applied sciences

MDPI

Article

A Digital-Twin-Based and RAG-Based LLM Agent Framework for Intelligent Bearing Diagnosis and Maintenance

Bingran Zhu, Hai Huang * and Zhengkui Chen *

School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2025210601026@mails.zstu.edu.cn

* Correspondence: haihuang1005@gmail.com (H.H.); chenzk@zstu.edu.cn (Z.C.)

Abstract

Intelligent bearing diagnosis and maintenance require question-answering systems that combine asset-specific condition evidence with engineering knowledge. However, existing bearing diagnosis studies mainly predict fault labels or health states, while document-based RAG systems generally lack access to the condition of a specific asset. Digital twin-based systems provide structured condition data but offer limited support for manual-grounded diagnostic interpretation. Consequently, there is still no unified route-aware framework and benchmark for determining the evidence that is required by different diagnostic questions and evaluating route selection, DT querying, document retrieval, and answer grounding separately. To address this gap, this paper proposes a route-aware digital twin (DT) and retrieval-augmented generation (RAG) large language model (LLM) agent framework for bearing diagnosis question answering (QA). The framework routes questions to four paths: unrelated, DT-only, RAG-only, and DT+RAG. A route-specific 140-question benchmark is constructed from run-to-failure bearing data and a technical maintenance manual corpus to evaluate the router and the four processing paths. The hybrid router achieves an accuracy of 0.979, while the DT-only module obtains 0.967 query accuracy and 0.900 joint accuracy. Hybrid + reranker performs best for RAG-only Hit@5, mean reciprocal rank (MRR), and context precision, whereas hybrid performs best for DT+RAG Hit@1, MRR, faithfulness, and context precision. These results demonstrate that route-aware evidence selection provides a transparent and evaluable approach to LLM-based bearing diagnosis QA.

Keywords: bearing diagnosis; digital twin; retrieval-augmented generation; large language model; intelligent maintenance

Received: 28 August 2026
Revised: 14 September 2026
Accepted: 15 September 2026
Published: 17 September 2026

1. Introduction

Rolling bearings are essential components in rotating machinery, and their health directly affects equipment reliability, maintenance cost, and operational safety. Bearing faults can cause abnormal vibration, noise, temperature rise, and eventually severe machine downtime. As a result, bearing fault diagnosis has been widely studied using vibration signal processing, statistical feature extraction, machine learning, and deep learning. Most existing methods focus on fault classification or health-state estimation from sensor data.

However, practical maintenance scenarios often require more than a predicted fault label. Engineers may ask questions such as which bearing is closest to failure, whether

the health index is increasing over time, why high kurtosis indicates impulsive behavior, or what inspection action is reasonable when a bearing enters a severe degradation state. These questions require an interactive question-answering system that can access measured condition evidence, retrieve engineering knowledge, and generate cautious diagnostic explanations.

Large language models provide a natural interface for such diagnostic interaction, but direct LLM generation is not sufficient for engineering diagnosis. Without external grounding, an LLM may produce unsupported explanations or overconfident maintenance suggestions. Retrieval-augmented generation reduces this risk by grounding answers in retrieved documents, but a document-only RAG system cannot answer questions about the current state of a specific bearing. Conversely, a digital twin database can provide measured condition evidence, but numerical features alone do not provide the manual-supported mechanical interpretation that is required for diagnosis and maintenance reasoning.

This motivates the integration of digital twins and retrieval-augmented generation for bearing diagnosis question answering. The objective of this work is not to develop another bearing fault classifier but to investigate how an LLM agent can select and combine heterogeneous evidence for maintenance-oriented diagnostic QA. To this end, the proposed framework combines two complementary knowledge sources: a lightweight queryable database that stores time-dependent bearing condition states and a structured technical-document corpus that provides engineering knowledge. The agent first determines whether a question is within the scope of bearing diagnosis and maintenance. Questions that are unrelated to this scope are rejected without querying either the digital twin database or the technical-document corpus, thereby preventing unsupported answers. For relevant questions, the agent selects one of three evidence-based routes: DT-only question answering for measured condition values and health trends, RAG-only question answering for general engineering knowledge such as bearing operation, mounting, and lubrication, or DT+RAG reasoning when an answer requires both the condition of a specific bearing and its engineering interpretation. In the DT+RAG route, the agent first retrieves the relevant digital twin state and then uses this state to guide technical-document retrieval. This route-specific design matches each question with the evidence it requires and enables the separate evaluation of query routing, structured querying, document retrieval, and answer grounding. The main contributions of this paper are summarized as follows:

- A four-route DT–RAG LLM agent framework is proposed for bearing diagnosis question answering. The framework classifies user queries into unrelated, DT-only, RAG-only, and DT+RAG routes and applies route-specific processing to each query type.
- A 140-question benchmark is constructed to evaluate the four query-handling routes. The benchmark includes unrelated, DT-only, RAG-only, and DT+RAG questions together with route labels and evidence annotations. This benchmark provides a unified basis for comparing route selection, structured condition-state querying, document retrieval quality, and answer grounding across heterogeneous diagnostic questions.
- A route-specific evaluation protocol is established to assess the complete four-route framework. The protocol evaluates routing accuracy, unrelated-query rejection, DT query and answer accuracy, document retrieval quality, and answer grounding while comparing the BM25, dense, hybrid, and hybrid + reranker retrieval methods.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed route-aware DT–RAG agent framework. Section 4 describes the benchmark, experimental settings, and results. Section 5 discusses the limitations and future work, and Section 6 concludes the paper.

