10.37965/jdmd.2026.1505

A Fault Diagnosis-Oriented Data-Driven Embedding Fine-Tuning and Domain-Aware Hybrid Retrieval Method for Intelligent Operation and Maintenance

Xueyi Li[1,2], Gang Li[1], Jiannan Dong[1,*], Yun Kong[3,**], Wenyang Hu[2], Tianyang Wang[2]

[1] College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin 150040, China

[2] Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China

[3] School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China

*Corresponding author: Jiannan Dong (e-mail: dl0827@nefu.edu.cn)

**Corresponding author: Yun Kong (e-mail: kongyun@bit.edu.cn)

Received 11 June 2026; Revised 07 July 2026; Accepted 28 July 2026; Published online 01 August 2026

Abstract: To improve the support of engineering knowledge resources for fault diagnosis and intelligent maintenance, this paper proposes a fault diagnosis-oriented data-driven embedding fine-tuning and domain-aware hybrid retrieval method. The proposed framework is intended for industrial fault diagnosis and intelligent operation and maintenance, and is instantiated and validated in wind turbine operation and maintenance. Domain-specific training data are constructed to fine-tune the embedding model, enhancing its representation of fault concepts, failure mechanisms, and maintenance knowledge. Query expansion, dense retrieval, sparse retrieval, and domain term matching are then integrated with weighted fusion and cross-encoder re-ranking to obtain relevant and low-redundancy diagnostic contexts. Compared with the general retrieval-augmented generation (RAG) baseline, the proposed method increases Recall@5 by 57.33 percentage points, achieves an MRR of 98%, increases the key point coverage rate by 4.8 percentage points, reaches a factual support rate of 97.83%, and reduces the hallucination rate to 2.17%. These results demonstrate its effectiveness in improving retrieval accuracy and generation reliability for fault diagnosis and maintenance decision-making in the wind turbine operation and maintenance scenario.

Keywords*:* fault diagnosis; intelligent maintenance; hybrid retrieval; domain knowledge; fine-tuning

I. INTRODUCTION

With the continuous development of industrial systems toward greater intelligence and efficiency, higher requirements have been placed on data processing capability, fault analysis efficiency, and intelligent decision-making in production operation, equipment maintenance, and technical services [1]. As an important enabling technology for industrial upgrading, artificial intelligence can optimize operation monitoring, fault diagnosis, and maintenance management processes in industrial systems through automated processing, intelligent

decision-making, and data-driven analysis, thereby improving system operating efficiency and reducing maintenance costs [2]. In complex industrial environments, equipment conditions are often affected by variations in operating conditions, environmental disturbances, multi-source data coupling, and uncertainties in operational states [3]. Therefore, traditional analysis methods that rely on manual experience and static rules can hardly meet practical engineering requirements in a timely and accurate manner [4]. In contrast, artificial intelligence can rapidly analyze large amounts of operational data and domain knowledge, thereby providing effective support for fault identification, risk warning, and operation and maintenance decision-making [5].

In recent years, large language models, with their strong capabilities in language understanding, knowledge reasoning, and text generation, have emerged as a promising technical approach for industrial knowledge services and intelligent operation and maintenance [6]. However, the training data of general-purpose large language models are mainly derived from open-domain corpora, and these models usually lack deep modeling of domain-specific industrial knowledge, technical terminology, and engineering scenarios [7]. When applied to highly specialized engineering problems, they are prone to issues such as biased interpretation of technical terms, inaccurate responses, insufficient factual support, and generated results that are disconnected from actual operating conditions [8]. This is especially important in industrial operation and maintenance, where equipment, faults, safety procedures, and maintenance workflows are highly specialized and context-dependent [9]. Therefore, integrating industrial domain knowledge into large language models has become a key challenge for accurate, reliable, and traceable industrial intelligence [10].

At present, methods for adapting large language models to specific domains mainly include model fine-tuning [11], prompt engineering [12] and retrieval-augmented generation [13]. Among these methods, model fine-tuning can enhance domain knowledge acquisition but requires substantial high-quality data and computational resources. Prompt engineering is flexible yet highly dependent on manual expertise and prompt quality. In contrast, retrieval-augmented generation introduces external domain knowledge through knowledge bases, improving response accuracy, interpretability, and traceability without extensive model retraining [14]. Its plug-and-play nature makes it particularly suitable for intelligent question answering and decision-support tasks in knowledge-intensive industrial scenarios [15].

The key to retrieval-augmented generation is to effectively match user queries with relevant documents in an external knowledge base and provide reliable context for the generative model. However, conventional RAG methods mostly rely on general-purpose embedding models and single vector-similarity-based retrieval, making it difficult to fully represent technical terms in industrial domains. As a result, the retrieved results may deviate from the user's true intent.

To address these issues, this study proposes a fault diagnosis-oriented RAG method that combines domain-specific embedding fine-tuning with domain-aware hybrid retrieval. Wind turbine operation and maintenance is used for validation. Domain knowledge is used to fine-tune the embedding model, while query expansion, dense and sparse retrieval, term matching, weighted fusion, and cross-encoder re-ranking are integrated to improve retrieval coverage, relevance, and domain consistency, providing more accurate contextual support

for generation. The main contributions of this study are summarized as follows:

(1) This study proposes a domain-specific embedding fine-tuning method to improve the representation of wind turbine equipment terms, fault symptoms, and maintenance semantics.

(2) A domain-aware hybrid retrieval mechanism is developed by integrating semantic retrieval, keyword retrieval, term matching, and re-ranking to improve retrieval accuracy and consistency.

(3) The proposed RAG framework enhances the reliability of question answering and decision support in wind turbine operation and maintenance, while providing a transferable framework for other industrial fault diagnosis scenarios.

II. RELATED WORKS

A. Retrieval-Augmented Generation for Industrial Knowledge Services

RAG provides an effective technical route for applying large language models to industrial knowledge services. Lin *et al.* [16] proposed MAKG, a Multi-Agent and Synergistic Knowledge Graph framework that integrates RAG with a multi-agent system for intelligent industrial equipment maintenance. The MAKG framework enables large language model agents to collaborate by drawing on information from a diverse knowledge base and generating coordinated responses to equipment faults. Chen *et al.* [17] proposed a RAG-based interactive industrial knowledge management system that supports technical service queries and internal regulation searches by retrieving and generating answers from heterogeneous industrial documents. Bahr *et al.* [18] proposed a knowledge graph-enhanced RAG framework for FMEA data, aiming to improve semantic question answering and failure-analysis support by standardizing FMEA information, constructing an FMEA knowledge graph, and generating vector embeddings from it.

These studies demonstrate the value of RAG for improving knowledge access and decision support in industrial scenarios; however, their effectiveness still depends heavily on retrieval quality, especially when dealing with domain-specific terminology, heterogeneous documents, and complex fault descriptions, which motivates further research on domain-specific embedding fine-tuning and knowledge-enhanced retrieval.

B. Domain-Specific Embedding Fine-Tuning and Knowledge-Enhanced Retrieval

The retrieval module largely determines the reliability of RAG systems. In industrial scenarios, general-purpose embeddings often fail to capture specialized terminology, fault semantics, and maintenance knowledge, making domain-specific embedding fine-tuning and knowledge-enhanced retrieval essential for improving retrieval relevance, factual grounding, and generation reliability. Liu *et al.* [19] proposed an Agent-driven framework for automated DFA evaluation, in which a domain-specific DFA-LLM is fine-tuned through domain-specific continual pre-training and DFA instruction data, and integrated with an AAS-based hybrid retrieval system to improve the accuracy and factual reliability of engineering assessments. Luo *et al.* [20] proposed CRGS-RAG, a retrieval-augmented generation framework that combines causal reasoning fine-tuning with a game-theory-inspired knowledge fusion strategy to improve evidence selection from noisy retrieved documents and enhance robustness when integrating internal parametric knowledge with external retrieved knowledge. Wang *et al.* [21] proposed a scientific RAG framework that integrates

contrastive learning and multi-level retrieval to improve multi-element, cross-document evidence retrieval from text, tables, formulas, and figures, while incorporating Chain-of-Thought reasoning for coherent knowledge synthesis and constructing the SciLitQA dataset for evaluation. Overall, existing studies have enhanced RAG reliability through domain adaptation, knowledge-enhanced retrieval, and evidence fusion. However, they remain limited in industrial fault semantic modeling, domain-aware hybrid retrieval, and retrieved-context quality control.

