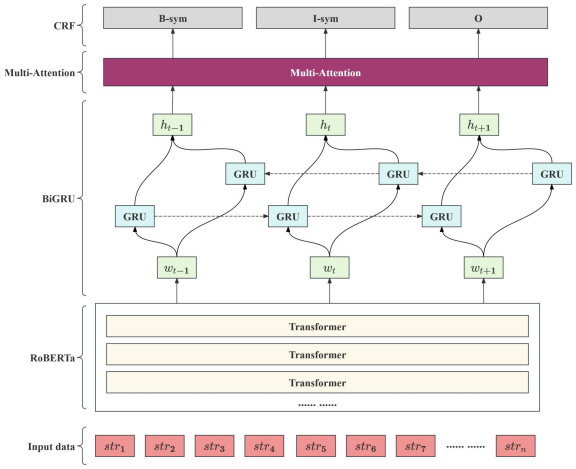

Figure 2. Network Structure of RoBERTa+BiGRU+Multi Attention+CRF Model

3.3 Problem Intent Recognition Test

In the Q&A process of recruitment and bidding,

using question intention recognition based on word label analysis provides question classification features for knowledge graph retrieval and generative models, which helps improve answer accuracy. Define the network structure of the RoBERTa+DPCNN model as shown in Figure 3, and initialize the model parameters as shown in Table 3. The training set, test set, and validation set are set at 7:2:1, and the questions are labeled as 0-7 according to topic types. The question statements are input into the RoBERTa+DPCNN model, and after generating word vector sequences, extracting local semantics from vectors, and fusing word vector features, classification labels are generated. Through iterative training and validation optimization, the model is optimized until it reaches the optimal level.

Table 3. Initialization Parameters of Problem Intent Recognition Model

Parameter Name	Parameter Value
optimizer	Adagrad
Learning_rate	1e-3
dropout_rate	0.3
Kernel_size	3
epochs	30
batch_size	40

After completing the training of the model, TextCNN and BERT TextCNN models were selected as the control groups. The test sets were input separately, and 5 senior (over 5 years old) tenderers were selected for double-blind evaluation of the test results. Accuracy, recall, and F1 were selected to evaluate the model results.

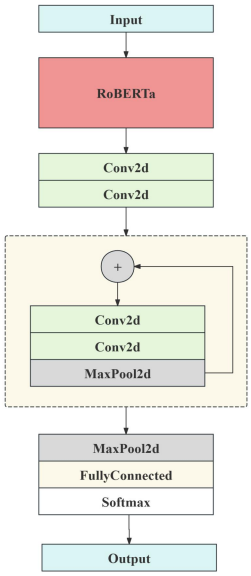

Figure 3. RoBERTa+DPCNN Network Structure

3.4 Intelligent Q&A Generation Test

Intelligent question answering generation was compared using traditional LLM, model fine-tuning LLM, and knowledge graph+RAG+LLM. The LLM model selected LLaMA2, which has the advantages of context information aggregation mechanism, strong language generation ability, high inference efficiency, low secondary training cost, and supports 128k long context windows. It has been applied in intelligent question answering systems in multiple fields. All 500 samples were selected and divided into a training group and a testing group according to a 4:1 ratio. The test results were evaluated by 5 senior (over 5 years old) tenderers in a double-blind manner, and answer accuracy, information completeness, text consistency, and irrelevant question rejection were selected as the evaluation model results. The indicators were set to 0-4 levels, and the higher the score, the better the performance of this item.

LLM control group: Directly input 100 question texts into LLaMA2, and have experts evaluate the output results.Model fine-tuning LLM group: LoRA (Low Rank Adaptation) was chosen as the LLaMA2 model fine-tuning algorithm, which can reduce the training cost of the LLM model and improve its generalization ability by decomposing the high-dimensional weight matrix of the model into a low rank matrix [4]. Firstly, gte-base-en-v1.5 is selected as the Embedding model, which has good performance in semantic classification, information retrieval, sorting, relationship extraction, and similarity comparison, as shown in Table 4 for initializing LoRA parameters; Input 400 samples into the training script to complete the LLaMA2 model training; Input 100 test samples into the LLaMA2 model and have experts evaluate the output results.