2. Related Work

2.1. Digital Twin-Based Bearing Diagnosis

Bearing fault diagnosis has traditionally relied on vibration signal processing, handcrafted features, and machine learning classifiers. Classical studies have emphasized envelope analysis, cyclostationarity, spectral kurtosis, and other signal-processing tools for extracting weak defect signatures from vibration signals [1,2]. Public benchmark studies, such as those based on the Case Western Reserve University bearing data, have also shaped reproducible evaluation practice for rolling-element bearing diagnosis [3]. Recent reviews show that machine learning and deep learning have substantially improved automatic machinery fault diagnosis but also highlight persistent challenges related to data scarcity, domain shift, interpretability, and deployment robustness [4–6]. Spirto et al. combined symmetrized dot pattern indices with a feedforward neural network for rolling bearing fault detection [7], while Jiang et al. developed a convolutional capsule network for rolling bearing fault diagnosis [8]. Digital twins have become an important tool for machinery health monitoring because they can represent equipment states, simulate operating conditions, and support predictive maintenance. For rolling bearings, recent digital twin studies often use physical modeling, simulation data, or domain adaptation to address the scarcity of labeled real fault data. Fan et al. constructed a lifecycle-oriented digital twin model for rolling bearing fault diagnosis using Modelica [9]. Qin et al. developed a faulty rolling bearing digital twin model for imbalanced-sample diagnosis [10], while Ming et al. proposed a digital twin-assisted framework that reduces the distribution discrepancy between dynamic-model responses and measured data under imbalanced conditions [11]. Jiang et al. used digital twin-assisted domain adaptation transfer learning to improve bearing diagnosis under small-sample conditions [12]. You et al. proposed a multi-margin physical-information digital twin framework that combines virtual and real sound-vibration signals for bearing diagnosis without real fault samples [13]. Fang et al. developed a digital twin-enabled domain adaptation network for cross-space roller bearing fault diagnosis between simulated and measured data [14]. Recently, this line of work further expanded toward resonance-aware data augmentation, physics-informed autoencoding, open-set recognition, missing-data repair, and industrial spindle datasets [15–19]. Other recent work has explored zero-shot transfer, few-shot data generation, digital twin-driven feature enhancement, and graph contrastive domain adaptation for data-scarce bearing diagnosis [20–23].

These studies demonstrate the value of digital twins for data augmentation, cross-domain adaptation, and fault classification. In contrast, our digital twin is designed as a lightweight queryable condition-state representation for QA. Rather than using the digital twin primarily to train a classifier, we use it to provide structured evidence for natural-language diagnostic questions and to guide retrieval from an engineering manual.

2.2. LLM-Based Bearing Diagnosis

Recent studies have begun to explore how large language models can support bearing fault diagnosis. Li et al. proposed FD-MVLLM, which integrates time series, time–frequency representations, statistical prompt engineering, and LLM fine-tuning for bearing fault diagnosis [24]. Xiao et al. proposed KG-SR-LLM, which converts vibration signals into structured semantic sequences and uses knowledge-guided prompt tuning for cross-domain bearing diagnosis [25]. Signal LLM further illustrates the use of cross-attention LLMs for fault perception and maintenance planning in turbine bearing systems [26].

These methods show that LLMs can be adapted to vibration-based diagnosis when sensor signals are transformed into semantic or multimodal representations. However, they mainly focus on fault classification, perception, or maintenance planning, whereas our

work focuses on evidence-grounded diagnostic question answering over both measured digital twin states and maintenance-manual knowledge.

2.3. Industrial RAG for Maintenance Question Answering

Retrieval-augmented generation has become a practical approach for reducing hallucination in knowledge-intensive QA by grounding LLM outputs in retrieved documents [27]. Recent RAG studies have further investigated modular retrieval-generation pipelines, adaptive retrieval, self-reflection, and corrective retrieval when the retrieved evidence is incomplete or noisy [28–30]. Recent industrial applications increasingly use RAG to support troubleshooting, maintenance guidance, and technical-document question answering. Lukens et al. proposed an evaluation framework for fault diagnosis using technical manuals in retrieval-augmented LLMs and emphasized that retrieval-first and task-specific evaluation designs are important for technical maintenance settings [31]. Tang et al. developed a digital twin- and LLM-driven intelligent maintenance system for laser-directed energy deposition, including a RAG-based maintenance Q&A assistant and a DAG-based evaluation framework for fault diagnosis [32]. Miao et al. proposed a Graph RAG-based diagnosis framework for train bogies that combines knowledge graphs and LLMs to improve explainability and attribution consistency [33]. Other studies have explored labeling-free RAG-enhanced LLM diagnosis, LLM-driven knowledge graph construction for fault diagnosis, and signal-LLM maintenance systems for turbine bearings [26,34,35]. Recent surveys also highlight RAG, knowledge graphs, and LLM–digital twin integration as emerging directions for machinery health monitoring [36].

