International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1057-1069

https://doi.org/10.70517/ijhsa46499

Design of Knowledge Base Q&A System Based on Retrieval Enhanced Generation Technique

Xiaohan Li[1,*]
[1] Computer School, Santa Clara University, Santa Clara, California, 95053, USA
Corresponding authors: (e-mail: xiaohanli1986@163.com).

Abstract In this paper, we propose a knowledge base Q&A system based on retrieval-enhanced generation (RAG) technology, which significantly improves the semantic comprehension ability and generation quality of the system by integrating bi-directional gated recurrent units (BiGRUs), neural state machines (NSMs), and hybrid RAG indexing optimization methods. To address the gradient decay problem of traditional recurrent neural networks (RNNs), BiGRU extracts deep features of the text through a bidirectional information transfer mechanism, while NSM simulates human causal thinking through probabilistic scene-graph reasoning to enhance the interpretability of the model. The hybrid RAG strategy further combines local knowledge base construction, context fusion and cue sample introduction to achieve dynamic knowledge enhancement. For indexing optimization, the HNSW-PQ composite indexing technique is adopted to significantly reduce the latency and storage overhead of high-dimensional vector retrieval. In the complex knowledge base quiz tasks (CWQ and WebQSP), the model F1 scores reach 70.34% and 83.31%, Hits@1 were 73.18% and 85.33% respectively, which are fully ahead of the traditional methods and pre-trained models. The experimental results show that through the semantic reasoning capability of BiGRU-NSM, the dynamic knowledge enhancement of hybrid RAG and the efficient retrieval of HNSW-PQ indexing, the system achieves breakthroughs in multi-hop reasoning, complex semantic parsing and generative accuracy, and provides an efficient, interpretable and adaptable solution for knowledge base quiz tasks.

Index Terms retrieval-enhanced generation technique, bi-directional gated loop unit, neural state machine, hybrid RAG index optimization, knowledge base quizzing

I. Introduction

With the rapid development of information technology, Q&A system, as an important application in the field of natural language processing, has been gradually integrated into people's daily life and become an indispensable part [1]. And with the continuous innovation of information technology and the growing demand for efficiency and accuracy of information acquisition, the domain-specific adaptability and Q&A accuracy of knowledge base Q&A systems become more important [2]-[4].

Artificial intelligence macrolanguage modeling has shown great potential in the field of Q&A systems by virtue of its superior language understanding and generation capabilities [5], [6]. Although many generalized bigram models and industry bigram models have emerged at home and abroad, these models all suffer from a certain degree of data obsolescence and corpus insufficiency [7], [8]. In contrast, retrieval-enhanced generation (RAG) technology can effectively solve the problems of data adaptability as well as corpus insufficiency by introducing external documents to enable the big models to access external knowledge bases so as to generate more authentic and reliable answers [9]-[11]. Therefore, how to utilize the existing big models and retrieval augmentation generation (RAG) techniques to quickly build a highly available knowledge base application system is an important way to achieve efficient and accurate retrieval of user-needed knowledge from massive information [12]-[14].

Focusing on the design of a knowledge base Q&A system that combines deep learning models with retrieval augmented generation (RAG) technology, this paper proposes an integrated approach that incorporates bidirectional gated recurrent unit (BiGRU), neural state machine (NSM), and hybrid RAG indexing optimization, aiming to improve the system's semantic comprehension ability, reasoning efficiency, and generation quality. The article firstly starts from the construction of the knowledge base Q&A model, and for the gradient decay problem of recurrent neural network (RNN) when dealing with long sequences, BiGRU model is introduced to extract the deep features of the text through the bidirectional information transfer mechanism. Meanwhile, in order to solve the defect of insufficient interpretability of deep learning models, the Neural State Machine (NSM) framework is proposed, which decomposes the inference process into two phases of learning and inference, simulates human causal thinking, and realizes the joint inference of semantics and images through probabilistic scene graphs. On this basis,

a hybrid RAG strategy is further proposed to enhance the real-time and accuracy of the generated model by constructing a local knowledge base, contextual information fusion and cue sample introduction. Finally, for the bottleneck problem of knowledge base retrieval efficiency, an optimization method based on sparse coding (backward indexing) and dense coding (HNSW-PQ composite indexing) is proposed to significantly reduce query latency and storage overhead.

II. Knowledge base Q&A model based on BiGRU and NSM and hybrid RAG index optimization approach

II. A.Knowledge Base Q&A Model

II. A. 1) Bi-directional gated circulation units

Due to the uncertainty of the time step, Recurrent Neural Networks (RNNs) experience gradient decay and explosion when dealing with the relationship between time step distances in a time series, leading to a decrease in the model's learning ability. To address this problem, gated recurrent neural network (GRU) is proposed, which is able to understand and recognize the relationship between different time-step distances in a time series and use it to control the information transfer. On the basis of GRU, Bidirectional Gated Recurrent Neural Network (BiGRU) was proposed for enabling the output of the current moment to be associated with both the state of the previous moment and the state of the next moment, extracting the deep features of the text.BiGRU is a neural network model consisting of two unidirectional, directionally opposing, neural networks whose outputs are jointly determined by the states of these two GRUs. At each moment, the input is provided with two GRUs in opposite directions at the same time and the output is jointly determined by these two unidirectional GRUs.The framework of BiGRU model is shown in Fig. 1.