III. RESEARCH METHODS

A retrieval-augmented knowledge generation framework is proposed for industrial fault diagnosis and intelligent operation and maintenance, as illustrated in **Fig. 1**. In this study, wind turbine operation and maintenance is selected as a representative scenario for framework instantiation and experimental validation. In the wind turbine case study, the framework constructs an external knowledge base from maintenance manuals, technical documents, and domain-specific materials. These documents are cleaned, chunked, and vectorized to construct a retrievable knowledge base. Domain-specific data are further used to fine-tune the embedding model, enabling it to more accurately represent the semantic relationships among technical terminology, equipment structures, fault descriptions, user queries, and knowledge texts in the target domain.

After receiving a user query, the system first preprocesses and expands it with domain terms, synonymous expressions, and related concepts to improve candidate recall. It then applies a domain-aware hybrid retrieval mechanism, where dense retrieval captures semantic relevance and sparse retrieval strengthens exact keyword matching. A domain term matching score is further introduced and fused with the retrieval scores to evaluate candidate chunks more comprehensively. Finally, a cross-encoder re-ranks the candidates to select highly relevant, informative, and low-redundancy contexts, which are combined with the query and input into the large language model to generate accurate, professional, and traceable operation and maintenance responses.

10.37965/jdmd.2026.1505

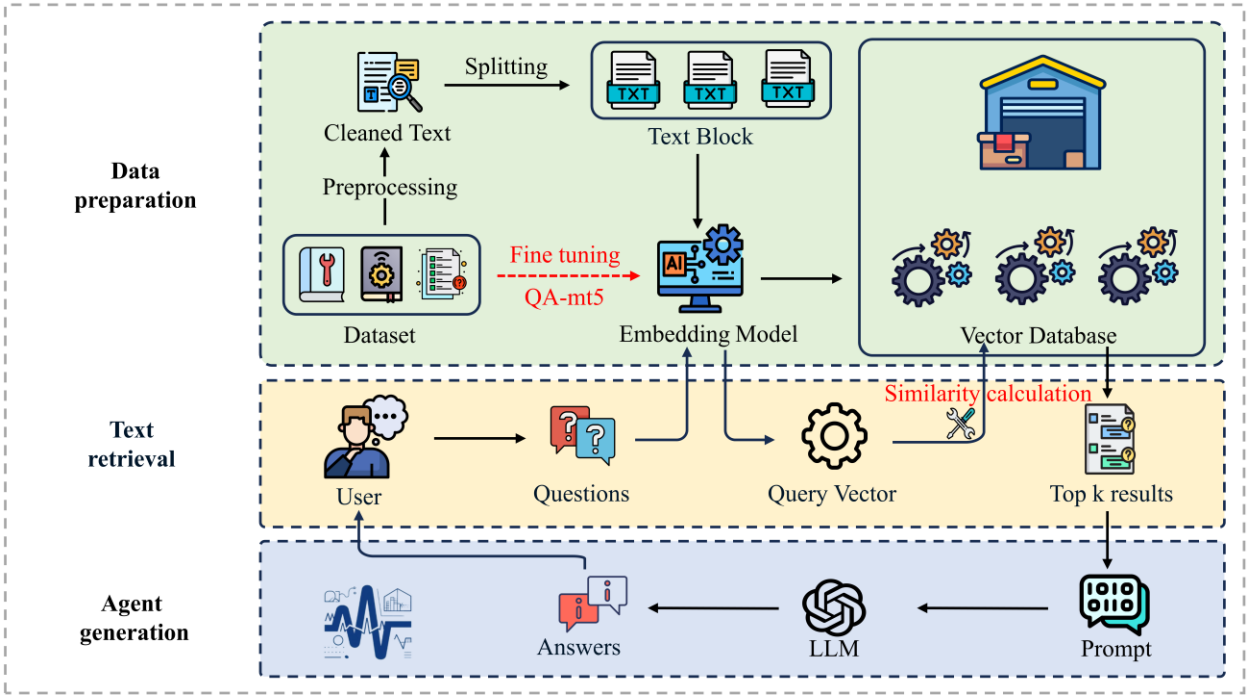

Fig. 1. Flowchart of RAG based on Fine-Tuning of embedding models.

A. Embedding Model Fine-Tuning Guided by Domain Question–Answer Data

To improve the retrieval model's semantic representation of wind turbine operation and maintenance knowledge, a domain question-answering-driven embedding fine-tuning method is proposed, as shown in **Fig. 2**. The method first generates domain-specific question–answer samples from topic information and local knowledge documents, and then constructs high-quality data through document-constrained generation. Finally, question–positive context pairs are used for contrastive fine-tuning, enabling the embedding model to better capture semantic relations among technical terms, fault descriptions, and maintenance knowledge fragments.

10.37965/jdmd.2026.1505

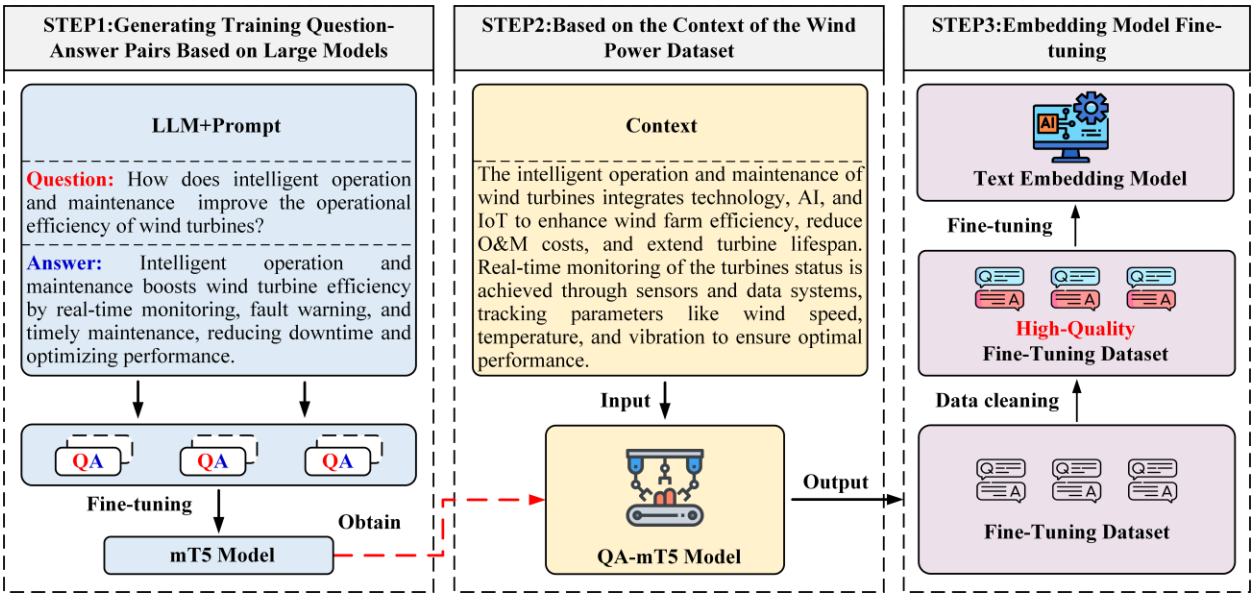

Fig. 2. Framework diagram of multi-turn question answer generation model.

1. Construction of Domain Question–Answer Samples

To alleviate the shortage of labeled question–answer data in the field of wind turbine operation and maintenance, this study first constructs a domain topic set and employs a local large language model to generate initial question–answer samples. The domain topic set is defined as follows:

$$\mathrm{T} = \{t_1, t_2, \cdots, t_i\} \tag{1}$$

where t_i denotes the i-th wind turbine operation and maintenance topic.