Compared with these works, our study focuses specifically on bearing diagnosis QA and evaluates four route-specific reasoning paths. The system does not only retrieve technical manual passages; it first decides whether the question requires DT evidence, manual evidence, both, or neither. For DT+RAG questions, the measured DT condition is explicitly transformed into engineering semantics before manual retrieval, which differentiates our framework from conventional document-only RAG systems.

2.4. Research Gap

The recent literature shows rapid progress in digital twin-driven bearing diagnosis, LLM-based fault classification, and RAG-based industrial maintenance QA. However, three gaps remain. First, most digital twin bearing studies focus on simulation, data augmentation, domain adaptation, or classification rather than interactive diagnostic question answering. Second, LLM-based bearing diagnosis studies generally transform sensor signals into semantic or multimodal inputs for classification, perception, or planning, but they rarely evaluate how an agent should choose among heterogeneous evidence sources before answering an engineering question. Third, industrial RAG studies often ground answers in manuals or knowledge graphs, but they do not explicitly combine measured bearing-state evidence with manual-supported interpretation for the same diagnostic query. To address these gaps, this work proposes a four-route DT–RAG framework for bearing diagnosis question answering. Table 1 summarizes the distinction between representative recent studies and the present work.

Table 1. Comparison with representative recent related work.

Research Direction	Representative Works	Evidence Source	Gap Addressed by This Work
Digital twin bearing diagnosis	Jiang et al.; Fang et al.; Zhang et al.; Ming et al. [12,14,15,17]	Simulated and measured bearing data	Focuses on fault classification, data augmentation, or domain adaptation rather than manual-grounded interactive QA.

Table 1. *Cont.*

Research Direction	Representative Works	Evidence Source	Gap Addressed by This Work
LLM-based bearing diagnosis	FD-MVLLM; KG-SR-LLM; Signal LLM [24–26]	Sensor-derived semantic or multimodal inputs	Uses LLMs mainly for fault labels, perception, or maintenance planning, not route-aware evidence-grounded QA.
Industrial RAG diagnosis	Lukens et al.; Miao et al.; Xu et al.; Tang et al. [31–34]	Manuals, logs, or knowledge graphs	Provides manual or graph-based reasoning but does not explicitly fuse bearing DT features with technical-document evidence.
This work	Proposed DT–RAG framework	Structured condition-state database and technical-document corpus	Evaluates routing, DT querying, retrieval, and answer grounding for bearing diagnosis QA.

3. A Route-Aware DT–RAG Bearing Diagnosis Question-Answering Framework

The proposed framework is designed as a route-aware diagnostic question-answering system for bearing condition analysis. Its goal is to answer natural-language questions by selecting the appropriate evidence source and reasoning path rather than forcing all the questions into a single retrieval or database query pipeline. The overall architecture consists of four main components: a question router, a digital twin query module, a technical-document retrieval module, and an evidence-grounded answer generation module.

Given a user question, the system first determines whether the question is relevant to the bearing diagnosis task and, if so, which type of evidence is required. Questions about measured bearing states, extracted vibration features, degradation trends, or comparisons among bearings are routed to the digital twin module. Questions about bearing operation, mounting, lubrication, inspection, damage mechanisms, or troubleshooting knowledge are routed to the technical-document retrieval module. Questions that require both measured condition evidence and engineering interpretation are routed to the DT+RAG module, where digital twin evidence is first retrieved and then used to guide document retrieval. Questions outside the bearing diagnosis domain are rejected or redirected as unrelated inputs. Figure 1 illustrates the overall workflow, including the user question, route selection, DT-only path, RAG-only path, DT+RAG path, and final evidence-grounded answer.