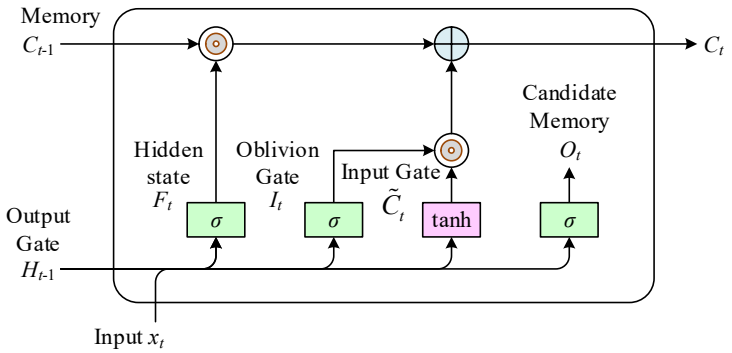

Figure 1: BiGRU model framework

II. A. 2) Neural state machines

Deep learning based neural network models have better performance, more adaptability and stronger robustness than traditional models. However, based on the black-box characteristics of deep learning models, its generalization ability is limited by unfixed parameters and difficult to implement, as well as the low degree of visualization of the model, which also leads to the model is difficult for human minds to learn the cause and effect relationships in the problem. At the same time, based on the huge scale of neural network models and the need for a large amount of data for training, it also reduces their interpretability, modularity and rationality. Based on this, Neural State Machine (NSM) has been proposed, which has the ability to miniaturize and simulate human operational thinking behavior. The Neural State Machine model is divided into two phases: a learning phase and a reasoning phase. In the learning phase, the model generates probabilistic scenario graphs capable of capturing its own latent semantic knowledge in a compact structure, based on the given knowledge. The model then views the graph as a state machine and simulates its iterative computation to answer questions or draw inferences.

During the modeling process, the researchers modeled images and language as abstract representations. Images are represented by probabilistic graphs that represent their semantics - including the goals, attributes, and relationships represented in the image. Questions, on the other hand, are converted into a sequence of inference instructions.

In the inference phase, the researchers treated the graph as a state machine, with nodes representing goals in the image, corresponding to states, and edges representing relationships between goals, corresponding to transfers. The researchers later initiate a sequence computation that iteratively feeds the instructions extracted from the problem into the machine and changes the state, allowing the model to perform semantic-image inference and eventually arrive at a result.

II. B.Retrieval Enhancement Generation Method

Although BiGRU and NSM models significantly improve semantic understanding and reasoning, the accuracy and real-time performance of the generated results still need to rely on the dynamic enhancement of external knowledge. For this reason, this section further proposes a hybrid RAG strategy to realize the accurate governance of knowledge base Q&A systems through the deep integration of retrieval and generation.

The Retrieval Augmented Generation (RAG) strategy is a natural language processing approach that combines retrieval and generation capabilities. It aims to augment the generative model by retrieving existing high-quality information to produce more accurate, richer, and coherent text. This section contains the detailed process of hybrid RAG strategy, including 3 key steps of RAG knowledge base construction, contextual information fusion, and cue sample introduction. The detailed steps of the hybrid RAG retrieval framework are shown in Fig. 2. The RAG strategy is at the core of the whole model architecture, which realizes the precise governance of online public opinion through the deep combination of dynamic information retrieval and generation process. The framework provides a feasible and efficient solution for handling complex and rapidly changing public opinion data.

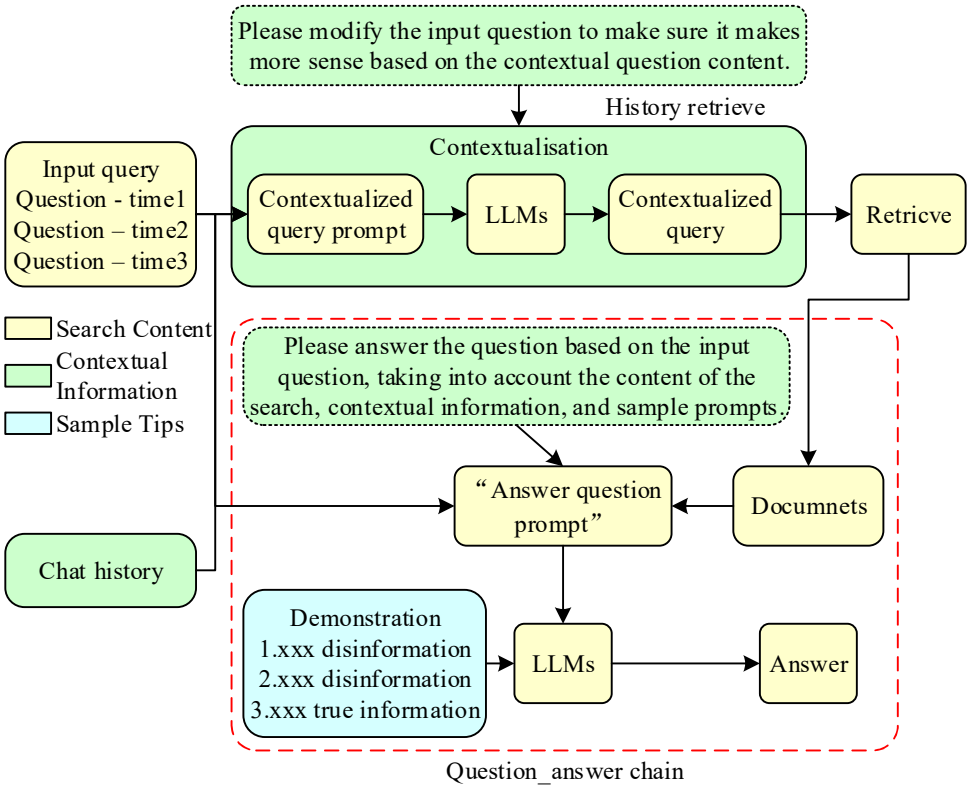

Figure 2: Hybrid RAG retrieval framework

(1) Constructing a local RAG knowledge base: By collecting and integrating structured and unstructured data in related fields, a high-quality knowledge base covering multiple public opinion events is formed. The knowledge base not only contains real-time information of current public opinion, but also includes historical data and related background information, providing rich retrieval resources for the model. Providing real-time information support for the model ensures that accurate and relevant information is obtained in real time when generating content, effectively enhancing the reliability and accuracy of the generated content.RAG retrieval uses cosine similarity, and the calculation process is as follows:

$$similarity(q,d_i)=\frac{q\cdot d_i}{\|q\|\|d_i\|} \tag{1}$$

where q is the embedding vector that converts user questions into embedding vectors, d_i is the embedding vector that converts each document in the retrieval repository, and i denotes the index of the document, and finally the Top-k documents with the highest similarity are selected.

(2) Contextual information fusion: in the input stage, the model combines contextual information. Through the integration of contextual information, the model's comprehension of the input data is significantly improved, so that it can better maintain logical coherence and information accuracy during content generation. It not only improves the model's ability to adapt to complex public opinion environments, but also enhances its performance in analysis and generation.

