2025 10th International Conference on Intelligent Computing and Signal Processing (ICSP) | 979-8-3315-3626-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICSP65755.2025.11086677

The Research on Re-ranking Algorithm for FAQ-based Systems in the Petroleum Domain

1st Jiaxiang Zhang
School of Computer Science
Xi'an Shiyou University
Xi'an, Shaanxi, China
941324625@qq.com

2nd Jiaxin Han*
School of Computer Science
Xi'an Shiyou University
Xi'an, Shaanxi, China
*Corresponding author:jxhan@xsyu.edu.cn

***Abstract*—In an FAQ-based system, precise information retrieval and recall are key to enhancing user experience. However, traditional retrieval methods struggle with professional terminology and complex semantic matching, making it difficult to achieve high-precision QA. This study explores the application of reranking algorithms in an oil industry FAQ system, aiming to optimize the ranking of recalled candidate answers and improve QA matching accuracy. We employ a deep learning-based dense retrieval (DPR) model for initial recall and integrate a pretrained reranking model (BAAI/bge-reranker) to refine the retrieved results. Experimental results demonstrate that incorporating a reranking algorithm significantly improves recall precision. Additionally, by fine-tuning the model with domain-specific petroleum data, the reranking task achieves better performance within the field, further validating the value of reranking algorithms in enhancing professional knowledge QA matching.**

***Index Terms*—Petroleum Domain FAQ-based Systems, Retrieval and Recall, Re-ranking Model**

I. Introduction

In the petroleum industry, the rapid advancement of technology and the continuous expansion of production scale have led to an exponential increase in knowledge complexity and specialization. This makes efficiently retrieving accurate information and knowledge from massive datasets a critical task. Traditional information retrieval methods, such as manual data queries and search engines, often struggle to meet highly specialized and domain-specific question-answering needs due to long retrieval cycles, insufficient semantic understanding, and low precision. To address these challenges, FAQ-based systems have emerged, with the core objective of leveraging retrieval and recall techniques to quickly match user queries with the most relevant answers from a pre-constructed domain-specific knowledge base. However, building a question-answering system in the petroleum domain not only requires handling vast and complex industry knowledge but also accommodating specialized terminology, industry standards, and the diverse and intricate nature of contextual semantics. Therefore, improving the accuracy of question-answering systems, particularly in highly specialized petroleum-related scenarios, has become a key research focus in intelligent question-answering systems.

In traditional FAQ-based retrieval systems, conventional retrieval algorithms usually provide multiple candidate answers related to user questions. However, these candidate answers may not necessarily be the most accurate or relevant. To address this issue, this study focuses on introducing a re-ranking algorithm model to optimize the retrieval process and further enhance user experience and the effectiveness of the question-answering system. The primary objective of this study is to explore the application of re-ranking algorithms in the petroleum field FAQ-based systems. Based on the textual knowledge within the petroleum vertical domain, we aim to investigate the optimization and improvement of recall accuracy and question-answer context matching brought about by re-ranking algorithms.

II. Related Work

In the research of vertical domain FAQ-based Systems, the main challenges include understanding professional terminology, data scarcity, retrieval recall optimization, and reranking strategies. Recent trends in research indicate that combining multi-stage retrieval, semantic matching, reranking optimization, contrastive learning, and LoRA fine-tuning techniques can effectively improve the system's retrieval accuracy and domain adaptability. To further optimize the performance of the oil industry FAQ question-answering system, this section will discuss the following key aspects: (1) Characteristics and challenges of the oil industry FAQ question-answering system; (2) Multi-stage retrieval strategies and the application of the DPR model; (3) Reranking techniques; (4) The role of contrastive learning and LoRA fine-tuning with domain knowledge in pre-trained models.

Compared to general question-answering systems, vertical domain FAQ-based systems typically have a more specific professional background and knowledge system. Since these systems need to understand domain-specific terminology and high-quality data is relatively scarce, their generalization ability is often limited. Moreover, the knowledge structure of vertical domains is complex. For example, it needs to integrate knowledge from multiple aspects such as geological exploration, drilling technology, and reservoir management to provide precise question-answering capabilities. For instance, Hojageldiyev et al. proposed the HSE AI Assistant at the Abu Dhabi International Petroleum Exhibition. This system is specifically designed for answering questions related to

occupational health, safety, and environmental protection regulations in the oil and gas industry. It uses the latest natural language processing techniques to retrieve information from the regulatory database and provide the most relevant answers [1].