Knowledge Graph+DAG+LLM: Firstly, based on the bidding theme definition schema, 1023 entity nodes and 11046 entity relationships were extracted through entity samples RoBERTa+BiGRU+Multi Attention+CRF. The graph data storage was implemented based on the Neo4j graph database, and the Neo4j database was constructed using Py2neo tool [10]. Implement entity and associated document queries using Cypher statements and a composite document retriever [11]. Input 500 training samples corresponding to documents

into RAG, process them through vectorization, and start the testing process with 100 test samples.

Table 4. LoRA and Training Initialization Parameters

Parameter Name	Parameter Value
low_rank_matrix_rank	16
scaling_factor	16
dropout_rate	0.2
learning_rate	1e-5
epochs	20
batch_size	10
weight_decay_rate	1e-2

4. Results

4.1 Entity Recognition Results

The entity recognition results are shown in Table 5. In the accuracy test, RoBERTa+ BiGRU+ MultiAttention+CRF had the highest accuracy, reaching 64.00%. BERT BiGRU and RoBERTa+ BiGRU+Multi Attention+CRF had accuracies 5.00% and 9.00% higher than the CRF model, respectively; In the recall test results, RoBERTa+ BiGRU+MultiAttention+ CRF had the highest accuracy, reaching 61.54%. Both BERT BiGRU and RoBERTa+ BiGRU+ Multi Attention+CRF had higher accuracy than the CRF model by 3.59% and 6.64%, respectively; In the RoBERTa+BiGRU +MultiAttention+CRF test, F1 was the highest, which was 3.22% and 8.55% higher than the BERT BiGRU and CRF models, respectively.

Table 5. Comparison Table of Entity Recognition Model Testing

model name	accuracy	recall	F1
CRF	0.5500	0.5490	0.5545
BERT-BiGRU	0.6000	0.5849	0.6078
RoBERTa+BiGRU+ Multi-Attention+CRF	0.6400	0.6154	0.6400

4.2 Problem Intent Recognition Results

The entity recognition results are shown in Table 6. In the accuracy test, RoBERTa+ DPCNN has the highest accuracy, reaching 74.00%. LatticeLSTM and RoBERTa+ DPCNN have higher accuracies than the TextCNN model by 8.00% and 13.00%, respectively; In the recall test results, RoBERTa+DPCNN had the highest accuracy, reaching 75.47%. LatticeLSTM and RoBERTa + DPCNN had higher accuracies than TextCNN models by 10.13% and 14.69%, respectively; In the RoBERTa+DPCNN test, F1 was the highest, 3.91% and 14.08% higher than the LatticeLSTM and TextCNN models, respectively.

Table 6. Test Comparison Table for Problem Intent Recognition Model

model name	accuracy rate	recall rate	F1
TextCNN	0.6100	0.6078	0.6139
LatticeLSTM	0.6900	0.7091	0.7156
RoBERTa+DPCNN	0.7400	0.7547	0.7547

4.3 Intelligent Q&A Generation Results

The results of intelligent Q&A generation are shown in Table 7. In terms of comparison of answer accuracy, the KG+ RAG+LLM test score is 4.7, which is higher than the LoRA+LLM group by 0.8. The RAG performance is slightly better than the fine-tuning model, but significantly higher than the LLM group score. KG+RAG+LLM provides relevant prompt word retrieval based on knowledge graph and RAG collaboration, providing LLM with inherent rules and dynamic domain knowledge information. At the same time, RAG embeds an automatic verification mechanism, which can return LLM generated content to the knowledge graph and document retrieval for comparison and verification, improving the accuracy of answers. In the information integrity comparison test, the KG+RAG+ LLM score was 3.3, which was higher than the LoRA+LLM group and LLM group by 0.2 and 0.5, respectively. RAG technology supports multiple iterative retrieval, which can be improved step by step through each retrieval result, reducing information omission and enhancing the completeness of generated answers [12]; In the information consistency test, the KG+RAG+LLM score was 4.5, which was higher than the LoRA+LLM group and LLM group by 0.7 and 1.4, respectively. RAG can achieve accurate local data matching and can form clear prompt word basis through document block indexing and query parsing [13-15]. The generated answers have a high degree of restoration and improve the consistency of answer output; However, the fine-tuning model based on LoRA technology has limited text alignment ability due to the limited number of updated parameters. When the question exceeds the knowledge of the retrieved document, the KG+RAG+LLM score is 3.4, significantly higher than the scores of the