For each topic, the corresponding domain knowledge background z_i is first generated, and then several question–answer samples are generated under the constraint of this background:

$$z_i = G_\phi(t_i) \tag{2}$$

where $G_\phi(\cdot)$ denotes the generation function of the local large language model. Accordingly, the initial domain question–answer set can be obtained as:

$$\mathrm{D}_{qa}^{0} = \{q_{ij}, a_{ij}, z_i\} \tag{3}$$

The initial question–answer set is used to transform domain knowledge into structured samples in the form of "question–answer–background", enabling the model to learn typical expressions of wind turbine operation and maintenance problems as well as the corresponding answer organization patterns. On this basis, the initial question–answer set is further used to perform domain adaptation for the mT5 generative model. Let $x_i = [q_i, z_i]$ denote the input sequence and $y_i = a_i$ denote the target output. The training objective of the generative model can then be expressed as:

$$\mathrm{L}_{\mathrm{gen}} = -\sum_{i=1}^{|\mathrm{D}_{qa}^{0}|}\sum_{k=1}^{|y_i|} \log P_\theta\left(y_{i,k} \mid y_{i,<k}, x_i\right) \tag{4}$$

where θ denotes the parameters of the generative model, and $y_{i,k}$ denotes the k-th token in the target answer sequence.

2. Refinement of Document-Constrained Question–Answer Data

Although the initial question–answer samples can reflect the expression patterns of domain-specific problems, their content is mainly generated by the model and may still deviate from the texts in the real knowledge base. To improve the factual constraints and knowledge traceability of the training data, this study further introduces local domain

documents and constructs high-quality question–answer samples based on the document content.

Let D denote the original domain document set. It is divided into multiple knowledge fragments using a text segmentation function, which can be expressed as:

$$\mathrm{C} = S(\mathrm{D}) = \{c_1, c_2, \cdots, c_i\} \tag{5}$$

where c_i denotes the i -th document knowledge chunk. For each knowledge chunk, the model first generates a question q_i related to its content, and then generates an answer a_i based on the same knowledge chunk.

Each answer is grounded in its corresponding knowledge chunk to ensure source consistency. Low-quality samples, including short, duplicate, unsupported, or prompt-contaminated entries, are then removed to obtain the final document-constrained question–answer set:

$$\mathrm{D}_{qa} = \mathrm{F}\left(\left\{\left(q_i, a_i, c_i\right)\right\}_{i=1}^{L}\right) \tag{6}$$

Compared with the initial question–answer samples, the document-constrained question–answer data not only include possible question forms that users may raise but also retain the original knowledge chunks on which the answers depend.

3. Embedding Model Fine-Tuning Based on Contrastive Learning

After constructing the document-constrained question–answer set, the data are used for domain-specific fine-tuning of the embedding model. The retrieval stage aims to identify knowledge chunks most relevant to user queries, rather than directly generating answers. Therefore, during the training of the embedding model, the question q_i and its corresponding knowledge chunk c_i are constructed as a positive sample pair, while the answer a_i is not directly used as the retrieval target.

Let $f(\cdot)$ denote the embedding model. The question and the corresponding knowledge chunk are encoded separately and then normalized using L_2 normalization, yielding:

$$\hat{h}_{q_i} = \frac{f_\omega(q_i)}{\| f_\omega(q_i) \|_2} \tag{7}$$

$$\hat{h}_{c_i} = \frac{f_\omega(c_i)}{\| f_\omega(c_i) \|_2} \tag{8}$$

Within a batch, the similarities between all question vectors and knowledge chunk vectors are calculated as follows:

$$s_{ij} = \tau \cdot \hat{h}_{qi}^{T} \hat{h}_{c_j} \tag{9}$$

where τ denotes the temperature scaling coefficient.

Based on the obtained similarity matrix, the embedding model is optimized using a contrastive learning loss with in-batch negative samples:

$$\mathrm{L}_{emb} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ij})} \tag{10}$$

where B denotes the batch size.

For reproducibility, the embedding model is fine-tuned with an in-batch contrastive learning objective using document-grounded question–context pairs, and the detailed training configuration is summarized in **Table 1**.

Table 1. Training configuration for embedding model fine-tuning.

Parameter	Configuration
Base model	BGE-base-zh
Embedding dimension	768
Pooling	[CLS] token
Normalization	L2 normalization
Batch size	16
Learning rate	2×10^{-5}
Max sequence length	256
Optimizer	AdamW
Gradient clipping	1.0

This loss function helps the embedding model learn domain-specific semantic matching by pulling positive pairs closer and pushing non-matching pairs apart.

B. Domain-Aware Hybrid Retrieval Mechanism

As shown in **Fig. 3**, the domain-aware hybrid retrieval mechanism improves wind turbine maintenance knowledge retrieval by combining query expansion, dense semantic retrieval, and sparse keyword retrieval [22]. A domain term consistency score is then introduced to enhance professional relevance, and the final results are obtained through weighted fusion and cross-encoder re-ranking [23]. This method jointly models semantic similarity, keyword matching, and domain term consistency to improve retrieval coverage, accuracy, and context quality.

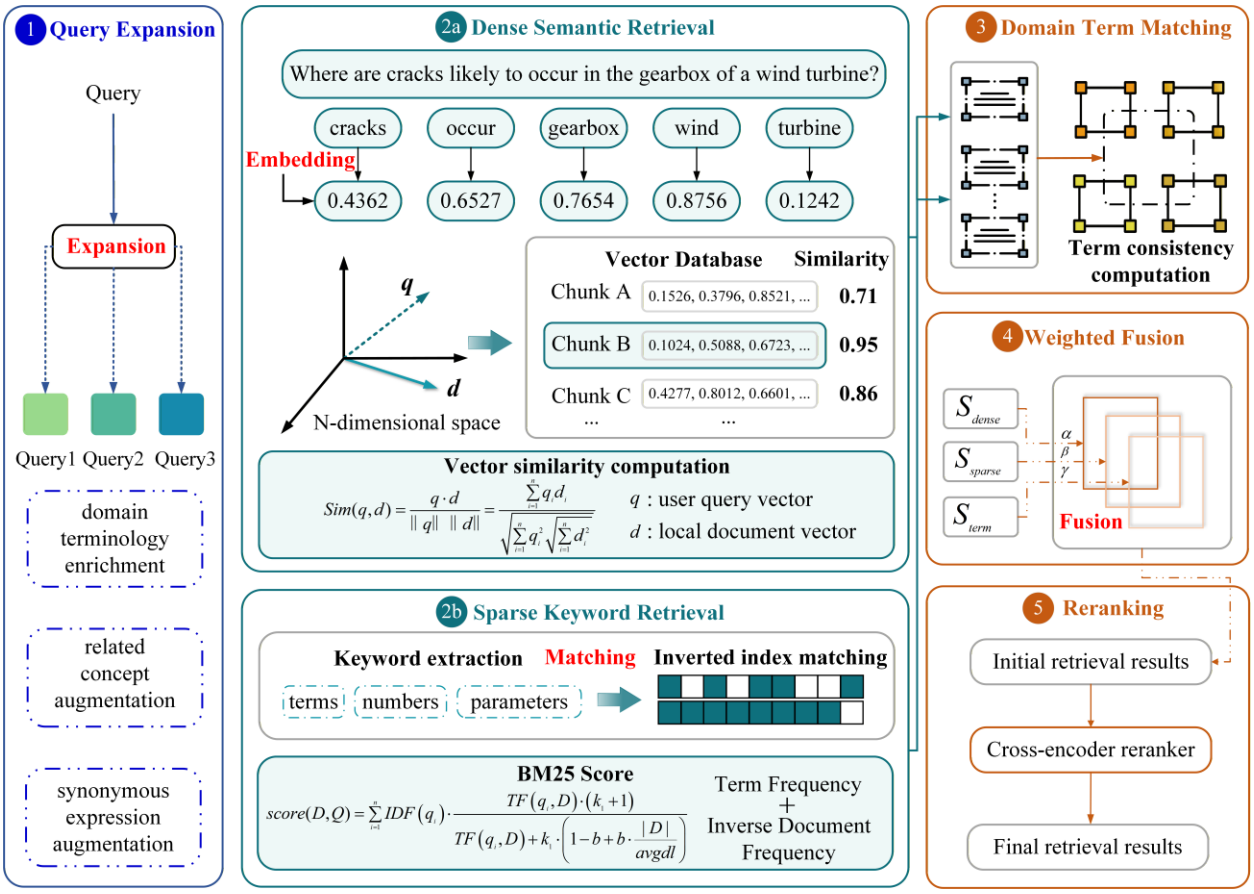

Fig. 3. Schematic diagram of the vector indexing process.