(3) Introduction of cue samples: By using appropriate cue samples, the model can generate outputs with specific styles and contents, thus improving the quality of the generated results. It helps to improve the performance of the model in the case of fewer samples and increases the applicability and flexibility of the model in diverse tasks.

The introduction of RAG strategies for knowledge enhancement based on contextual learning and cued samples has become a new paradigm for solving natural language processing problems using large language models. By basing on contextual instruction commands and a small number of examples, task instructions in different domains can be learned and text prediction can be accomplished. The complete hybrid RAG strategy, context learning commands and cue samples can be formally defined as follows:

$$C = \left\{ I, (x_1^t, d, k_1, y_1^t), (x_2^t, d, k_n, y_2^t), \cdots, (x_n^t, d, k_n, y_n^t) \right\} \tag{2}$$

where C is the complete instruction command used in this paper, I is the instruction command in different downstream tasks of online opinion analysis, d is the detailed description of the domain to which the downstream task belongs, k_n is the cueing samples, and y_n^t is the content of the knowledge retrieved by the RAG strategy. Finally, the RAG retrieved knowledge, contextual information, and cue samples are integrated and inputted into the large language model.

II. C.Knowledge Base Retrieval Enhancement Design - Indexing Optimization Approach

The efficient operation of hybrid RAG strategies relies on the fast retrieval capability of the knowledge base, while traditional indexing methods are difficult to cope with the real-time query demand of massive high-dimensional vectors. In this section, we focus on the indexing optimization techniques of sparse and dense coding to provide low-latency and high-precision retrieval support for the RAG strategy through HNSW-PQ composite index design.

In the knowledge base retrieval enhancement approach for language modeling, since both BM25 and encoder-based retrieval models do not need to encode documents in real time, the encoding of documents and the encoding of queries are set to work in parallel.

In the experiments, the prepared knowledge bases are sparsely and densely encoded separately in advance. For sparse vectors an inverted index is used as shown in Fig. 3, where each indexed lexical item (or keyword) is mapped to a list of documents containing that lexical item. In other words, it is an indexing structure that associates the lexical items in the document collection with document IDs, rather than associating document IDs with lexical items (orthogonal indexing).

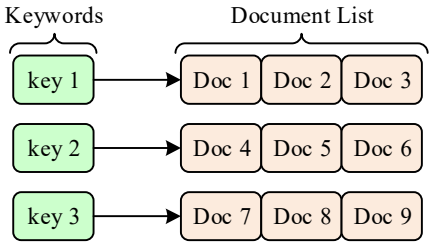

Figure 3: Inverted index

For dense vector indexing is constructed based on a vector retrieval method combining Hierarchical Navigable Small World (HNSW), Inverted File Indexing (IVF) and Product Quantization (PQ). The idea of navigable small world (NSW) comes from the connectivity in social networks, which speeds up the query by constructing networks with small-world navigability properties by creating local neighbors and global neighbors, but may fall into error local minima.HNSW separates the connectivity relationships based on the length of the connections using a hierarchical NSW graph, where the top layer of the graph is a subset of the bottom layer and the bottom layer contains all of the elements, such that the structure of the network is capable of maintain the local structure in low-dimensional space while having global connectivity in high-dimensional space. When querying, the greedy algorithm is used to start from the top layer and gradually query to the lower layers, and the nearest neighbor is selected as the entry point

of the next layer in each layer, so that the complexity of the search process is reduced to the logarithmic level, and the working principle of the HNSW is shown in Fig. 4.

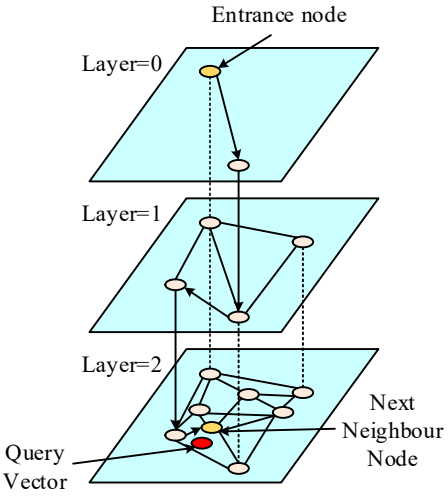

Figure 4: The working principle of HNSW

The disadvantage of HNSW is that the construction and storage of the graph structure leads to a relatively large memory overhead, PQ is developed from the basis of vector quantization (VQ), the indexes of composite IVF and PQ decompose the vectors in the high-dimensional space into several lower dimensional subvectors and then quantize each of the subvectors independently, the principle of the PQ operation is shown in Fig. 5.

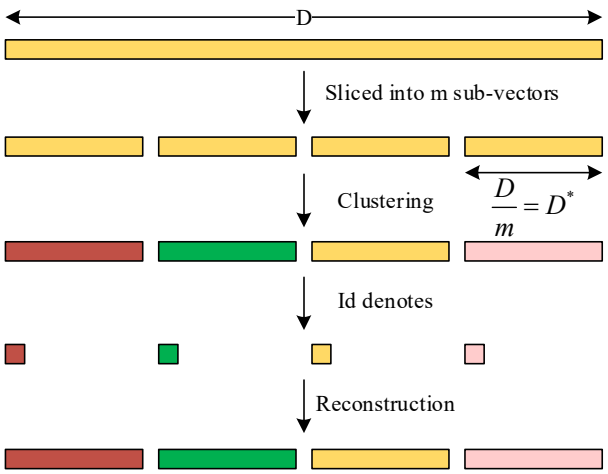

Figure 5: The working principle of PQ

PQ divides the input high-dimensional vector $X = (x_1, \ldots, x_D)$ into m sub-vectors by dimension, each with dimension $D^* = D / m$, where D is a multiple of m, and the sub-vectors are processed using m vector quantifiers $u_1, \ldots, u_m$, with the formal representation shown in equation (3).

$$\begin{cases} u_1(x) = x_1, \ldots, x_{D^*} \\ \cdots \\ u_m(x) = x_{D-D^*+1}, \ldots, x_D \end{cases} \rightarrow q_1(u_1(x)), \ldots, q_m(u_m(x)) \tag{3}$$

where q_j is the quantizer of the j th subvector, and each subvector space has k^* centers of mass, which PQ calls reproducible values, denoted by c_{ji}, and for q_j the corresponding set of indexes l_j and the codebook

$C_j = \{c_{ji} \mid i = 1,\ldots,k^*\}$, the Cartesian product of the m sub-vector codebook constitutes the total codebook $C = C_1 \times \ldots \times C_m$, which is used to map to the complete center of mass. When $m = 1$, the PQ degenerates into a normal k-Means VQ, and when $m = D$, the PQ becomes of scalar quantized (SQ) form.