LLM and LoRA+LLM groups, indicating that the model can target the semantics and intention of the question, and can combine knowledge graph and RAG document index to retrieve whether the question exceeds its own knowledge scope, without directly providing fuzzy responses, showing better refusal ability; LLM and LoRA+LLM groups tend to provide fuzzy responses that lack targeted knowledge details when faced with problems related to knowledge beyond knowledge [16].

Table 7. Comparison Table for Intelligent Q&A Generation Test

group	accuracy	completeness	consistency	refusal to irrelevant
LLM	3.3	2.8	3.1	1.3
LoRA+LLM	3.9	3.1	3.8	1.8
KG+RAG+LLM	4.7	3.3	4.5	3.4

5. Discussion

5.1 Discussion on Entity Recognition Analysis

In entity recognition comparison tests, compared with traditional CRF models, BERT BiGRU and RoBERTa+BiGRU+Multi Attention+CRF models adopt the BERT model infrastructure, which has bidirectional encoding ability and good transferability, and can simultaneously capture contextual information before and after text. The CRF model has a high dependence on feature engineering. Although it can capture the correlation between entities, it lacks the ability to recognize contextual semantics, resulting in low accuracy in recognizing long text entities and limited semantic acquisition capabilities; At the same time, regularization methods need to be added during the recognition process to reduce the risk of overfitting, which requires high tuning requirements and limits the stability of model performance. The RoBERTa model improves language modeling performance by removing NSP tasks, while also increasing the training data size. Compared with BERT BiGRU, it improves the accuracy of text context information extraction, with an F1 improvement of 3.22% in comparative testing. The BiGRU model inherits the advantage of reducing gradient explosion from the GRU model, and also has bidirectional understanding and memory capabilities. It can simultaneously connect the contextual environment features and dependencies before and after, deeply extract text features, and enhance the robustness of handling default values and noisy data [17]. The Multi Attention structure not only further enhances the overall processing range of contextual information in the model, but also facilitates the extraction of global sentence features. By embedding the CRF structure, it deeply extracts hidden text features and accurately obtains text index weight information.

5.2 Discussion on Problem Intention Recognition

The study introduces convolutional neural network models into the task of problem intention recognition. Compared with the TextCNN model, the DPCNN model has the advantages of simple structure, strong interpretability, and the ability to mine contextual relationships between text sequences over longer distances [18]. Therefore, the study attempts to use the RoBERTa+ DPCNN method to solve the problem of problem intention recognition. By comparing the results, RoBERTa DPCNN outperforms other control group test results. The reasons for this include: firstly, the RoBERTa+DPCNN model performs better than TextCNN and LatticeLSTM, indicating that the RoBERTa model can capture contextual relationships and deeper semantic features between texts at longer distances, which is more conducive to intent analysis; Secondly, the DPCNN model adopts a deeper convolutional neural network structure and draws on ResNet's shortcut design, which can further expand the network structure and enhance the gradient transfer between network layers, which is beneficial for complex text feature analysis; Thirdly, the DPCNN model adopts a convolutional double-layer stacking strategy and region embedding to improve the performance of word position embedding. It combines adaptive weighting with dual view information to improve text feature representation. That is, the adaptive dual view weighted clustering algorithm is used to obtain the weights of word sequence relationships and attributes, and then an unsupervised method is used to construct the feature space between word vectors; Fourthly, DPCNN adopts equal

length convolution to iteratively capture word vectors and their contextual features, and gradually mine higher-level semantic features layer by layer.