1. Query Recall

For the original user query q_0, query expansion is first performed to obtain a query set containing domain terms, related concepts, and synonymous expressions:

$$Q = \{q_0, q_1, q_2, \cdots, q_m\} \tag{11}$$

where q_m denotes the expanded query expression. Through query expansion, the user's natural language question can be transformed into a retrieval form that is more consistent with the expression patterns of the domain knowledge base, thereby improving the coverage of the initial recall stage.

In the dense semantic retrieval stage, the embedding model is used to map the query and document chunks into the same vector space. Let q denote the query vector and d_i denote the candidate document chunk vector.

The semantic similarity between them is calculated as:

$$S_{\text{dense}}(q,d_i)=\frac{\sum_{j=1}^{N}q_j d_{ij}}{\sqrt{\sum_{j=1}^{N}q_j^2}\sqrt{\sum_{j=1}^{N}d_{ij}^2}} \quad (12)$$

This score measures the proximity between the query and the document chunk in the semantic space, thereby capturing latent semantic associations across different expressions.

To enhance the matching capability for explicit information such as equipment names, fault locations, and technical terms [24] , sparse keyword retrieval is introduced. This process is implemented based on keyword extraction and an inverted index, and BM25 is used to calculate the matching score between the query Q and the document chunk D :

$$S_{sparse}(Q,D)=\sum_{t\in Q}IDF(t)\cdot\frac{TF(t,D)(k_1+1)}{TF(t,D)+k_1\left(1-b+b\frac{|D|}{avgdl}\right)} \quad (13)$$

where t denotes a keyword in the query, $TF(t,D)$ denotes the term frequency of keyword (t) in the document chunk, $IDF(t)$ denotes the inverse document frequency, $|D|$ is the length of the document chunk, $avgdl$ is the average length of all document chunks, and k_1 and b are tuning parameters. Dense semantic retrieval focuses on semantic-level similarity, whereas sparse keyword retrieval emphasizes exact matching at the term level. Together, they form a complementary dual-channel recall mechanism.

2. Fusion Ranking

After the initial candidate document chunks are obtained, a domain term consistency score is calculated to constrain the domain-specific relevance of the retrieval results. Let T_Q denote the set of domain terms identified in the query and T_D denote the set of domain terms contained in the candidate document chunk. The domain term matching score is defined as:

$$S_{term}(Q,D)=\frac{\sum_{\tau\in T_Q\cap T_D}w_\tau}{\sum_{\tau\in T_Q}w_\tau+\varepsilon} \quad (14)$$

where τ denotes a domain term, w_τ denotes the weight of term, and ε is a smoothing factor. This score is used to measure the matching degree between the candidate document chunk and the user query.

Considering that the dense semantic score, sparse keyword score, and domain term score are distributed on different numerical scales, this study first normalizes the three scores and then performs weighted fusion to obtain the comprehensive retrieval score:

$$S=\alpha\hat{S}_{dense}+\beta\hat{S}_{sparse}+\gamma\hat{S}_{term} \quad (15)$$

where $\hat{S}_{dense}$, $\hat{S}_{sparse}$, and $\hat{S}_{term}$ denote the normalized semantic similarity score, keyword matching score, and domain term consistency score, respectively; and α , β , and γ are the corresponding weight coefficients.

Finally, a cross-encoder [25] is employed to re-rank the initial candidate results [26]. Let C denote the set of candidate document chunks obtained after weighted fusion. The cross-encoder jointly encodes the query Q and each candidate document chunk D_i , and outputs the final relevance score:

$$R_i=g_\phi(Q,D_i),\quad D_i\in C \quad (16)$$

where g_ϕ denotes the cross-encoder re-ranking model, and R_i denotes the final relevance score of the candidate document chunk D_i . The candidate document chunks are re-ranked according to R_i , and the top-K document chunks are selected as the input context for the generative model:

$$C_{topK}=TopK_{D_i\in C}(R_i) \quad (17)$$

In the implemented system, cross-encoder reranking is performed by Qwen2.5-

3B-Instruct, which jointly reads the user query and a numbered list of candidate passages within a single prompt. After hybrid retrieval and domain-term reweighting, the retained candidates are reranked, and the final Top-5 passages are selected according to the model-generated ranked index list. The detailed settings are summarized in **Table 2**.

Table 2. Settings of the cross-encoder reranking module

Parameter	Configuration
Model	Qwen2.5-3B-Instruct
Input	Query + numbered passages
Candidates	15
Output	Top-5
Decoding	Greedy

Through the above process, the domain-aware hybrid retrieval mechanism can further select knowledge chunks that are most relevant to the user query, contain more complete information, and have lower redundancy from the initially retrieved results, thereby providing high-quality contextual support for subsequent retrieval-augmented generation.

IV. CASE STUDY

A. Comparative experiments

To verify the overall effectiveness of the proposed method, comparative experiments are conducted against several representative baselines, including the base large language model, conventional retrieval-augmented generation, embedding-model-fine-tuned retrieval-augmented generation, and conventional hybrid-retrieval retrieval-augmented generation. This comparison is designed to evaluate the contribution of domain-specific embedding fine-tuning and domain-aware hybrid retrieval to retrieval quality and answer generation performance.

To ensure a fair and unambiguous comparison, the roles of the model components are specified as follows. Qwen2.5-3B-Instruct is the locally deployed instruction-tuned causal language model used for final answer generation in all compared systems (M1–M5); in M5, the same model is additionally used for query expansion and LLM-based candidate reranking. Keeping this generator unchanged across systems ensures that performance differences mainly reflect retrieval enhancement, embedding fine-tuning, and the proposed M5 modules. mT5-small, fine-tuned as QA-mT5, is used only in offline data preparation: it is trained on seed QA pairs produced by Qwen and serves as a lightweight domain QA generator during data construction, but is not used for final answer generation or online inference. Document-grounded refined QA pairs for embedding training are generated directly by Qwen from local domain documents. BGE-base-zh is the backbone embedding model, fine-tuned with question–context pairs for dense retrieval only and is not involved in answer generation.

The five comparative models used in this study are listed in **Table 3**. M1 denotes the base large language model without access to an external knowledge base. M2 represents standard RAG using only dense vector retrieval. M3 incorporates embedding model fine-tuning on the basis of M2. M4 adopts conventional hybrid retrieval by combining dense retrieval and sparse retrieval. M5 is the complete proposed method, which further integrates query expansion, domain term matching, and LLM-based candidate reranking.

Table 3. Comparative methods and key module settings.

Method	Embedding Fine-Tuning	Query Expansion	Dense Retrieval	Sparse Retrieval	Term Matching	Re-ranking
M1	×	×	×	×	×	×
M2	×	×	√	×	×	×

M3	√	×	√	×	×	×
M4	×	×	√	√	×	×
M5	√	√	√	√	√	√

1. Retrieval Performance Evaluation

To evaluate the knowledge retrieval capability of different methods, Recall@1, Recall@3, Recall@5, Precision@5, MRR, and nDCG@5 are used as evaluation metrics, and the results are presented in **Table 4**. Among these metrics, Recall@K measures the coverage of relevant documents within the top-(K) retrieved results, reflecting the retrieval method's ability to provide sufficient candidate knowledge. Precision@5 evaluates the accuracy of the top five results, while MRR indicates the ranking position of the first relevant document. nDCG@5 further considers both ranking position and relevance grade. Together, these metrics enable retrieval performance to be analyzed in terms of recall capability, matching accuracy, and ranking quality, providing an objective basis for comparing different methods.