PQ has the advantage of representing a large number of centroids using multiple small sets of centroids, which reduces storage requirements and speeds up distance computation. After encoding the raw vectors using the PQ algorithm, the computation of distances to neighboring points during the insertion process of the HNSW algorithm needs to be modified accordingly to the distance computation of PQ.

After the indexing of the knowledge base is completed, the query is performed by calculating the similarity between the query encoding and the document encoding, and the retrieval of enhanced documents is returned for a set k , and the query latency increases with k during the query evaluation.

III. Experimental Validation of Knowledge Base Q&A System Based on Hybrid RAG with HNSW-PQ Index Optimization

The BiGRU-NSM model and hybrid RAG strategy proposed in Chapter 2 lays a theoretical foundation for knowledge base Q&A systems, while its practicality and generalization ability need to be verified through multi-dimensional experiments. In this chapter, we will focus on the railroad domain and complex knowledge base Q&A scenarios, and systematically evaluate the comprehensive performance of the model in dynamic retrieval, semantic reasoning, and generation quality through comparative experiments, parameter analysis, and ablation studies.

III. A. Railroad field knowledge quiz search

In order to verify that the knowledge base Q&A model and retrieval enhancement generation method based on bi-directional gated recurrent neural network proposed in the article have better performance on knowledge retrieval tasks, using the railroad domain Q&A system as a dataset, this section sets up experiments on the constructed railroad domain Q&A dataset and compares it with several different models to verify the effectiveness of the proposed model.

III. A. 1) Data sets

The knowledge Q&A retrieval dataset in the railroad domain constructed in the article is derived from a dedicated dataset constructed based on the Q&A system in the railroad domain, which contains structured knowledge base data (e.g., railroad timetable, station information, operation rules, etc.) and unstructured natural language Q&A pairs. It covers multiple types of questions in the railroad domain, such as train timetable query, distance calculation between stations, troubleshooting process, and so on.

III. A. 2) Experimental setup

The knowledge retrieval model based on bidirectional gated recurrent neural networks employs BERTbase as query encoder and paragraph encoder for obtaining vector representations of query questions and knowledge paragraphs. The experiments were conducted through a grid search method to determine the optimal hyperparameter configuration: batch size B $\in$ {8,16,32,64,128} and learning rate lr $\in$ {1e-5,1e-4,1e-3,1e-2}. The number of training rounds is 500, in which the warm-up phase contains 800 iteration steps, during which the learning rate is gradually increased from zero to the preset value lr, and then gradually decreased to zero as the training progresses. The specific hyper-parameter settings for this experiment are shown below, with a word vector dimension h of 783, a maximum sentence length K of 524, a learning rate of 0.001, an epoch of 500 training rounds, and a batch size of 16.

III. A. 3) Evaluation indicators

In order to validate the accuracy of the knowledge retrieval model, this chapter uses Recall and EM as the basis for evaluating the performance of the model.

Recall is an important evaluation metric in the field of knowledge retrieval, which is used to measure the ability of a retrieval system to return relevant knowledge. Recall is a measure of comprehensiveness, i.e., how much relevant knowledge the retrieval system is able to find, and the higher the recall, the less the retrieval model misses relevant knowledge.

EM is used as a basis for evaluating the performance of the model.The EM metric indicates the percentage of generated answers that exactly match the standard answer, and is used to measure the Q&A accuracy.

III. A. 4) Comparison of baseline models

In order to better evaluate the performance of other knowledge retrieval models on the Q&A dataset in the railroad domain, six classical models are selected for the comparative experiments of knowledge retrieval evaluation metrics, which are described in detail as follows.

(1) The DPR model, a classical model for intensive retrieval tasks, is a deep learning-based retrieval model that achieves encoded representations and matching retrieval of query questions and knowledge passages by encoding them with two separate encoders.

(2) The ANCE model adopts the same idea and loss function as the DPR model, and mainly focuses on the construction of negative samples during the training of dense retrieval models. The model optimizes the learning process of intensive retrieval by constructing negative samples from approximate nearest neighbor (ANN) indexes.ANCE updates ANN indexes in parallel during the learning process to select more effective negative training samples. This approach alleviates the mismatch between the dense retrieval data distribution in the training and testing phases.

(3) The RocketQA model introduces the adoption of a recaller and a rescheduler in knowledge retrieval, and adopts the strategy of cross-batch negative sample acquisition and difficult negative sample acquisition in the recaller, and adopts the data enhancement approach, which utilizes the cross-coder to generate pseudo-labels on the large-scale unsupervised data for the training of the Twin Towers encoder.

(4) The RocketQAv2 model is an improved version of RocketQA, and the main improvement strategy is to jointly train the recaller and rearranger in RocketQA, and unify the training of the two into listwise mode. Dynamic listwise distillation and data augmentation are introduced in the strategy of joint training, in which data augmentation is based on the results of RocketQA, using the recaller to construct random samples, and using the rearranger to construct noise reduction samples, and the two are unified to construct hybrid training data.

(5) The PAIR model introduces paragraph-centric contrast loss based on the traditional DPR model, which is used to assist in optimizing the vector representation of the passages in the pre-training phase, and only the traditional query-centric contrast loss is used in fine-tuning. In addition, PAIR also borrows the idea of RocketQAv2, which utilizes cross-coders for data re-labeling to achieve data enhancement.

(6) ConvGQR model focuses on the improvement of the query question side and proposes a new framework that utilizes generative macromodels to rewrite the query question and generate potential answers to enhance the retrieval of the searcher. In addition to this ConvGQR improves retrieval by utilizing a generator to generate query questions related to knowledge passages and later using the generated query questions.