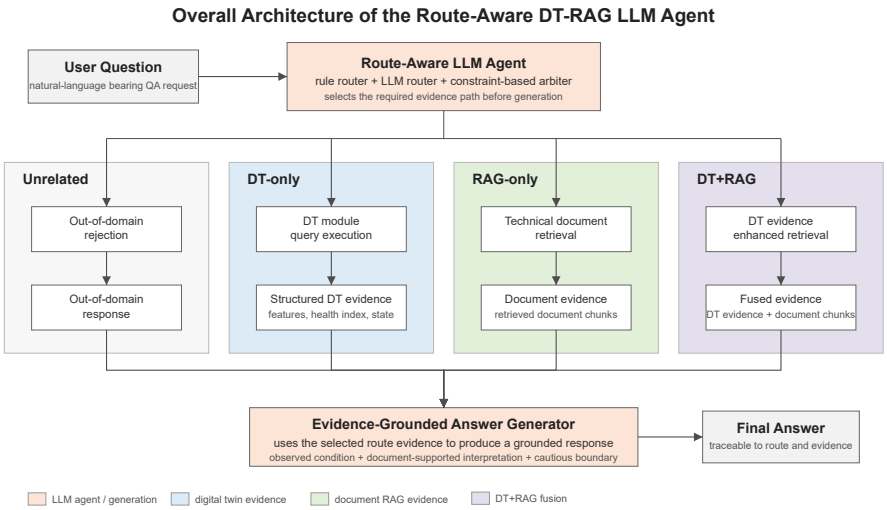

Figure 1. Overall architecture of the proposed route-aware DT–RAG bearing diagnosis question-answering framework.

3.1. Question Routing

Question routing determines whether an input is within the scope of the system and identifies the processing path required to answer it. The proposed framework classifies each input into four routes: unrelated, DT-only, RAG-only, and DT+RAG. Unrelated questions are rejected without invoking the evidence modules. DT-only questions use the digital twin query module, RAG-only questions use the technical-document retrieval module, and DT+RAG questions invoke both modules sequentially, with the digital twin query performed first.

The *unrelated* route is used for questions that are not related to bearing diagnosis or technical maintenance. Such questions may involve general knowledge, translation, travel, weather, or other topics outside the scope of the system. In this route, the system does not attempt to retrieve bearing evidence and instead returns an out-of-domain response.

The *DT-only* route is used when the question can be answered by structured digital twin evidence alone. Typical examples include querying the latest RMS value of a bearing, comparing health indices across bearings, identifying the bearing closest to failure, or describing the degradation trend of a feature. These questions require accurate mapping from natural language to database query parameters.

The *RAG-only* route is used when the question asks for technical knowledge contained in the document corpus but does not require current bearing state information. Examples include questions about mounting procedures, lubrication principles, inspection recommendations, bearing damage causes, or troubleshooting guidelines. In this route, the system retrieves relevant document chunks and generates an answer grounded in the retrieved context.

The *DT+RAG* route is used when the question requires both measured evidence and manual-supported interpretation. For example, a user may ask what a high kurtosis value means for a specific bearing, what fault type may be associated with abnormal high-frequency vibration energy, or what maintenance action should be considered when a bearing is close to failure. In this case, the system first obtains DT evidence and then uses the engineering semantics of that evidence to retrieve relevant technical-document knowledge.

Three router variants are considered: a rule-based router using domain cues, an LLM router using question semantics, and a hybrid router that combines both. The hybrid router accepts an agreed route directly and resolves disagreements with domain constraints involving bearing entities, measurement intent, document knowledge, and diagnostic or maintenance intent. More specifically, the rule-based router extracts bearing identifiers, digital twin feature terms, document terms, explanation or maintenance terms, and unrelated-domain terms. DT signals include condition and feature names such as `health_index`, RMS, kurtosis, crest factor, spectral entropy, high-band energy ratio, degradation, trend, and anomaly. Document signals include mounting, lubrication, inspection, troubleshooting, misalignment, bearing damage, load, life, selection, seals, and related technical concepts. Explanation signals include terms such as “why”, “explain”, “interpret”, “diagnose”, and “how should”. The LLM router is prompted to return strict JSON with the route, confidence, reason, DT entities, and RAG query; temperature is set to 0. For the hybrid router, route conflicts are resolved as follows: if no condition or document signal is present and unrelated signals are detected, the final route is `unrelated`; if a condition entity appears together with explanation or maintenance intent, the final route is `dt_rag`; if a condition entity appears with measurement, state, or trend intent but without explanation intent, the final route is `dt_only`; if no condition entity is present but document intent is detected, the final route is `rag_only`. If none of these constraints applies, the rule-based result is used as the fallback.

The corresponding processing workflows for the unrelated, DT-only, and RAG-only routes are summarized in Figure 2 and the DT+RAG workflow is described separately in Figure 3.