Table 4. Retrieval performance of different comparative methods.

Method	Recall@1	Recall@3	Recall@5	Precision@5	MRR	nDCG@5
M1	-	-	-	-	-	-
M2	5.33	14.67	24.00	14.40	28.13	20.80
M3	24.00	58.67	70.67	42.40	78.80	68.25
M4	33.33	69.33	70.67	43.45	91.00	76.68
M5	32.00	73.33	81.33	48.80	98.00	83.44

To intuitively compare the retrieval performance differences among different methods, the evaluation results in **Table 4** are visualized, as shown in **Fig. 4**. The visualization compares different methods across Recall@K, Precision@5, MRR, and nDCG@5, which respectively reflect retrieval completeness, result relevance, the ranking position of the first relevant document, and overall ranking quality. Compared with the tabular results, it more clearly reveals the performance gaps among different methods and the improvement trend from the baseline model to the proposed method. In particular, the recall-related metrics indicate whether the retrieval system can cover sufficient domain knowledge, while MRR and nDCG@5 further reflect its ability to rank highly relevant evidence in top positions. Therefore, the visualization provides a direct basis for analyzing how embedding fine-tuning, hybrid retrieval, and re-ranking contribute to retrieval effectiveness.

10.37965/jdmd.2026.1505

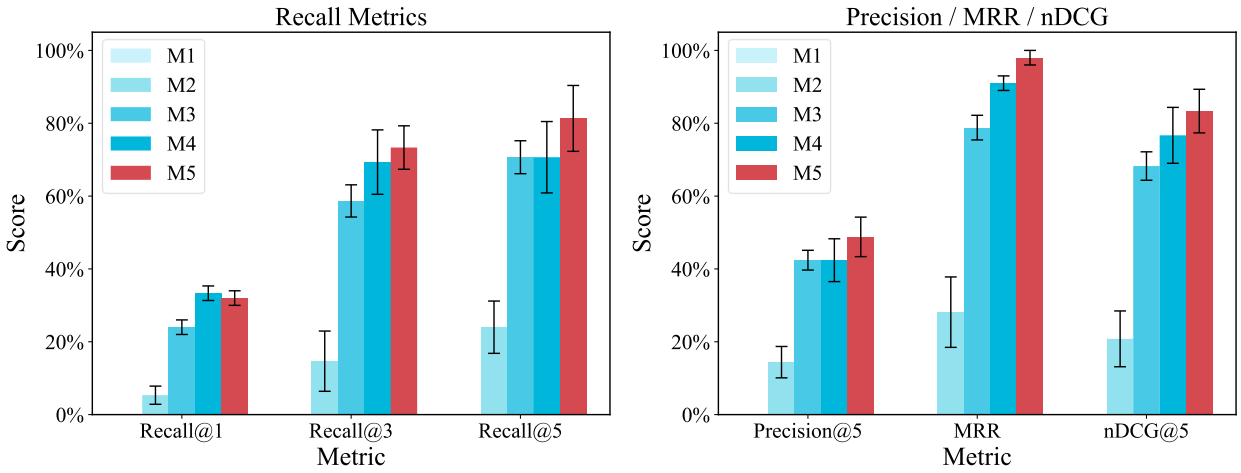

Fig. 4. Comparison of retrieval performance metrics across different methods.

All retrieval metrics of M1 are zero because M1 is a pure language model baseline without a retrieval module and thus cannot retrieve relevant content from the document database. The overall performance of M2 is relatively low, indicating that general-purpose embeddings have obvious limitations in domain-specific retrieval. After introducing the fine-tuned embedding model, M3 achieves a substantial performance improvement, with Recall@1 reaching 24% and MRR reaching 78.8%. M4 further improves the ranking quality through hybrid retrieval, increasing MRR to 91%. M5 achieves the best performance in terms of both recall depth and overall ranking quality, with Recall@5 of 81.33%, MRR of 98%, and nDCG@5 of 83.44%, which verifies the effectiveness of the proposed retrieval strategy.

To further highlight the overall variation trends of different methods across multiple retrieval metrics, a line chart is additionally plotted, as shown in **Fig. 5**. The final model, M5, achieves the best performance on most metrics. The test questions are further divided into five categories, namely basic principles, equipment structure, fault diagnosis, safety specifications, and inspection and maintenance, to examine the adaptability of the model across different knowledge scenarios. The results are shown in **Fig. 6**.

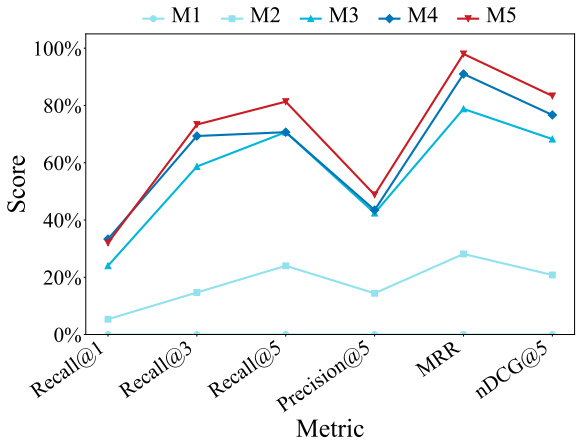

Fig. 5. Comparison of retrieval performance metrics across models.

The bubble chart further illustrates retrieval differences across knowledge categories. M2 achieves limited and uneven performance, indicating that general-purpose embeddings struggle to generalize to different wind turbine maintenance scenarios. After embedding fine-tuning, M3 shows clear improvements across categories, with a more balanced bubble distribution and better cross-category consistency. M4 performs well in several categories due to the combination of dense and sparse retrieval, but its results remain uneven. In contrast, M5 presents larger and more uniform bubbles, demonstrating more stable retrieval performance and stronger cross-category generalization.