III. A. 5) Analysis of the results of comparative experiments

In order to verify the superiority of the validated model in this paper, comparative analysis experiments on the proposed railroad domain knowledge base and the proposed domain Q&A dataset with the mainstream models mentioned above on the knowledge retrieval task are conducted, and Fig. 6 demonstrates the recall rate under recalling 50 and 100 relevant documents and the EM metrics under different candidate lists.

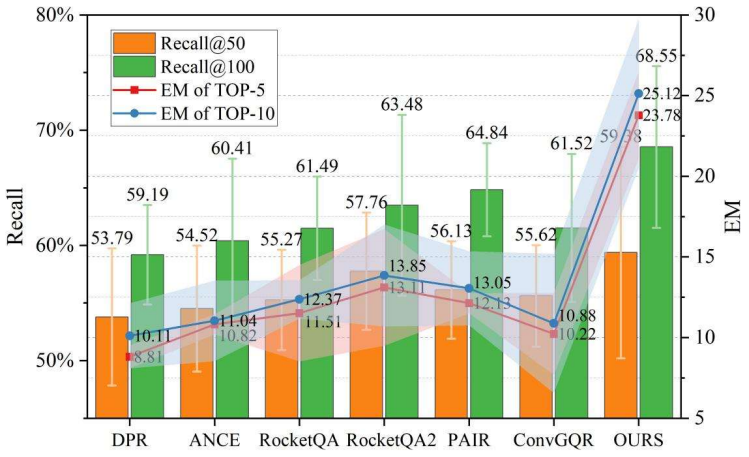

Figure 6: Recall rates and EM metrics of different models

The proposed model reached 59.38% and 68.55% in Recall@50 and Recall@100, respectively, which was significantly better than other baseline models (such as 53.79% and 59.19% of DPR). This shows that by combining

the inference capabilities of BiGRU and the neural state machine NSM, the model can more comprehensively capture the relevant information in the knowledge base. In terms of generation quality, the TOP-5 EM and TOP-10 EM of this paper's model reach 23.78% and 25.12%, respectively, far exceeding the comparison models (e.g., 11.51% and 12.37% for RocketQA). This gap highlights the advantage of the hybrid RAG strategy in generating answer accuracy - by dynamically retrieving real-time knowledge and fusing contexts, the model is able to generate answers that are more in line with users' needs. It is worth noting that the RocketQA series model performs well in terms of recall, with the Recall@100 of RocketQA2 reaching 63.48%, but there is still a significant gap between the model and the model in this paper in terms of EM indicators, which further shows that it is difficult to directly improve the semantic matching degree of the answer by relying only on retrieval optimization, and it needs to be combined with the in-depth optimization of the generation process.

III. B. Complex Knowledge Base Q&A Study

To further validate the superiority of the introduced indexing optimization method, this paper conducts related experiments on the more complex knowledge base Q&A datasets CWQ and WebQSP.

III. B. 1) Data set construction

CWQ: Contains complex multi-hop reasoning questions that need to be answered by combining multi-step logical relationships in the knowledge base.

WebQSP (WebQuestionsSP): extended from WebQuestions, adding more complex semantic parsing requirements, questions need to be reasoned through entities, attributes and relationship chains in the knowledge base.

The complex datasets are constructed to support complex queries and contain a large number of question-answer pairs that require joint reasoning across multiple entities.

III. B. 2) Parameterization and evaluation indicators

For the experimental parameter settings, this paper uses the Adam optimizer to learn the parameters. For both datasets, the embedding dimension dim = {100,150,200,250,300} is adjusted, and the optimal dim settings are 200 and 250 on the CWQ and WebQSP datasets, respectively.The batch size of training batch = {128,256,512,1024}, and the optimal batch on the CWQ and WebQSP datasets is set to 512. The attention weight α is searched in the range of [0,1], and the optimal α is set to 0.7 on the CWQ and WebQSP datasets.

In this section, F1 and Hits@1 metrics are introduced to evaluate the model performance.HITS@1 denotes the percentage of entities correctly predicted by the model ranked in the first position in the test set. Specifically, HITS@1 is the accuracy when the model predicts an entity ranked in the first place.

III. B. 3) Introduction to the baseline model

In order to validate the effectiveness of the method proposed in this paper, the following 11 representative ones are selected as baseline systems for comparison. For the English datasets CWQ and WebQSP, the KBPL model proposed in this paper is compared with two types of baselines: traditional complex knowledge base quizzing methods based on traditional complex knowledge base quizzing methods and complex knowledge base quizzing methods based on pre-trained language models.

Among the traditional complex knowledge base Q&A based methods are ReifKB, UHop, NSM, TUL, ARL and PullNet.

Pre-trained model-based complex knowledge base Q&A applies pre-trained models to complex knowledge base Q&A, which enhances the ability to learn powerful representations from text corpus and improves the semantic parsing and retrieval matching performance of traditional methods, and its representative models are mainly UniKGQA, KD-CoT, Keqing-ChatGPT, KB-BINDER-Codex and BeamQA.

III. B. 4) Analysis of the results of comparative experiments

Table 1 shows the experimental results for the CWQ and Web QSP datasets, which are visualized in Figure 7. The methods in the table are mainly categorized into traditional complex knowledge base Q&A based methods and pre-trained language model based methods.