Traditional FAQ-based systems usually rely on sparse retrieval methods based on keywords, such as TF-IDF, BM25, and inverted index. However, these methods are inadequate when it comes to semantic understanding and long-text processing. In recent years, the DPR model based on deep learning has become the mainstream method in FAQ-based systems due to its excellent semantic understanding capabilities. Compared with traditional retrieval methods, DPR can capture the deep semantic relationships between queries and candidate answers, optimize the initially retrieved answers, and improve the overall accuracy of the system [2][3].

Rodrigo Nogueira et al. published a multi-stage document ranking method based on the BERT model, aiming to improve ranking accuracy in information retrieval systems[4]. This method divides the document ranking task into multiple stages. First, it uses traditional sparse retrieval methods (BM25) to preliminarily filter candidate documents. Then, it employs the BERT model to deeply rank the filtered documents, effectively combining the efficiency of sparse retrieval methods and the semantic understanding capabilities of dense retrieval models. Through this multi-stage strategy, ranking accuracy can be significantly improved, especially in large-scale document retrieval tasks, reducing computational costs and enhancing the system's efficiency and accuracy.

Deep neural embedding models capture the semantic features of text by converting it into low-dimensional vector representations and then computing the similarity between texts based on these vectors. The core of neural retrieval lies in accurately calculating the semantic similarity between a query and candidate answers. Among many embedding-based retrieval methods, dense retrieval is the most common. This method computes the similarity between texts by aggregating the output of text encoders, thereby recalling the most relevant answers.In conversational question-answering tasks, a dual-tower structure is commonly used to calculate the similarity between user queries and answers in the database, effectively recalling the most matching answers[5]. The input question and query text are mapped to multidimensional vectors using different encoders EQ(·). The relevance between them is measured by calculating the dot product (normalized) or cosine similarity between the question and query paragraph vectors. Finally, the top k paragraphs with the highest similarity are sorted and recalled.

$$\text{sim}(q, p) = \mathbf{E}_Q(q)^\top \cdot \mathbf{E}_P(p) \tag{1}$$

The research progress in re-ranking techniques has also provided significant support for improving the performance of question-answering systems. Re-ranking technology improves the accuracy of the final output by further ranking the candidate answers initially retrieved. This technique typically relies on deep learning models, especially those based on the Transformer architecture (BERT, GPT, etc.), which are capable of feature learning on large-scale data, thereby enhancing the accuracy and effectiveness of the ranking[6]. For domain-specific question-answering systems, such as those in the petroleum industry, the task of re-ranking models is to optimize the ranking of retrieval results through more refined feature learning, making it better aligned with the user's actual needs. Question-answering systems in the petroleum field often face challenges with specific terminology and complex concepts, making the design of customized re-ranking models particularly important. The integration of retrieval-based recall and re-ranking techniques has become a key direction for improving the overall performance of private domain question-answering systems.

To enhance the accuracy and professionalism of information retrieval models in specific knowledge domains, contrastive learning and model fine-tuning methods have been widely applied in the field of natural language processing in recent years. Contrastive learning, especially in vector representation learning for information retrieval and question-answering systems, aims to optimize the embedding representations by maximizing the similarity between representations of similar samples while minimizing the similarity between dissimilar samples[7]. For the DPR retrieval model, contrastive learning can effectively enhance the model's semantic modeling capabilities in unsupervised or weakly supervised environments. By constructing positive and negative sample pairs, the distribution of the embedding space can be optimized to be more discriminative, thereby improving the model's retrieval performance.

Meanwhile, LoRA fine-tuning technology provides a new paradigm for the efficient tuning of large-scale language models. It reduces the parameter overhead during the fine-tuning process by adding low-rank adaptation matrices to the weight matrices in the Transformer architecture, enabling large models to efficiently adapt to new tasks under limited resource conditions[8]. In FAQ-based systems, LoRA can be used to quickly adapt the model to domain-specific data, avoiding the computational and storage overhead associated with full fine-tuning.