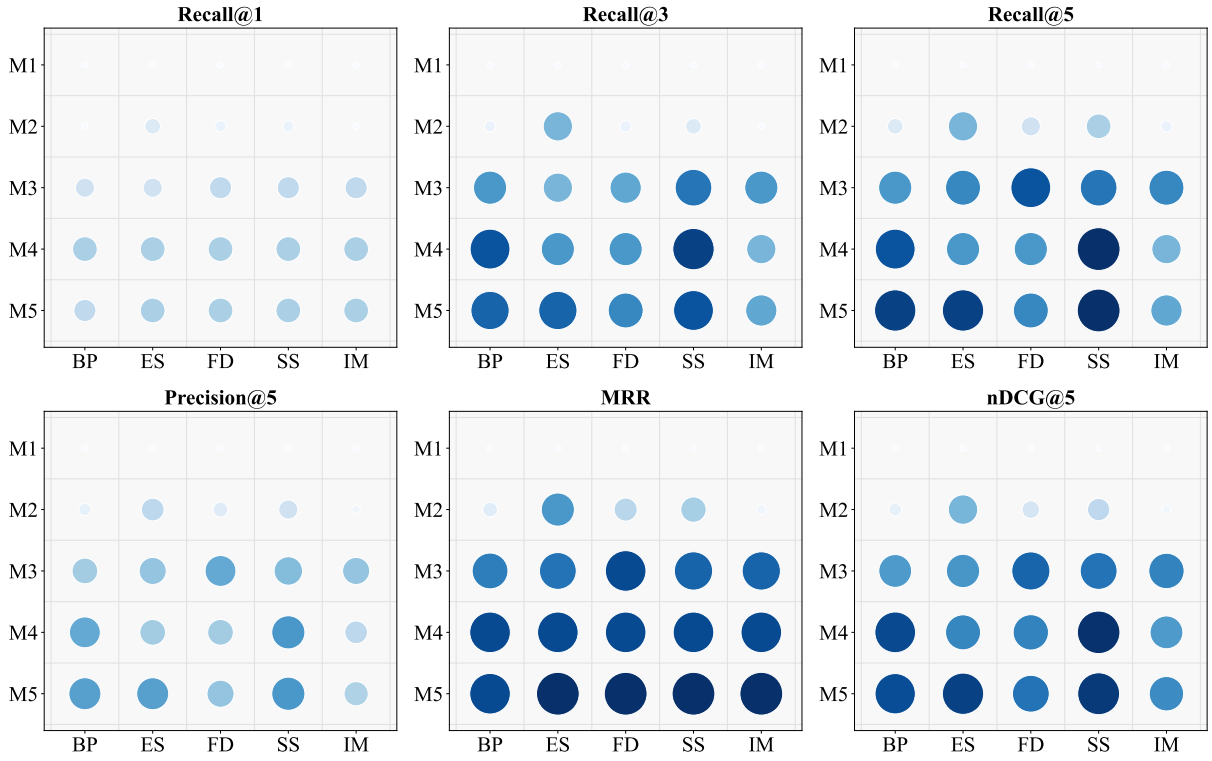

Fig. 6. Retrieval performance bubble matrix across multi-category question scenarios.

2. Generation Quality Evaluation

This study adopts five automated objective metrics to evaluate generation quality: SemSim, KPC, FSR, HR, and DTF1, as listed in **Table 5**. These metrics assess the generated answers in terms of semantic consistency, key-point coverage, factual support, hallucination level, and domain terminology usage. Specifically, SemSim is computed as the cosine similarity between the sentence embeddings of the generated and reference answers. KPC measures the proportion of predefined reference key points covered by the generated answer. FSR is obtained by checking whether the atomic factual claims in the generated answer are supported by the retrieved contexts, while HR measures the proportion of unsupported claims. DTF1 is calculated from the precision and recall of domain terms extracted from the generated and reference answers. Higher values of SemSim, KPC, FSR, and DTF1 indicate better generation quality, whereas a lower HR indicates fewer hallucinations. The final score of each metric is averaged over all test questions.

Table 5. Objective evaluation metrics for generation performance.

Abbreviation	Full Name	Purpose
SemSim	Semantic Similarity	Evaluates the semantic consistency between the generated answer and the reference answer
KPC	Key-point Coverage	Evaluates whether the generated answer covers the key knowledge points
FSR	Fact Support Rate	Evaluates whether the generated answer is supported by the retrieved context
HR	Hallucination Rate	Evaluates the proportion of unsupported generated content
DTF1	Domain Term F1	Evaluates the accuracy of domain-specific terminology usage

Table 6 and **Fig. 7** present the comparison results of the five systems on five generation quality metrics. In terms of semantic similarity, the M1 baseline system achieves a score of 75.91%, which increases to 79.53% after retrieval augmentation is introduced in M2. Supported by fine-tuned embeddings and hybrid retrieval, M3 and M4 reach 80.22% and 79.97%, respectively, showing only slight differences compared with M2. The proposed method, M5, achieves the highest value of 80.85%.

Key point coverage is the most discriminative metric among the five systems. M1 achieves only 26.47%, indicating that the large language model has difficulty covering domain-specific knowledge points without retrieval support. After retrieval is introduced, M2 increases substantially to 62.60%. M3 and M4 further improve this metric to 65.80% and 65.60%, respectively, while M5 reaches 67.40%, verifying the effectiveness of the proposed retrieval strategy in improving knowledge coverage.

In terms of factual support rate and hallucination rate, M1 obtains a factual support rate of 91.21% and a hallucination rate of 8.79%. With improved retrieval capability, M2 slightly increases the factual support rate and reduces the hallucination rate to 8.19%. M3 and M4 greatly reduce the hallucination rate to 4.57% and 4.33%, respectively, through more accurate retrieval. M5 achieves the highest factual support rate.

For domain terminology F1, M3 and M5 achieve 63.28% and 64.76%, respectively, indicating that fine-tuned embeddings positively contribute to the accurate use of domain-specific terminology.

Table 6. Generation quality evaluation results of different comparative methods.

Method	SemSim	KPC	FSR	HR	DTF1
M1	75.91	26.47	91.21	8.79	41.66
M2	79.53	62.60	91.81	8.19	58.75
M3	80.22	65.80	95.43	4.57	63.28
M4	79.97	65.60	95.67	4.33	58.73
M5	80.85	67.40	97.83	2.17	64.76

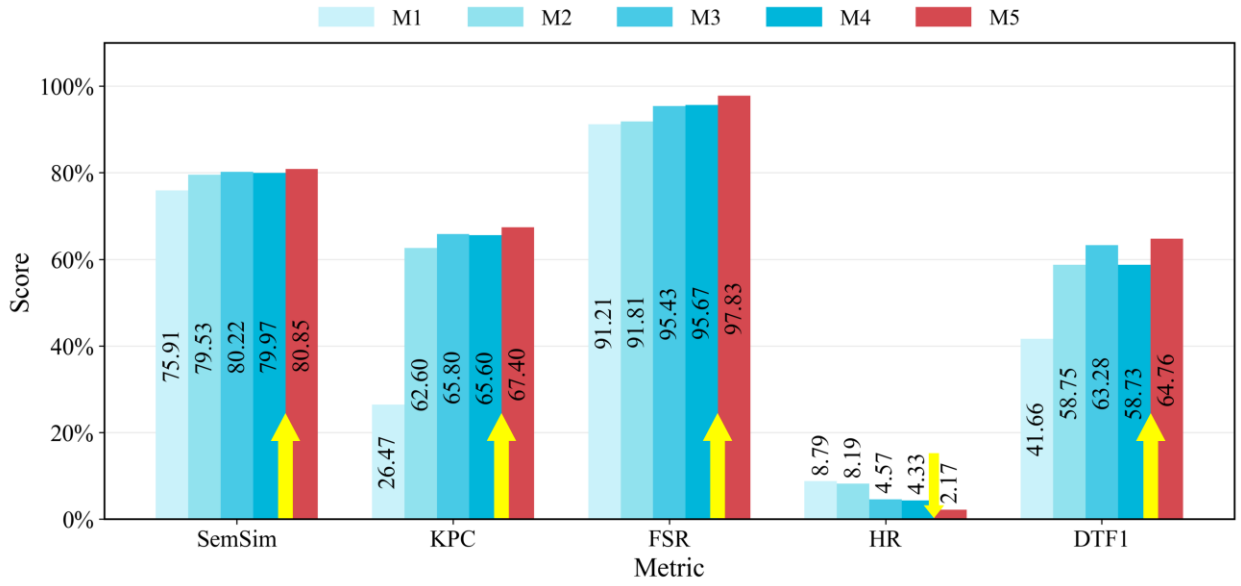

Fig. 7. Comparison of objective generation performance metrics across different methods.

B. Ablation experiments

To further examine the contribution of each module in M5, ablation experiments were conducted by removing embedding fine-tuning, query expansion, sparse retrieval, domain term matching, and cross-encoder re-ranking, respectively. Under the same data and evaluation metrics, retrieval performance and generation quality were compared to reveal each module's impact on knowledge recall, ranking quality, factual support, and domain-term accuracy. The ablation settings are shown in **Table 7**.

Table 7. Configuration of ablation models obtained by removing one module from the full M5 system.

Method	Embedding Fine-Tuning	Query Expansion	Dense Retrieval	Sparse Retrieval	Term Matching	Re-ranking
M5	√	√	√	√	√	√
w/o-FT	×	√	√	√	√	√
w/o-QE	√	×	√	√	√	√
w/o-SR	√	√	√	×	√	√
w/o-DT	√	√	√	√	×	√
w/o-CE	√	√	√	√	√	×

1. Retrieval Performance Evaluation

Retrieval performance is evaluated using Recall@1, Recall@3, Recall@5, Precision@5, MRR, and nDCG@5, which comprehensively reflect recall capability, matching accuracy, and ranking quality. The results are shown in **Table 8**, with the corresponding visualization presented in **Fig. 8**.