Table 1: Experimental results of the datasets WebQSP and CWQ

Model	CWQ		WebQSP	
	F1	Hits@1	F1	Hits@1
ReifKB	35.97	41.62	53.48	52.25
UHop	30.22	35.12	68.66	71.68
NSM	42.02	48.20	62.99	68.84
TUL	37.57	40.88	67.84	68.44
ARL	43.79	48.60	72.32	72.24
PullNet	41.64	47.44	65.73	68.23
UniKGQA	51.14	49.03	77.82	71.67
KD-CoT	60.89	55.06	52.04	68.74
Keqing-ChatGPT	64.35	58.96	75.26	71.83
KB-BINDER-Codex	67.23	60.76	74.63	67.78
BeamQA	51.73	50.75	72.79	74.87
OURS	70.34	73.18	83.31	85.33

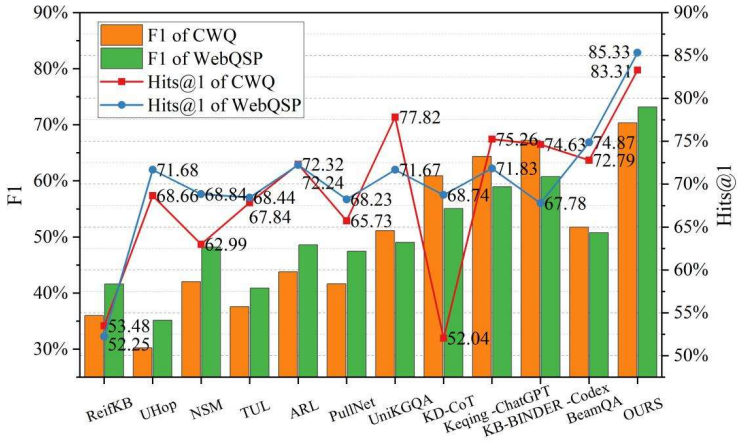

Figure 7: The F1 and Hits@1 metrics of the model on the CWQ and WebQSP datasets

The proposed model reached 70.34% and 73.18% on the F1 and Hits@1 indicators of the CWQ dataset, respectively, and 83.31% and 85.33% on the WebQSP dataset, which is comprehensively ahead of the traditional methods (such as the 42.02% F1 of NSM) and the pre-trained model (such as the 64.35% F1 of Keqing-ChatGPT). This result shows that combining HNSW-PQ index optimization with hybrid RAG strategy, the model can efficiently handle complex semantic associations in multi-hop inference tasks. Among the traditional methods, the ARL model performs better on WebQSP with F1 = 72.32%, but it is still significantly lower than that of the model in this paper, which is 83.31%, suggesting that there is a bottleneck in the dynamic knowledge fusion and retrieval efficiency of the traditional methods. In addition, pre-trained models such as KB-BINDER-Codex (CWQ F1=67.23%) are able to utilize large-scale corpus to enhance semantic representations, but their performance is still limited by the real-time nature of knowledge base retrieval. In contrast, the author's model achieves a better balance in complex reasoning tasks through the dual strategy of indexing optimization and generative enhancement, verifying its universal advantage in cross-domain, multimodal knowledge base querying.

III. B. 5) Effect of hyperparameters on experiments

In order to obtain the optimal values of the model parameters, in this paper, the number of iterations Epoch is set to 2000 and the optimal values of the parameters are determined based on the performance on the validation set. There exist the following 2 parameters that affect the effectiveness of the model: (1) the attention weight and (2) the embedding dimension dim.

Attention weight and embedding dimension are searched in the range of [0,1], {50,100,150,200,250,300} respectively on the validation set to set the optimal values to make the model optimal. The F1 values and Hits@1 of the dataset for different values of attention weights and embedding dimensions are shown in Fig. 8 and Fig. 9, respectively.

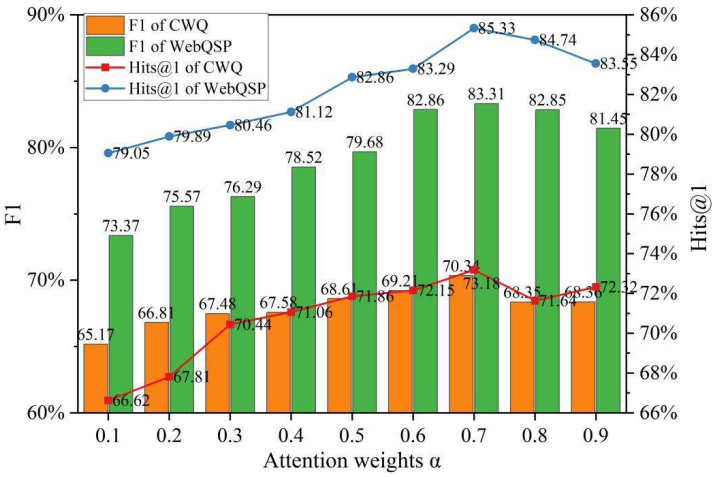

Figure 8: F1 values under different attention weights α and Hits@1

In the CWQ dataset, when the α gradually increased from 0.1 to 0.7, the F1 score increased from 65.17% to 70.34%, and the Hits@1 increased from 66.62% to 73.18%, indicating that the increase of attention weight can effectively enhance the model's ability to capture complex semantic relationships. However, when the α continues to increase to 0.8 and 0.9, both F1 and Hits@1 decrease, and F1=68.35% when α=0.8, indicating that the excessive attention weight may lead the model to pay too much attention to local features and ignore the global semantic association. In the WebQSP dataset, F1 and Hits@1 peaked at 83.31% and 85.33%, respectively, at α=0.7, which was significantly better than the other settings, and at α=0.6, F1 reached 82.86%. This phenomenon verifies the key role of attention mechanism in dynamically adjusting the weight of information fusion—moderate attention allocation can balance the local and global features in multi-hop inference, while extreme weight setting can disrupt this balance and lead to performance degradation.