III. Methodology

A. Experiments Design

The selection and application of retrieval and re-ranking pre-trained models are a crucial part of the experimental design in this study. We mainly used the BAAI/bge series of pre-trained models for the relevant experiments. Based on the design requirements of the petroleum field FAQ-based systems and the research focus of the re-ranking model, we selected multiple pre-trained models for comparative experiments and chose the better-performing models for fine-tuning experiments. We set different hyperparameters for training to verify the improvement of the re-ranking results.

In the initial retrieval recall, this experiment uses the BAAI/bge-m3 dense retrieval embedding model, which is

based on the XLM-RoBERTa architecture. As an important optimization variant of the BERT single encoder structure, XLM-RoBERTa adopts a larger-scale cross-lingual unsupervised pretraining, removes the next-sentence prediction task from BERT, and focuses on dynamic masked language modeling. It also incorporates longer training steps and larger batch sizes. These optimizations significantly enhance the performance of embedding models, providing a solid foundation for the retrieval capabilities of m3 [9]. Additionally, bge-m3 further utilizes a large training corpus for pretraining and fine-tuning, enabling better dense passage retrieval (DPR) capabilities. This includes a substantial amount of domain-specific data (such as S2ORC, PubMedQA, etc.), which allows it to better capture the semantics of professional terminology. Moreover, the m3 model supports inputs up to 8192 tokens, which effectively handles long documents and technical reports in the oil industry. In contrast to traditional BERT or Transformer models, which are typically limited to token lengths of 512 or 1024, the professional documents in the oil industry are often lengthy. The m3 model's long-text processing ability allows it to better capture the overall semantics of documents [10].

To enhance computational efficiency and meet question-answering demands, this experiment employs pooling techniques to reduce the dimensionality of feature vectors, improving the recall efficiency of question-answer computations. The output vectors uniformly use the normalized [CLS] token as the semantic representation (embedding_dims=1024). Both the test set questions and the entire candidate answer corpus are encoded into vectors [10]. In this study, 500 test set questions and 43,330 corresponding candidate answers are individually converted into feature vectors. A dual-tower architecture is then used to compute the dot product between the normalized feature vectors of the questions and all candidate answer texts, generating similarity scores. These scores are sorted to obtain the top 5 results, and relevance is determined based on whether the corresponding real answers match the test questions.

After the initial recall, the BAAI/bge-reranker series models are used to rerank the candidate answers to further improve retrieval accuracy. Specifically, this series of models employs a pretrained Transformer architecture to jointly encode the question and candidate answers. The input format is constructed as "CLS + prompt + SEP + question + SEP + answer + SEP." This cross-encoder structure allows the model to capture both the question and answer's semantic information within a single sequence, and the CLS token is used to capture the global semantics, enabling more precise calculation of the relevance between the two.

The model first inputs the user's question and candidate answers into an interactive Transformer, generating context-rich feature representations for fine-grained semantic matching between the question and answer. Then, a deep neural network is used to match the question-answer pair, and a Sigmoid activation function maps the output relevance scores to the [0, 1] range. The higher the score, the stronger the semantic match between the question and answer. Finally, the similarity scores are used to rank the multiple initially recalled candidate answers, ensuring that the most relevant answers are placed at the top [11]. The process is illustrated in Figure 1.

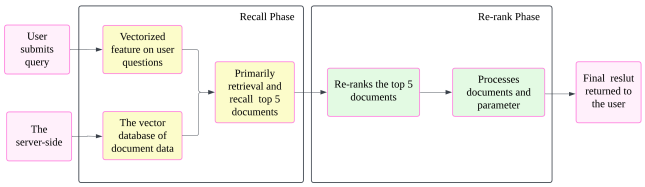

Fig. 1. Validation Process Flowchart of Re-ranking Model.

This paper will test the improvement effects of different BGE-reranker models on the recall results and select the optimal pre-trained model base for fine-tuning based on the experimental results. Through comparative experiments, we systematically evaluate the advantages and disadvantages of different models and further optimize the performance of the petroleum field FAQ-based systems. The specific model selection is shown in Table 1.