Table 8. Retrieval performance of different ablation models.

Method	Recall@1	Recall@3	Recall@5	Precision@5	MRR	nDCG@5
M5	32.00	73.33	81.33	48.80	98.00	83.44
w/o-FT	26.31	51.45	61.33	36.80	79.85	61.86
w/o-QE	31.33	62.51	78.67	47.20	86.24	82.74
w/o-SR	28.31	52.65	62.67	37.60	77.81	62.67
w/o-DT	29.33	66.67	73.33	44.00	94.58	75.14
w/o-CE	31.20	54.67	64.12	38.40	74.89	63.98

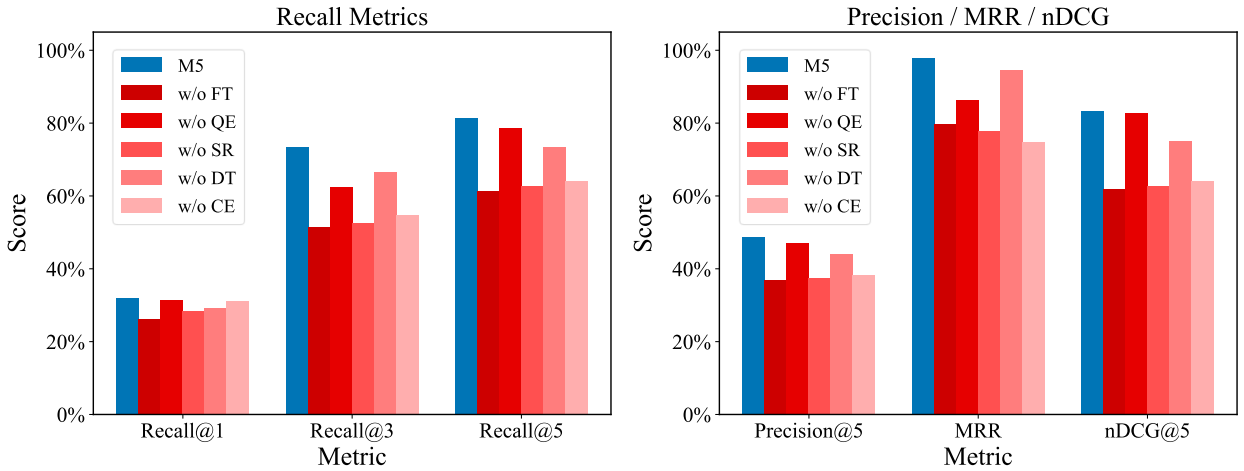

Fig. 8. Comparison of retrieval performance metrics across different ablation variants.

The retrieval ablation results show that removing any component degrades M5 performance, confirming the complementarity of the proposed modules. Removing embedding fine-tuning or sparse retrieval markedly reduces recall, precision, and ranking quality, indicating the importance of domain semantic representation and keyword matching. The absence of cross-encoder re-ranking causes the largest MRR drop, highlighting its role in prioritizing relevant evidence. Query expansion and domain term matching further enhance recall stability and domain consistency.

2. Generation Quality Evaluation

Generation performance is evaluated using SemSim, KPC, FSR, HR, and DTF1, which measure semantic consistency, key-point coverage, factual support, hallucination level, and domain terminology usage, respectively. The results are shown in **Table 9** , and the performance differences among methods are visualized in **Fig. 9**.

Table 9. Generation quality evaluation results of different ablation variants.

Method	SemSim	KPC	FSR	HR	DTF1
M5	80.85	67.40	97.83	2.17	64.76
w/o-FT	79.22	62.61	95.73	4.27	61.53
w/o-QE	79.64	65.87	97.40	2.60	62.83
w/o-SR	79.82	65.89	97.38	2.62	60.71
w/o-DT	80.45	66.54	96.56	3.44	62.72
w/o-CE	79.24	62.62	95.79	4.21	61.53

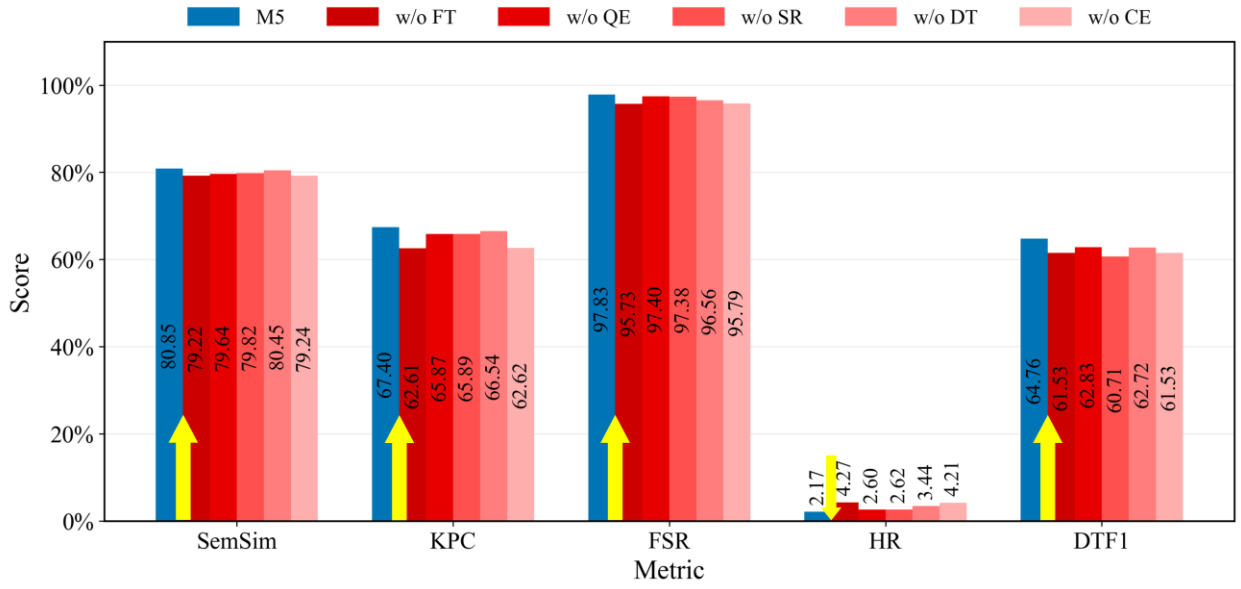

Fig. 9. Comparison of objective generation performance metrics across different ablation variants.

The generation-quality ablation results show that each module contributes to answer reliability. M5 achieves the best overall performance, with the highest SemSim, KPC, FSR, and DTF1 and the lowest HR. Removing embedding fine-tuning or cross-encoder re-ranking reduces answer completeness and factual support, while removing sparse retrieval or domain term matching weakens domain-term accuracy.

These results confirm that the proposed modules jointly improve knowledge coverage, factual grounding, hallucination suppression, and domain-specific expression.

V. CONCLUSIONS

To address the limited domain understanding of general-purpose models, insufficient retrieval relevance, and weak answer traceability in industrial fault diagnosis and intelligent operation and maintenance, this study proposes a retrieval-augmented generation framework that integrates data-driven embedding fine-tuning and domain-aware hybrid retrieval. The framework is intended for industrial fault diagnosis and intelligent maintenance, and is instantiated and validated in wind turbine operation and maintenance as a representative application scenario. By constructing document-constrained question–answer data and combining query expansion, dense retrieval, sparse retrieval, domain term matching, weighted fusion, and cross-encoder re-ranking, the proposed method improves knowledge recall, ranking quality, and generation reliability. The ablation experiments further demonstrate the effectiveness of each key module: embedding fine-tuning enhances domain semantic representation, hybrid retrieval improves both semantic and keyword-level matching, and the integration of domain term matching and re-ranking further strengthens retrieval relevance, factual support, and answer traceability [27].