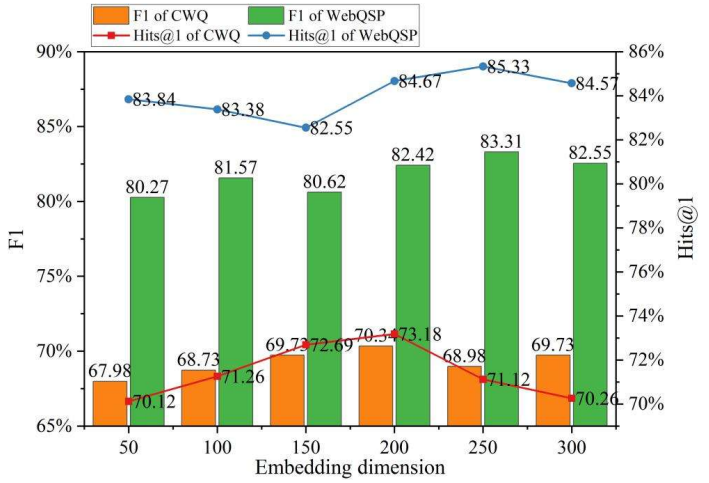

Figure 9: F1 values under different embedding dimension and Hits@1

In the CWQ dataset, when the dim increases from 50 to 200, the F1 score gradually increases from 67.98% to 70.34%, and the Hits@1 increases from 70.12% to 73.18%, indicating that increasing the embedding dimension can more fully represent complex semantic information. However, when the dim is further increased to 250 and 300, F1 drops to 68.98% and 69.73%, respectively, indicating that too high dimensions may introduce redundant noise and reduce the generalization ability of the model. In the WebQSP dataset, F1 and Hits@1 reached their highest values at dim=250 (83.31% and 85.33%), while dim=200 showed slightly lower performance (F1=82.42%)=. This difference indicates that there are differences in the sensitivity of different datasets to embedding dimensions—WebQSP requires a higher-dimensional vector space to distinguish fine-grained semantic relationships because it involves more complex multi-entity inference; However, the inference logic of CWQ is relatively centralized, and the medium dimension (dim=200) can meet the needs. In addition, the F1 of WebQSP drops to 82.55% when dim=300, which further confirms the negative impact of dimensional redundancy on performance.

The above figures collectively reveal the need for balance in the model parameter settings. The optimal combination of attentional weights (α = 0.7) and embedding dimensions (CWQ-dim = 200, WebQSP-dim = 250) verifies the effectiveness of the hybrid RAG strategy proposed in this paper in dynamic parameter adaptation, and provides scalable optimization directions for complex knowledge base quizzing tasks.

III. B. 6) Ablation experiments

While the hyperparametric analysis clarifies the optimal configuration of the model, the actual contribution of each component still needs to be quantified by ablation experiments. In order to verify the effectiveness of the components in the hybrid RAG strategy proposed in this paper, this section designs multiple sets of ablation experiments to evaluate the impact on the model performance by gradually removing or replacing key modules. The specific experimental settings are as follows.

Remove BiGRU module (Ab-BiGRU): the bidirectional gated recurrent unit (BiGRU) is replaced with a unidirectional GRU, and the remaining components (NSM, hybrid RAG, index optimization) remain unchanged.

Disable Neural State Machine (Ab-NSM): remove the probabilistic scene graph inference mechanism of the Neural State Machine, and use only the BERT encoder to generate answers directly without semantic-image joint inference. The retrieval and generation process still uses the hybrid RAG strategy.

Remove hybrid RAG strategy (Ab-RAG): disable the dynamic knowledge retrieval function, rely only on the static data in the initial knowledge base to generate answers, and do not perform contextual information fusion and hint sample introduction. The index optimization method still adopts the HNSW-PQ composite index.

Replacement index optimization method (Ab-Index): the HNSW-PQ composite index is replaced with the traditional BM25 sparse retrieval method, and the rest of the modules (BiGRU, NSM, and hybrid RAG) remain unchanged.

The results of the ablation experiments under each experimental setup are shown in Table 2.

Table 2: Ablation experiment results

	CWQ		WebQSP	
	F1	Hits@1	F1	Hits@1
Ab-BiGRU	34.69	35.61	41.28	44.11
Ab-NSM	40.49	43.02	46.09	49.24
Ab-RAG	47.07	49.51	54.95	60.93
Ab-Index	62.3	63.98	65.67	68.75
Baseline	70.34	73.18	83.31	85.33

In the CWQ dataset, the removal of the BiGRU module (Ab-BiGRU) caused F1 and Hits@1 to drop to 34.69% and 35.61%, respectively, compared with 70.34% and 73.18% of the complete model, and the performance decreased by more than 50%, indicating that the bidirectional feature extraction ability of BiGRU is the core of the model to understand complex semantics. When the neural state machine (Ab-NSM) is disabled, the F1 and Hits@1 are 40.49% and 43.02%, respectively, indicating that although the inference mechanism of NSM can improve the performance, its dependence on the model as a whole is lower than that of BiGRU. After removing the hybrid RAG strategy (Ab-RAG), the F1 and Hits@1 decreased to 47.07% and 49.51%, which further verified the significant improvement effect of dynamic knowledge retrieval and context fusion on the generation quality. After replacing the index optimization method (Ab-Index), F1 and Hits@1 still maintain 62.3% and 63.98%, indicating that although the HNSW-PQ composite index can optimize the retrieval efficiency, it is not the only decisive factor for performance.

In the WebQSP dataset, the trend is consistent with CWQ, but the performance degradation is relatively small. For example, the F1 and Hits@1 of Ab-RAG are 54.95% and 60.93%, respectively, while those of the full model are 83.31% and 85.33%, with a gap of about 30%, indicating that WebQSP relies more on RAG's dynamic knowledge enhancement due to the higher complexity of the problem. Overall, the complete model was significantly better than the ablation group on both datasets, which verified the synergistic necessity of BiGRU, NSM, mixed RAG and HNSW-PQ index optimization.

III. C. Knowledge Base Q&A System Evaluation Results

To validate the knowledge base Q&A system based on deep learning model with retrieval augmented generation (RAG) technique designed in this paper, practical comparison experiments with knowledge graph based Q&A system and large language model based Q&A system are conducted on WebQSP dataset.

III. C. 1) Evaluation indicators

ROUGE is often used as an evaluation metric in question and answer system generation tasks. Among them, ROUGE-1 is the recall between the model output and the reference summary at the 1-tuple word level, where the closer its value is to 1, the higher the agreement between the model output and the reference answer is.The ROUGE-L metric, on the other hand, calculates the recall based on the longest common subsequence between the model output and the reference summary. The value of this metric ranges from 0 to 1, where 1 represents perfect agreement and 0 indicates a complete mismatch, and is suitable for evaluating the fluency and coherence of the text.

III. C. 2) Experimental results

In order to ensure the reliability and statistical significance of the test results, 10 independent experimental tests of the three systems were conducted in this study, and their mean values were taken for comparison. The scores of each index of the system evaluation are shown in Figure 10.