TABLE I
PRETRAINED MODEL INFORMATION

Model Category	Pretrained Model Information	
	Pretrained Model	**Size (GB)**
Encoder-only	-	-
Retrieval	bge-m3	2.3
Reranker	bge-reranker-base	1.1
Reranker	bge-reranker-large	2.2
Reranker	bge-reranker-v2-m3	2.3
Decoder-only	-	-
Reranker	bge-reranker-v2-gemma	10.1
Reranker	bge-reranker-v2.5-gemma2-lightweight	37.4
Reranker	bge-reranker-v2-minicpm-layerwise	10.9

B. Dataset Preparation

The petroleum domain question-answering dataset used in this study is derived from authoritative Chinese resources in the petroleum industry, including petroleum knowledge, journals of petroleum science, and publicly available content from official websites. The dataset comprises 1,720 articles from January 2020 to December 2024, covering multiple important topics in the petroleum field, such as exploration and development, refining and chemical processing, oil and gas storage and transportation, and petrochemical enterprises. Through these articles, a multi-topic rich dataset has been constructed, providing a solid foundation for the training and fine-tuning of the question-answering system.

The generation of the dataset combines automated techniques with manual review. Relying on a pre-trained large-scale language model (Qwen2.5-72B-Instruction), raw question-answer pairs were generated and then manually annotated and cleaned to ensure data quality. A total of 43,330 question-answer pairs were collected, with redundant and illegal content removed. The test set randomly selected 500 data points to validate the recall and re-ranking model performance.

This dataset not only covers common petroleum knowledge questions but also focuses on domain-specific terminology, industry standards, and complex concepts, effectively supporting the precise ranking and fine-tuning optimization of the re-ranking model. An example of the data is shown in Table 2 below:

TABLE II
PROFESSIONAL Q&A IN THE PETROLEUM FIELD.

Question	Answer
What are the existing problems with geothermal energy utilization?	There are sustainability issues with geothermal energy utilization, including insufficient understanding of comprehensive resource utilization and industrial development, low-level and low-benefit development, serious resource wastage, and unclear understanding of geothermal resource characteristics by developers. Additionally, environmental issues such as water pollution, thermal pollution, air pollution, soil pollution, and ground subsidence may arise during the development and utilization of geothermal energy.
What are the two basic methods of oil extraction?	The two basic methods of oil extraction include: one is relying on natural or artificial water injection, gas supplementation to produce energy through self-pressurization; the other is using mechanical lifting, that is, installing oil extraction machinery in oil wells that cannot self-pressurize to produce oil. Currently, about 90% of oil wells worldwide use mechanical lifting methods.
What is mechanical drilling speed and the factors affecting mechanical drilling speed?	Mechanical drilling speed refers to the length (meters per hour) of continuous drilling progress obtained per hour, which is an important indicator for measuring drilling efficiency. There are many factors affecting mechanical drilling speed, including: • Rock properties (drillability, abrasivity) • Process parameters (drilling pressure, rotation speed) • Hydraulic parameters (pumping power, pump pressure, flow rate) • Drilling fluid properties (density, viscosity, solid content, etc.) • Bit condition (structure type, wear degree)

IV. EXPERIMENTS AND RESULTS

A. Experiments Setup

This experiment was conducted using the Google Colab cloud service for setting up configurations and the experimental setup. GPU resources were rented, and a cloud drive was mounted for data storage and access. The key hardware and software configurations are summarized in Table 3.

B. Evaluation Metrics

In this study, several evaluation metrics are utilized to assess the performance of the retrieval and re-ranking models for the FAQ-based systems in the oil domain. These metrics provide a clear understanding of how well the models perform in retrieving and ranking relevant answers.For ranking tasks, evaluation metrics typically focus on recall hit rate (Hit Ratio), Mean Reciprocal Rank (MRR), and other ranking-specific measures, rather than commonly used metrics such as F1 score and accuracy, which are more prevalent in classification tasks.