Nevertheless, the current study is mainly validated using wind turbine-related textual knowledge resources, and its scalability, cross-domain generalization, deployment efficiency, and performance in complex fault scenarios require further investigation [28]. Future work will focus on adapting the framework to other industrial fault diagnosis domains by reconstructing domain-specific knowledge bases, terminology dictionaries, and fine-tuning datasets [29]. In addition, dynamically updated knowledge bases, lightweight retrieval and re-ranking mechanisms, and multi-source multimodal information such as SCADA data, vibration signals, images, maintenance work orders, fault cases, and expert experience will be further incorporated to improve deployment efficiency and enhance diagnostic capability under complex operating conditions.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (52505091).

REFERENCES

[1] X. Li, B. Kang, J. Tang, Q. Li, T. Wang, and M. Zhang, "A Single-Device Environment-Adaptive Mixed Reality Framework for Real-Time Industrial Fault Diagnosis," *Journal of Dynamics, Monitoring and Diagnostics,* vol. 5, no. 1, pp. 64-73, 2026.

[2] X. Li *et al.*, "A fault diagnosis data augmentation method integrating multimodal non-Gaussian denoising diffusion generative adversarial network," *Advanced Engineering Informatics,* vol. 68, p. 103776, 2025.

[3] K. Su *et al.*, "Harmonising convolutional networks and Transformer for transfer learning in bearing fault diagnosis," *Nondestructive Testing and Evaluation,* pp. 1-25, 2026.

[4] J. Zhuang *et al.*, "A neuro-fuzzy approach with hypergraph convolution for fault diagnosis in industrial devices," *Journal of Reliability Science and Engineering,* vol. 1, no. 3, p. 035301, 2025.

[5] R. Huang, J. Xia, B. Zhang, Z. Chen, and W. Li, "Compound fault diagnosis for rotating machinery: State-of-the-art, challenges, and opportunities," *Journal of dynamics, monitoring and diagnostics,* pp. 13-29, 2023.

[6] S. Yu, X. Li, Y. Lei, B. Yang, N. Li, and K. Feng, "Multimodal Data-enabled Large Model for Machine Fault Diagnosis Towards Intelligent Operation and Maintenance," *Journal of Industrial Information Integration,* p. 101061, 2026.

[7] C. Zhang *et al.*, "A survey on potentials, pathways and challenges of large language models in new-generation intelligent manufacturing," *Robotics and Computer-Integrated Manufacturing,* vol. 92, p. 102883, 2025.

[8] K. Su, X. Kong, X. Li, T. Wang, and F. Chu, "BI-Sandwich: a cross-domain model for fault diagnosis of aircraft engine main shaft bearings," *Structural Health Monitoring,* p. 14759217261426251, 2026.

[9] X. Chen *et al.*, "Large models for machine monitoring and fault diagnostics: Opportunities, challenges, and future direction," *Journal of Dynamics, Monitoring and Diagnostics,* vol. 4, no. 2, pp. 76-90, 2025.

[10] X. Zhang, T. Wang, L. Ma, and S. Mahadevan, "Reliability engineering, risk management, and trustworthiness assurance for AI systems," *Journal of Reliability Science and Engineering,* vol. 1, no. 2, p. 022001, 2025.

[11] M. Jiang, Y. Yang, X. Xie, P. Ke, and G. Liu, "Safe and effective post-fine-tuning alignment in large language models," *Knowledge-Based Systems,* p. 114523, 2025.

[12] J. Yang, L. Yao, T. Zhang, C.-Y. Tsai, Y. Lu, and M. Shen, "Integrating prompt techniques and multi-similarity matching for named entity recognition in low-resource settings," *Engineering Applications of Artificial Intelligence,* vol. 144, p. 110149, 2025.

[13] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems,* vol. 33, pp. 9459-9474, 2020.

[14] X. Li, S. Li, G. Zhang, Y. Xie, T. Wang, and F. Chu, "A novel interpretable dynamic weighted domain adaptation network for cross-domain fault diagnosis of bearings under time-varying speeds," *Engineering Applications of Artificial Intelligence,* vol. 170, p. 114173, 2026.

[15] Z. Gao, Z. Liao, and C. Li, "Better interaction experience: human-machine interface for soft robotic systems," *Intelligence & Robotics,* vol. 5, no. 3, pp. 520-40, 2025.

[16] J. Lin, C. Chen, T. Wang, Z. Wang, and L. Cheng, "A Multi-Agent and synergistic Knowledge Graph retrieval-augmented generation framework for intelligent maintenance," *Journal of Manufacturing Systems,* vol. 85, pp. 557-570, 2026.

[17] L.-C. Chen, M. S. Pardeshi, Y.-X. Liao, and K.-C. Pai, "Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model," *Computer Standards & Interfaces,* vol. 94, p. 103995, 2025.

[18] L. Bahr, C. Wehner, J. Wewerka, J. Bittencourt, U. Schmid, and R. Daub, "Knowledge graph enhanced retrieval-augmented generation for failure mode and effects analysis," *Journal of Industrial Information Integration,* vol. 45, p. 100807, 2025.

[19] J. Liu *et al.*, "LLM Agent-driven Automated DFA Assessment with Fine-tuning and AAS-based RAG," *Robotics and Computer-Integrated Manufacturing,* vol. 102, p. 103325, 2026.

[20] X. Luo, Y. Chen, Y. Tu, and W. Qiu, "Causal Reasoning Meets Heuristic Strategies: Enhancing RAG through Fine-Tuning and Knowledge Interaction," *Knowledge-Based Systems,* p. 114976, 2025.

[21] S. Wang, Z. Wang, C. Qu, and Z. Yin, "Multi-Level Retrieval with Representation Alignment Enhances Cross-Document Evidence Synthesis for Scientific Knowledge Generation," *Applied Soft Computing,* p. 114649, 2026.

[22] Y. Gao, L. Bai, R. Shi, and H. Jiang, "Time-weaver query in sparse temporal knowledge graphs leveraging complex space embedding and adjacent timestamp relations," *Expert Systems with Applications,* p. 130177, 2025.

[23] D. Yang, B. Cao, S. Qu, F. Lu, S. Gu, and G. Chen, "Retrieve-then-compare mitigates visual hallucination in multi-modal large language models," *Intelligence & Robotics,* vol. 5, no. 2, pp. 248-275, 2025.

[24] X. Li, G. Li, T. Wang, Z. Dong, Z. Qin, and F. Chu, "Knowledge extraction and retrieval-augmented generation for intelligent maintenance of wind power equipment based on graph attention networks," *Chinese Journal of Mechanical Engineering,* p. 100141, 2025.

[25] Z. Wang, Z. Chen, Z. Chen, T. Xia, and E. Pan, "An intelligent diagnosis framework for quality defects in aircraft assembly using weighted text embedding and a stacked sparse autoencoder," *Journal of Reliability Science and Engineering,* vol. 1, no. 3, p. 035401, 2025.

[26] A. Kazemi, A. Toral, A. Way, A. Monadjemi, and M. Nematbakhsh, "Syntax-and semantic-based reordering in hierarchical phrase-based statistical machine translation," *Expert systems with applications,* vol. 84, pp. 186-199, 2017.

[27] X. Ba, X. Zhang, S. Li, J. Yuan, and J. Hu, "Multimodal semantic communication system based on graph neural networks," *Intelligence & Robotics,* vol. 5, no. 3, pp. 805-26, 2025.

[28] X. Li *et al.*, "A fault diagnosis method for bearings based on multiscale graph convolutional network under non-stationary speed conditions," *Mechanism and Machine Theory,* vol. 217, p. 106267, 2025.

[29] W. Feroze *et al.*, "Enhancing Temporal commonsense Understanding using disentangled attention-based method with a hybrid data framework," *Intell. Rob,* vol. 5, no. 1, pp. 228-247, 2025.