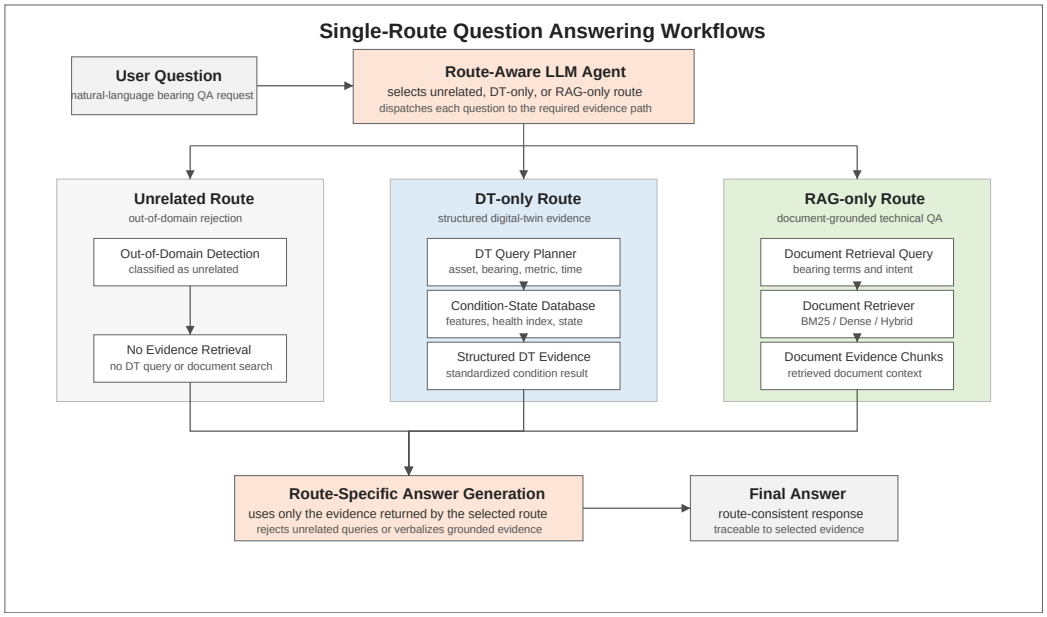

Figure 2. Single-pass question-answering workflows for the unrelated, DT-only, and RAG-only routes.

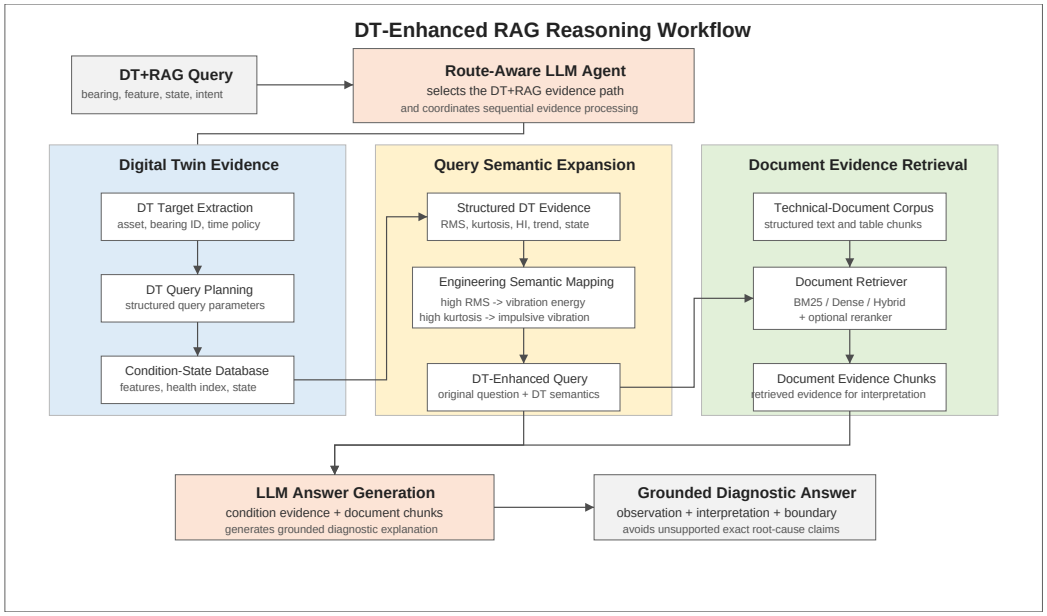

Figure 3. Single-pass DT-enhanced RAG reasoning workflow for combining measured bearing condition evidence with technical-document knowledge.

3.2. Digital Twin Construction and Querying

The digital twin (DT) module provides a lightweight queryable representation of time-dependent bearing conditions for the diagnosis QA agent. It stores asset and bearing identities, temporal condition records, extracted vibration features, health indices, health states, and evidence summaries. The proposed DT is intended for structured condition-state querying rather than full industrial digital twin functions such as online synchronization, physics-based simulation, or closed-loop control.

The DT evidence source accepts time-ordered multi-channel vibration records together with an asset-specific mapping from sensor channels to bearings. Each timestamped record

is treated as one condition snapshot. Feature extraction is performed independently for each channel, and channel-level records associated with the same bearing and timestamp are then aggregated into bearing-level snapshots. For each channel-level vibration segment, the framework extracts standard time-domain and frequency-domain features that are commonly used in bearing condition monitoring [1,2]. The time-domain features include mean, standard deviation, RMS, absolute mean, peak amplitude, peak-to-peak value, skewness, kurtosis, crest factor, shape factor, impulse factor, clearance factor, signal energy, and zero-crossing rate. Frequency-domain features are derived from the one-sided Fourier spectrum and include dominant frequency, spectral centroid, normalized spectral entropy, and low-, middle-, and high-band energy ratios.