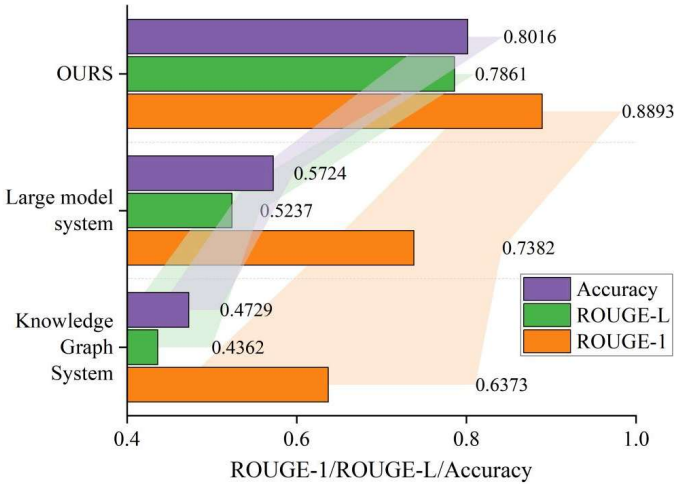

Figure 10: System evaluation results

The system in this paper is significantly ahead in ROUGE-1, ROUGE-L and Accuracy, with ROUGE-1 reaching 0.8893, far exceeding 0.6373 in the knowledge graph system and 0.7382 in the big model system, indicating that the answers generated by the system in this paper are closer to the reference answers in terms of vocabulary coverage, and the ROUGE-L of 0.7861 is improved by more than 50% compared with 0.5237 of the big model system is 0.7861, which is more than 50% higher than the 0.5237 of the large model system, indicating that the model is able to generate more logically coherent answers for long texts through context fusion and dynamic knowledge enhancement by the hybrid RAG strategy. The exact matching rate Accuracy reaches 0.8016, which is significantly higher than other systems, and 0.4729 for the knowledge graph system, verifying the central role of BiGRU and NSM in semantic reasoning and exact matching.

IV. Conclusion

The knowledge base question answering system based on retrieval enhancement generation technology proposed in this paper has made significant breakthroughs in semantic understanding, dynamic retrieval and generation quality by integrating BiGRU, Neural State Machine (NSM) and hybrid RAG index optimization methods. The experimental results show that on the railway dataset, the model comprehensively surpasses the baseline model in terms of recall rate Recall@50=59.38%, Recall@100=68.55%, generation accuracy TOP-5 EM=23.78%, TOP-10 EM=25.12%, and the TOP-5 EM=11.51% of RocketQA, which verifies the effectiveness of BiGRU's bidirectional feature extraction and NSM causal inference.

In the complex knowledge base Q&A datasets CWQ and WebQSP, the F1 scores of the models reached 70.34% and 83.31%, respectively, and the Hits@1 were 73.18% and 85.33%, respectively, which were significantly ahead of traditional methods (such as 42.02% F1 of NSM) and pre-trained models (such as 64.35% F1 of Keqing-ChatGPT).

The results of the ablation experiment show that after removing the hybrid RAG strategy (Ab-RAG), the F1 of CWQ decreases to 47.07%, and the F1 of WebQSP decreases to 54.95%, while the disabling of BiGRU (Ab-BiGRU)

results in a performance decrease of more than 50% (F1=34.69%), highlighting the core role of dynamic knowledge enhancement and bidirectional semantic modeling.

On the WebQSP dataset, the ROUGE-1 of the proposed system, ROUGE-L of 0.7861 and the accuracy of 80.16% are significantly better than the knowledge graph system (ROUGE-1=0.6373) and the large language model system (ROUGE-1=0.7382), which proves its comprehensive advantages in multi-hop inference and complex semantic parsing.

References

[1] Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W. X., & Wen, J. R. (2022). Complex knowledge base question answering: A survey. IEEE Transactions on Knowledge and Data Engineering, 35(11), 11196-11215.

[2] Cui, W., Xiao, Y., & Wang, W. (2016, July). KBQA: An Online Template Based Question Answering System over Freebase. In IJCAI (pp. 4240-4241).

[3] Pavlić, M., Han, Z. D., & Jakupović, A. (2015). Question answering with a conceptual framework for knowledge-based system development "Node of Knowledge". Expert systems with applications, 42(12), 5264-5286.

[4] Kafle, S., de Silva, N., & Dou, D. (2020). An overview of utilizing knowledge bases in neural networks for question answering. Information Systems Frontiers, 22, 1095-1111.

[5] Jiang, F., Qin, C., Yao, K., Fang, C., Zhuang, F., Zhu, H., & Xiong, H. (2024, July). Enhancing question answering for enterprise knowledge bases using large language models. In International Conference on Database Systems for Advanced Applications (pp. 273-290). Singapore: Springer Nature Singapore.

[6] Yu, Z., Ouyang, X., Shao, Z., Wang, M., & Yu, J. (2025). Prophet: Prompting large language models with complementary answer heuristics for knowledge-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[7] Cao, J., & Cao, J. (2024). CPEQA: A Large Language Model Based Knowledge Base Retrieval System for Chinese Confidentiality Knowledge Question Answering. Electronics, 13(21), 4195.

[8] Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? on the calibration of language models for question answering. Transactions of the Association for Computational Linguistics, 9, 962-977.

[9] Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. Transactions of the Association for Computational Linguistics, 11, 1-17.

[10] Xu, Z., Cruz, M. J., Guevara, M., Wang, T., Deshpande, M., Wang, X., & Li, Z. (2024, July). Retrieval-augmented generation with knowledge graphs for customer service question answering. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2905-2909).

[11] Xu, L., Lu, L., Liu, M., Song, C., & Wu, L. (2024). Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. Heritage Science, 12(1), 118.

[12] He, X., Tian, Y., Sun, Y., Chawla, N., Laurent, T., LeCun, Y., ... & Hooi, B. (2024). G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. Advances in Neural Information Processing Systems, 37, 132876-132907.

[13] Muludi, K., Fitria, K. M., & Triloka, J. (2024). Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model. International Journal of Advanced Computer Science & Applications, 15(3).

[14] Li, Y. (2025). A Dynamic Knowledge Base Updating Mechanism-Based Retrieval-Augmented Generation Framework for Intelligent Question-and-Answer Systems. Journal of Computer and Communications, 13(1), 41-58.