5.3 Intelligent Q&A Generation Discussion

The study selected traditional LLM and fine-tuning LLM based on LoRA method for comparative research. Through testing, the KG+RAG+LLM method performed better in terms of accuracy, completeness, consistency, and refusal ability of answer generation. At the same time, the combination of knowledge graph and RAG can integrate traditional domain knowledge rules and document update information, ensuring the integrity of domain knowledge. Compared to using LoRA fine-tuning LLM method, its iterative update cost is lower; In terms of development efficiency, the KG+RAG+LLM method is more efficient, consumes less computing power, and is conducive to expanding knowledge in multiple fields and transferring applications. Due to the complex network structure and large parameter scale of LLM models, using LLM for secondary training or fine-tuning to achieve domain transfer has problems such as high development costs and complex sample preparation in the early stage; At the same time, when using LLM directly to generate Q&A responses, there are deficiencies in the consistency of answer facts, information completeness, and rejection of irrelevant questions. When faced with professional Q&A needs, the answers lack accuracy and cannot directly meet the objective and rigorous knowledge Q&A scenarios in fields such as bidding and medical consulting. Currently, research generally combines knowledge graphs with LLM to generate high-quality and accurate responses and improve their performance.

5.4 Localized Application Testing and Discussion

This study systematically compared two model enhancement techniques, fine-tuning and RAG, in localized application scenarios such as bidding Q&A and technical consulting. Through experimental evaluation, it was found that the RAG method performs slightly better than the fine-tuning model in overall performance, specifically in terms of answer accuracy, text alignment quality, and refusal ability; At the same time, it also has significant advantages in data preparation and implementation complexity. In terms of resource consumption, RAG exhibits significant computational efficiency.

However, due to the fact that the data used in this experiment is mainly linear structured documents and the number of non relational structured documents is relatively small, the generalization ability of the two methods in complex corpus types still needs further verification. In further research, we plan to further explore the following directions: on the one hand, optimize text segmentation strategies and develop more efficient Embedding models to enhance the retrieval and generation performance of RAG; On the other hand, combining knowledge graphs, RAG, and various fine-tuning strategies can further improve model performance while controlling computational resource investment.

6. Conclusion

The research focuses on the shortcomings of bidding Q&A and technical consultation in the bidding process, aiming to significantly improve the transparency, efficiency, and quality of the bidding process by eliminating information asymmetry, unifying the starting point of competition, and promoting effective coordination between the bidding and procurement parties. With the continuous expansion of global bidding scale and the rapid growth of related information, traditional manual consulting and general search engines are no longer able to cope with the increasingly complex professional information needs. The intelligent question answering system supported by LLM and text mining technology provides new possibilities for efficiently responding to professional consultations in the field of bidding and tendering. However, traditional big language models suffer from high secondary training costs, the phenomenon of "illusion" in generated content, and insufficient accuracy in professional contexts in practical deployment, which limits the deep application of LLM in this business field. To address the above challenges, a technical method combining knowledge graph and RAG is proposed to enhance the comprehensive efficiency of GLLM in intelligent Q&A in bidding and tendering. The study first introduces a knowledge graph to construct a domain specific knowledge base, and based on a large language

model for joint learning, strengthens the understanding and expression ability of bidding terms, processes, and norms; Secondly, by utilizing the RAG mechanism, external knowledge sources are dynamically introduced during the generation process to enhance the model's ability to reason and answer complex problems. Through comparative experiments, it has been found that compared to traditional manual consultation and search engine methods, the intelligent question answering system that integrates knowledge graph and RAG has significantly improved response accuracy and efficiency, providing feasible ideas and practical references for building a new generation of intelligent bidding consultation system.

References

[1] Zhang L , Chen L , Liu W .Research and Application of a Question and Answer System for Bidding Based on Knowledge Graph. 2024 2nd International Conference on Signal Processing and Intelligent Computing (SPIC), 2024:928-932.DOI:10.1109/spic62469.2024.10691468.

[2] Zhang M , Yang G , Liu Y ,et al.Knowledge graph accuracy evaluation: an LLM-enhanced embedding approach. International Journal of Data Science and Analytics, 2025, 20(3):3021-3035.DOI:10.1007/s41060-024-00661-3.