Among the extracted features, five degradation-sensitive indicators, namely RMS, kurtosis, crest factor, high-band energy ratio, and spectral entropy, are selected to construct a heuristic health index (HI). For each asset, bearing, and channel, the first 10% of records, with a minimum of five records, are used as the baseline. The one-sided normalized deviation of feature f at sequence index t is defined as

$$z_f(t) = \max\left(\frac{x_f(t) - \mu_f^{base}}{\sigma_f^{base}}, 0\right), \tag{1}$$

where μ_f^{base} and σ_f^{base} are the baseline mean and standard deviation, and the normalized term follows the standard z-score form [37]. The $\max(\cdot, 0)$ operation retains only positive deviations because the selected degradation-sensitive features are interpreted mainly through increases relative to the baseline. If the baseline standard deviation is zero, a small numerical floor is applied.

To provide a compact representation of bearing condition, the selected features are fused into a single HI [38–40]. The health index is then calculated as

$$\mathrm{HI}(t) = \mathrm{clip}\left(\frac{\sum_{f \in F} w_f z_f(t)}{8.0}, 0, 1\right), \tag{2}$$

where F denotes the five selected features, and $\mathrm{clip}(u, 0, 1) = \min(\max(u, 0), 1)$ bounds the HI to $[0, 1]$. The corresponding weights are 0.35, 0.25, 0.15, 0.15, and 0.10. They are heuristic settings that give greater emphasis to overall vibration severity and impulsive behavior while retaining complementary peak-related and spectral evidence. The constant 8.0 is a fixed score scale that maps the weighted deviation to the unit interval after clipping. The resulting HI is mapped to discrete health states using the heuristic thresholds in Table 2.

Table 2. Heuristic health index thresholds used in the condition-state database.

Health Index Interval	Health State
$\mathrm{HI} < 0.20$	`normal`
$0.20 \leq \mathrm{HI} < 0.40$	`early_degradation`
$0.40 \leq \mathrm{HI} < 0.70$	`degradation`
$0.70 \leq \mathrm{HI} < 0.90$	`severe_degradation`
$0.90 \leq \mathrm{HI} \leq 1.00$	`failure_near`

The condition data are organized into three database tables, as summarized in Table 3. The `experiments` table stores experiment metadata, the `feature_records` table stores channel-level features and health indicators, and the `bearing_snapshots` table stores bearing-level condition summaries aggregated from channel-level records. This structure supports feature lookup, health-state queries, cross-bearing comparison, abnormality ranking, degradation trend analysis, and degradation-start detection.

Table 3. Digital twin database structure.

Table	Main Fields	Purpose
`experiments`	`experiment`, `channels`, `failure_description`	Stores experiment metadata and known run-to-failure descriptions.
`feature_records`	`experiment`, `bearing_id`, `channel`, `timestamp`, `sequence_index`, vibration features, baseline means/stds, z-scores, `health_index`, `health_state`, `evidence`	Stores channel-level feature extraction and health index calculation results.
`bearing_snapshots`	`experiment`, `bearing_id`, `timestamp`, `sequence_index`, aggregated vibration features, `health_index`, `health_state`, `evidence`	Stores bearing-level condition snapshots used by the DT query tools.

For DT-related questions, an LLM-based planner first generates an initial structured query containing the action, experiment, bearing, target feature or metric, scope, and temporal information. Deterministic normalization then standardizes the query fields and completes schema-dependent parameters before the query executor performs the corresponding SQL operation and returns structured DT evidence for answer generation. This separation enables independent evaluation of query planning, query normalization, database execution, and answer generation.

3.3. Technical-Document Corpus Construction and Querying

The technical-document corpus provides the unstructured engineering knowledge used by the RAG-only and DT+RAG routes. Since technical manuals contain a mixture of paragraphs, headings, figures, captions, tables, and enumerated procedures, direct page-level retrieval would produce contexts that are either too coarse or insufficiently aligned with question intent. Therefore, the source document is first parsed and converted into a structured intermediate representation before retrieval indexing.

The source document is parsed using LlamaParse, and the parsed output is cleaned and stored as structured Markdown JSON. The parsed content is segmented according to chapter titles, section headings, subsection hierarchy, table boundaries, and semantic continuity. The chunking process preserves both textual and tabular information. Textual passages are grouped into semantic text units and then merged when adjacent units belong to the same heading path and remain within the maximum length constraint. Tables are retained as independent chunks when appropriate because many maintenance-oriented answers depend on tabular procedures, tolerance values, troubleshooting symptoms, or cause–action mappings. Each chunk is assigned a unique chunk ID and metadata, including chapter name, chapter number, section, subsection, heading path, page range, content type, and source file. The metadata-aware corpus design allows the retriever to return not only raw text but also contextual information about where the evidence appears in the source document. Such metadata is useful for answer grounding, evidence inspection, and later citation-level analysis.