TABLE III
KEY HARDWARE AND SOFTWARE CONFIGURATIONS

Category	Configuration Details
Hardware	
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz (6 cores)
GPU	NVIDIA L4 (22.5GB VRAM)
Memory	53 GB
Storage	236 GB
Software	
Operating System	Ubuntu 22.04.3 LTS
Programming Language	Python 3.10.12
DeepLearning Framework	Torch==2.5.1
CUDA Version	12.2.1
Additional Libraries	FlagEmbedding==1.3.3, Faiss-gpu==1.7.3, Peft==0.14.0, Transformers==4.47.1

1) Recall Hit Rate: Recall hit rate is an important metric for evaluating the ability of a retrieval model to identify relevant answers. It measures the proportion of relevant answers that appear in the top N retrieved results. The recall hit rate for the top N positions is calculated as follows in formula 2:

$$\text{Recall Hit Rate@N} = \frac{\text{Number of relevant answers in top N results}}{\text{Total number of relevant answers}} \tag{2}$$

2) Mean Reciprocal Rank (MRR): The MRR is another critical metric that evaluates the ranking quality of the model. It calculates the average of the reciprocal ranks of the first correct answer across all queries. It is defined as:

$$\text{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\text{rank}_i} \tag{3}$$

C. Experimental Results

1) Preliminary Recall & Re-ranking Results: In this section, we present the retrieval recall and reranking results of different models evaluated using FP16 half-precision floating-point numbers. The model performance is measured using Hit@1-5 , MRR and GPU utilization. These metrics effectively assess the model's ability to retrieve relevant information and optimize rankings.

The test dataset consists of 500 randomly selected samples from previously self-collected question-answer pairs in the oil industry, covering various question-answer scenarios to ensure diversity and representativeness in the testing. In the experiment, we first use the recall model to retrieve candidate document sets for these 500 queries, then apply the reranking model for fine-grained ranking to evaluate the performance improvements of different pretrained reranking models.

Table 4 summarizes the experimental results, showing the performance of different pretrained reranking models based on the key evaluation metrics mentioned above.

According to the experimental results, compared to the baseline retrieval (82.6%), the selected pretrained reranker models improved the recall precision by approximately 3%-6%, with a particularly significant increase in the Top-1 hit rate, confirming the effectiveness of the reranker models.

TABLE IV
MODEL PERFORMANCE COMPARISON

Model	Top1_HIT	Top2_HIT	Top3_HIT	Top4_HIT	Top5_HIT	MRR	GPU utilization
Retrieval Result	82.6%	88.6%	91.2%	93.6%	95.2%	0.873867	-
bge-reranker-base	85.2%	90.6%	93.2%	94.0%	95.2%	0.892067	0.9GB
bge-reranker-large	86.4%	91.0%	93.6%	94.6%	95.2%	0.899367	1.4GB
bge-reranker-v2-m3	**87.2%**	**92.2%**	**94.0%**	**94.8%**	95.2%	**0.905800**	1.6GB
bge-reranker-v2-gemma	88.2%	92.0%	93.4%	94.4%	95.2%	0.909767	15.3GB
bge-reranker-v2.5-gemma2-lightweight	88.0%	**93.2%**	**94.4%**	**95.0%**	95.2%	0.911900	20.3GB
bge-reranker-v2-minicpm-layerwise[10]	85.8%	91.2%	93.8%	94.8%	95.2%	0.896967	5.8GB
bge-reranker-v2-minicpm-layerwise[28]	88.4%	92.6%	94.0%	94.8%	95.2%	0.912467	6.3GB
bge-reranker-v2-minicpm-layerwise[40]	**88.4%**	92.8%	94.2%	94.8%	95.2%	**0.912967**	6.6GB

BAAI/bge-reranker-v2-m3 improved the Top-1 hit rate by **4.6%**, and the MRR increased from 0.8739 to 0.9058.BAAI/bge-reranker-v2-minicpm-layerwise [40] showed excellent performance, with the Top-1 hit rate increasing by **5.8%**, and the MRR improving from 0.8739 to 0.91297.

While BAAI/bge-reranker-v2.5-gemma2-lightweight outperformed the minicpm model in terms of hit rates for Top-2 to Top-4 by 0.4%-0.2%, it has a model size and memory usage that are 2-3 times higher.For the bge-reranker-v2-minicpm-layerwise model, the reranking effect was similar when the decoder layer count was truncated at the 28th and 40th layers, with average processing times of 0.13 seconds and 0.19 seconds, respectively.

2) Model Fine-tuning & Re-ranking Results: Based on the aforementioned experimental results, this study utilizes a self-collected question-and-answer knowledge dataset in the field of petroleum, employing methods such as contrastive learning and low-rank adaptation to fine-tune two pre-trained models BAAI/bge-reranker-v2-m3 and BAAI/bge-reranker-v2-minicpm-layerwise. The number of training epochs is adjusted according to the main parameter to conduct tests, in order to verify the enhancement effect of fine-tuning on the re-ranking results.