[3] Kim B , Yang J .Comparative Analysis Study on Automated Dataset Generation Frameworks for RAG System Performance Evaluation. The Journal of Korea Institute of Information, Electronics, and Communication Technology, 2025, 18(2):143-154.DOI:10.17661/jkiiect.2025.18.2.143.

[4] HAYOU S,GHOSH N,YU B. LoRA+:efficient low rank adaptation of large models.（2024-02-19）[2024-08-01]. https://arxiv. org/abs/2402. 12354.

[5] Zhan L , Huang C .Research on Computer Natural Language Processing Intelligent Question Answering System Based on Knowledge Graph. 2024 International Conference on Machine Intelligence and Digital Applications, 2024:70-74.DOI:10.1145/3662739.3664744.

[6] Kai Z , Qibo W , Lining W ,et al.Sentiment Analysis of Chinese Bullet Comments Based on Enhanced Deep Learning Model. 2024 IEEE 5th International Conference on Pattern Recognition and Machine Learning (PRML), 2024:481-489.DOI:10.1109/prml62565.2024.10779914.

[7] Wang H , Zhang S .Research on the Application of Improved BERT-DPCNN Model in Chinese News Text Classification. CONCURRENCY-PRACTICE AND EXPERIENCE, 2025, 37(3).DOI:10.1002/cpe.8338.

[8] Zeng, P., Yuan, LJ., Intelligent Q&A System Based on Bee Knowledge Graph. Information Technology and Informatization, 2023 (7): 108 - 111. doi: 10.3969/j.issn.1672-9528.2023.07.027

[9] Gao W, Zheng X, Zhao S. 2021. Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF. Journal of Physics: Conference Series. doi: 10.1088/1742-6596/1848/1 /012083.

[10] Shaoxiong J, Shirui P, Erik C, et al. A survey on knowledge graphs: representation,acquisition, and applications. IEEE transactions on neural networks and learning systems, 2021, 33(2): 494-514.

[11] Wang Q , Li C , Liu Y ,et al.An Adaptive Framework Embedded With LLM for Knowledge Graph Construction. IEEE transactions on multimedia, 2025:27.DOI:10.1109/ TMM.2025. 3557717.

[12] Gürkan ahin, Varol K , Pak B K .LLM and RAG-Based Question Answering Assistant for Enterprise Knowledge Management. 2024 9th International Conference on Computer Science and Engineering (UBMK), 2024:1-6.DOI:10.1109/ubmk63289.2024.10773564.

[13] Hamdhana D .A LOW CODE APPROACH TO Q&A ON CARE RECORDS USING FLOWISE AI WITH LLM INTEGRATION AND RAG METHOD. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 2024, 9(4):2435-2445.DOI:10.29100/jipi.v9i4.6978.

[14] Mukhopadhyay P .Designing Conversational Search for Libraries: Retrieval Augmented Generation through Open Source Large Language Models. DESIDOC Journal of Library & Information Technology, 2025,

45(2).DOI:10.14429/djlit.20206.

[15]Liao L , Mo W , Wen Y ,et al.Safety Consultation for Prefabricated Construction: A Localized Retrieval-Augmented Generative Question-Answering System. Journal of Computing in Civil Engineering, 2025, 39(5).DOI:10.1061/JCCEE5.CPENG-6315.

[16]LIU Z, KUNDU S,LI A,et al. AFLoRA:adaptive freezing of low rank adaptation in parameter efficient fine-tuning of large models.（2024-03-20）[2024-08-01]. https://arxiv. org/abs/2403.13269.

[17] Xing Y , Tan J , Zhang R ,et al.Robust Anomaly Detection of Multivariate Time Series Data via Adversarial Graph Attention BiGRU. BIG DATA AND COGNITIVE COMPUTING, 2025, 9(5):122.DOI:10.3390/bdcc9050122.

[18] Guo X , Wang J , Gao G ,et al.Efficient Agricultural Question Classification With a BERT-Enhanced DPCNN Model. IEEE Access, 2024, 12:109255-109268.DOI:10.1109/access. 2024. 3438848.