Four retrieval strategies are considered for technical-document retrieval. BM25 is used as a sparse lexical baseline and is effective when the question contains explicit technical terms that also appear in the source. Dense retrieval uses `bge-m3` embeddings to capture semantic similarity beyond exact word overlap [41]. Hybrid retrieval combines lexical and dense candidates to balance exact term matching and semantic recall. Finally, hybrid + reranker applies `bge-reranker-v2-m3` to reorder the hybrid candidate set before selecting the final contexts. In all the RAG-only configurations, the final number of chunks provided to the answer model is fixed, so performance differences mainly reflect retrieval and ranking quality rather than context-budget differences.

3.4. DT-Enhanced RAG Reasoning

The DT+RAG route is designed for questions that cannot be answered reliably by either the digital twin or the technical-document corpus alone. Digital twin evidence provides measured condition information, but numerical indicators such as RMS, kurtosis, or high-band energy ratio do not by themselves provide a complete mechanical explanation. Conversely, technical documents provide general engineering knowledge, but they do not contain the current condition state of a specific bearing. The DT-enhanced RAG route therefore uses the digital twin to ground the question in measured evidence and then uses the technical-document corpus to support interpretation.

The reasoning process consists of five steps, as illustrated in Figure 3. *First*, the system identifies the DT query target from the user question, including the experiment, bearing ID, relevant feature or state, and intended diagnostic task. *Second*, the DT planner generates structured query parameters and retrieves the corresponding condition evidence from the database. The returned DT evidence may include the latest health state, health index, abnormal feature values, timestamp, sequence index, and a short structured evidence summary. *Third*, the system converts the numerical DT evidence into engineering semantics. For example, high RMS is interpreted as increased vibration amplitude or vibration energy, high kurtosis as impulsive vibration, high crest factor as peak impact vibration, high-band energy ratio as high-frequency vibration energy, and a `failure_near` state as severe degradation. This conversion does not assert a precise root cause; rather, it reformulates numerical observations into retrieval-oriented diagnostic concepts. *Fourth*, the DT-derived semantics are combined with the original question to construct a DT-enhanced retrieval query. This query is used to retrieve document chunks that discuss relevant inspection principles, vibration monitoring, bearing damage mechanisms, troubleshooting clues, or maintenance implications. Depending on the retrieval configuration, the system may use BM25, dense retrieval, hybrid retrieval, or reranking-enhanced hybrid retrieval to obtain the final evidence contexts. *Finally*, the answer generator receives both the structured DT context and the retrieved document chunks. The generated answer is required to include three types of information: the observed DT condition, the document-supported mechanical interpretation, and a cautious diagnostic boundary. This boundary is important because vibration features alone may indicate abnormal behavior but may not be sufficient to identify an exact defect location, damage mechanism, or corrective action without additional inspection. By explicitly combining measured DT evidence and document evidence, the DT+RAG route aims to produce answers that are both condition-specific and technically grounded.

3.5. Benchmark Construction

A route-specific benchmark is constructed to evaluate both the router and the downstream processing capabilities of the proposed framework. The benchmark contains four query types: unrelated, DT-only, RAG-only, and DT+RAG. The router evaluation uses these four types to determine whether the system can identify the evidence requirement of each question. The downstream benchmark further evaluates out-of-domain rejection, digital twin querying, technical-document retrieval, and diagnostic reasoning based on combined digital twin and manual evidence. All four types of benchmark questions were generated by an LLM and subsequently checked by human reviewers to ensure that each question matched its intended query type and evidence requirement.

For router evaluation, each question is manually given a reference route label corresponding to one of the four classes: `unrelated`, `DT-only`, `RAG-only`, or `DT+RAG`. These reference labels are used to compare the router's predicted route with the intended route.

Human reviewers check whether each question is consistent with its assigned route and whether the route reflects the evidence required to answer the question.

For DT-only questions, reference queries are defined using manually designed templates based on the question semantics and the digital twin database schema. The templates specify the required query parameters, including the action, experiment, bearing, target feature or metric, scope, order, and temporal settings. Each reference query is executed programmatically against the database to obtain the standard digital twin result and the corresponding reference answer. Human reviewers then check the consistency between the question, the reference query, the database result, and the reference answer.

For RAG-only questions, the question and each candidate passage from the technical maintenance manual corpus are provided to an LLM for relevance annotation. Each passage is assigned a relevance score from 0 to 3, where a score of 0 indicates irrelevant evidence and a positive score indicates candidate evidence. The positively scored passages are then provided with the question to a second LLM selection step, which identifies one primary gold evidence passage. A reference answer is generated from the selected passage. The relevance annotations, selected evidence, and reference answers are subsequently checked by human reviewers.

For DT+RAG questions, the digital twin part is defined using the same manually designed query templates as the DT-only questions. The reference query is executed against the database to obtain the corresponding DT evidence. This evidence is then converted into engineering semantics and combined with the original question to form a DT-enhanced retrieval query. The question, DT evidence, enhanced query, and candidate passages from the technical maintenance manual corpus are provided to an LLM for relevance annotation and gold-evidence selection using the same procedure as for the RAG-only questions. The final reference answer combines the observed DT condition with the manual-supported engineering interpretation and an appropriate diagnostic boundary. These DT queries, evidence annotations, selected passages, and reference answers are also checked by human reviewers.