The fine-tuning process utilized QA pairs to construct the training dataset, formatted in JSONL. Each question was saved as a dictionary containing three keys: query, pos, and neg, and a prompt field was added for decoder-only models. The total volume of training data was **43,330** entries samples. The query represents the question, pos is the corresponding answer document, and neg consists of multiple irrelevant negative example documents.

To generate the negative examples, a hard negative mining method was employed. Initially, 100 soft negative examples were randomly selected from other answers. Then, a smaller model (BAAI/bge-large-zh-v1.5) was used to compute the similarity and rank them, from which the top 15 hard negative examples were chosen for the training data. The final fine-tuning dataset size was 260MB.The results are shown in Table 5 below:

Based on the training methods and scripts from the official website, and considering the characteristics of text length and other factors in the petroleum domain knowledge data, the key training hyperparameters for fine-tuning BAAI/bge-reranker-v2-minicpm-layerwise are set as follows:

```
torchrun --nproc_per_node 1 \
    --model_name_or_path
    BAAI/bge-reranker-v2-minicpm-layerwise \
    -lora_rank 32 \
    -lora_alpha 64 \
    --use_flash_attn True \
    --target_modules
    q_proj k_proj v_proj o_proj \
    --save_merged_lora_model True \
    --model_type decoder \
    --start_layer 8 \
    --head_multi True \
    --trust_remote_code True \
    --cache_path ./cache/data \
    --train_group_size 8 \
    --query_max_len 256 \
    --passage_max_len 1024 \
    --pad_to_multiple_of 8 \
    --knowledge_distillation False \
    --output_dir
    ./test_decoder_bge-reranker \
    --learning_rate 2e-4 \
    --bf16 \
    --num_train_epochs 20 \
    --per_device_train_batch_size 10 \
    --gradient_accumulation_steps 1 \
    --warmup_ratio 0.1 \
    --weight_decay 0.01 \
    --logging_steps 1 \
    --save_steps 4333
```

Based on petroleum domain knowledge data and industry content, the fine-tuned models include two types of superior pre-trained models, namely encoder-only and decoder-only, for validation experiments. A comparison of the model performance was conducted, and the experimental results are shown in Table 5:

The experimental results show that the re-ranking model, combined with petroleum domain expertise, performs better in question-answer matching tasks. Specifically, the fine-tuned models exhibit significant improvements in HITRatio and

TABLE V
FINE-TUNED MODEL EXPERIMENTAL RESULTS

Model	Epochs	Top1_HIT	Top2_HIT	Top3_HIT	Top4_HIT	Top5_HIT	MRR
bge-reranker-v2-m3	–	87.2%	92.2%	94.0%	94.8%	95.2%	0.905800
bge-reranker-v2-m3-finetuned	1	88.4%	93.4%	94.4%	95.2%	95.2%	0.914333
bge-reranker-v2-m3-finetuned	2	87.4%	93.2%	94.8%	95.2%	95.2%	0.909333
bge-reranker-v2-m3-finetuned	3	87.6%	94.0%	94.4%	94.8%	95.2%	0.911133
bge-reranker-v2-m3-finetuned	5	88.8%	93.2%	94.8%	95.0%	95.2%	0.916233
bge-reranker-v2-m3-finetuned	10	90.6%	94.4%	95.2%	95.2%	95.2%	0.927667
bge-reranker-v2-m3-finetuned	15	91.2%	94.0%	95.2%	95.2%	95.2%	0.930000
bge-reranker-v2-m3-finetuned	20	**91.4%**	**94.8%**	**95.2%**	**95.2%**	**95.2%**	**0.932333**
bge-reranker-v2-minicpm-layerwise[40]	–	88.4%	92.6%	94.0%	94.8%	95.2%	0.912467
bge-reranker-v2-minicpm-layerwise[40]-finetuned	1	92.2%	94.6%	95.0%	95.2%	95.2%	0.935833
bge-reranker-v2-minicpm-layerwise[40]-finetuned	2	92.6%	94.8%	95.0%	95.2%	95.2%	0.938167
bge-reranker-v2-minicpm-layerwise[40]-finetuned	3	92.6%	95.0%	95.2%	95.2%	95.2%	0.938667
bge-reranker-v2-minicpm-layerwise[40]-finetuned	5	92.4%	94.2%	95.0%	95.2%	95.2%	0.936167
bge-reranker-v2-minicpm-layerwise[40]-finetuned	10	92.8%	94.8%	95.2%	95.2%	95.2%	0.939333
bge-reranker-v2-minicpm-layerwise[40]-finetuned	15	**92.8%**	**95.2%**	**95.2%**	**95.2%**	**95.2%**	**0.940000**
bge-reranker-v2-minicpm-layerwise[40]-finetuned	20	92.8%	95.2%	95.2%	95.2%	95.2%	0.940000

MRR metrics for Top-1 to Top-5, compared to the base pre-trained models. These improvements are especially notable in the Top-1 hit rate. Specifically, the m3 model performed optimally at epoch=20, with a **4.2%** improvement, while the minicpm-layerwise model reached its best performance at epoch=15, with a **4.4%** increase. These results demonstrate that fine-tuning on a vertical domain-specific dataset can effectively improve the accuracy of question-answer systems in the domain, proving the positive impact of domain-specific fine-tuning on model performance.

V. DISCUSSION AND CONCLUSION

This study verifies the effectiveness of deep learning-based re-ranking models in question-answering systems within the oil field domain. The experimental results indicate that under a multi-stage retrieval framework, employing re-ranking strategies based on pre-trained models such as the BGE series significantly enhances the accuracy of retrieval results. Improvements in MRR (Mean Reciprocal Rank) and Top-k recall rates further demonstrate their generalization capability on specialized domain data, particularly in question-answering tasks involving lengthy texts and dense professional terminology. The study found that deep learning re-ranking models can more effectively enhance the accuracy of answers after preliminary recall compared to traditional retrieval methods. Additionally, using domain-adapted text embedding models combined with fine-tuning training can further enhance their retrieval performance in vertical domains like oil. Moreover, the advantages of the multi-stage retrieval architecture were also validated. By integrating semantic understanding capabilities of neural networks like DPR and adopting a ”primary retrieval and recall + re-ranking” strategy, it is possible to improve the system’s accuracy while ensuring practicality. These results suggest that for specialized question-answering tasks, the key to enhancing the performance of question-answering systems lies in the rational selection of pre-trained models, conducting domain-adaptive fine-tuning, and employing multi-stage retrieval strategies.

REFERENCES

[1] D. Hojageldiyev, “Artificial intelligence in HSE,” in *Proceedings of the SPE Abu Dhabi International Petroleum Exhibition & Conference*, Abu Dhabi, UAE, Nov. 12-15, 2018, Paper SPE-192820-MS.

[2] B. Mitra and N. Craswell, “An introduction to neural information retrieval,” *Foundations and Trends® in Information Retrieval*, vol. 13, no. 1, pp. 1-126, 2018.

[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, “Dense passage retrieval for open-domain question answering,” arXiv preprint arXiv:2004.04906, 2020.

[4] Nogueira R, Yang W, Cho K, et al. Multi-stage document ranking with BERT. arXiv preprint arXiv:1910.14424, 2019.

[5] H. Tao, J. Zeng, Y. Yu, Z. Wang, and X. Hu, “SynC: A Dense Retrieval Method based on Syntactical Contrastive Learning,” in 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, pp. 1–8, 2023.

[6] A. Yates, R. Nogueira and J. Lin, ”Pretrained transformers for text ranking: BERT and beyond”, Proc. 14th ACM Int. Conf. Web Search Data Mining, pp. 1154-1156, Mar. 2021.

[7] T. Gao, X. Yao, and D. Chen, “SimCSE: Simple contrastive learning of sentence embeddings,” *arXiv preprint arXiv:2104.08821*, 2021.

[8] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.

[9] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[10] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation,” *arXiv preprint arXiv:2402.03216*, 2024.

[11] X. Ma, X. Zhang, R. Pradeep, and J. Lin, “Zero-shot listwise document reranking with a large language model,” *arXiv preprint arXiv:2305.02156*, 2023.