The resulting benchmark provides reference route labels, structured DT queries, standard database results, evidence annotations, selected gold passages, and reference answers. These reference materials are used only for evaluation and are not provided to the system during inference. This design enables the router and each of the four query-handling routes to be evaluated under their intended evidence requirements while keeping route selection, DT query generation, database execution, document retrieval, and answer grounding analytically separate.

4. Experimental Results and Analysis

4.1. Evaluation Benchmark

This work evaluates the proposed framework using a benchmark that was constructed specifically for this work. The evaluation benchmark contains 140 English questions and is designed to cover the four reasoning paths of the proposed system. The benchmark uses the IMS (Intelligent Maintenance Systems) run-to-failure bearing vibration dataset [42] to build the digital twin evidence source and the SKF (Svenska Kullagerfabriken) bearing maintenance handbook [43] to build the technical-document retrieval source. Each question is assigned an expected route: unrelated, DT-only, RAG-only, or DT+RAG. The benchmark is used to evaluate whether the system can reject out-of-domain questions, query structured condition evidence, retrieve technical-document evidence, and combine both sources for diagnostic explanation. The benchmark construction and annotation procedures are described in Section 3.5. The distribution of the evaluation benchmark is shown in Table 4, and representative examples from the four routes are given in Table 5. The bench-

mark questions and reference annotations are constructed separately with the assistance of `GPT-4o-mini`.

Table 4. Distribution of the evaluation dataset.

Question Type	Number	Description
Unrelated	30	Out-of-domain questions
DT-only	30	Digital twin database questions
RAG-only	50	SKF manual knowledge questions
DT+RAG	30	DT evidence plus manual reasoning
Total	140	Complete agent benchmark

Table 5. Representative benchmark question examples.

Route	Example Question	Expected Evidence
Unrelated	What is the capital of France?	None; out-of-domain rejection.
DT-only	What is the latest health state of bearing_3 in 1st_test?	Latest available DT condition snapshot for the specified bearing.
RAG-only	What are common SKF manual recommendations for bearing lubrication?	SKF manual chunk.
DT+RAG	What does the abnormal kurtosis of bearing_3 in 1st_test imply, and what inspection guidance is relevant?	DT feature evidence plus SKF manual evidence.

4.2. Evaluation Metrics

The evaluation metrics are defined for four aspects of the proposed framework: route classification, digital twin querying, technical-document retrieval, and answer grounding. Let N denote the number of questions in the evaluated subset and $\mathbb{I}(\cdot)$ be the indicator function.

Router performance is measured by overall accuracy and per-route accuracy. For question i, let y_i be the expected route and $\hat{y}_i$ the predicted route. The overall routing accuracy is

$$\mathrm{Acc}_{\mathrm{route}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i). \tag{3}$$

For a specific route class c, the per-route accuracy is

$$\mathrm{Acc}_c = \frac{1}{N_c}\sum_{i:y_i=c}\mathbb{I}(\hat{y}_i = c), \tag{4}$$

where N_c is the number of benchmark questions whose expected route is c.

DT-only performance is evaluated using four metrics: query accuracy, execution accuracy, answer accuracy, and joint accuracy. Query accuracy measures whether the normalized planned DT query matches the reference query in the required fields, including action, experiment, bearing, feature or metric, scope, order, state, and window parameters when applicable. Execution accuracy measures whether the planned query executes successfully in the database and returns a structured DT result. Answer accuracy measures whether the final natural-language answer contains the required key facts, including the correct experiment, bearing, health state, sequence index, trend direction, and numerical values within the predefined tolerance. Let q_i, e_i, and a_i denote the binary query, execution, and answer correctness indicators for question i. Then

$$\mathrm{Acc}_{\mathrm{query}} = \frac{1}{N}\sum_{i=1}^{N}q_i, \quad \mathrm{Acc}_{\mathrm{exec}} = \frac{1}{N}\sum_{i=1}^{N}e_i, \quad \mathrm{Acc}_{\mathrm{answer}} = \frac{1}{N}\sum_{i=1}^{N}a_i. \tag{5}$$

The joint accuracy requires all three stages to be correct:

$$\mathrm{Acc}_{\mathrm{joint}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(q_i = 1 \wedge e_i = 1 \wedge a_i = 1). \tag{6}$$

RAG-only and DT+RAG performance are evaluated from retrieval and grounding perspectives. Each question has one reviewed gold SKF chunk g_i. Given the ranked retrieved list $R_i = [r_{i,1}, \ldots, r_{i,K}]$, the gold rank is

$$\mathrm{rank}_i = \min\{j : r_{i,j} = g_i\}. \tag{7}$$

If the gold chunk is not retrieved, rank_i is undefined and the retrieval score for that question is 0. Hit@k and mean reciprocal rank (MRR), which are standard information-retrieval metrics [44], are computed as

$$\mathrm{Hit@}k = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\mathrm{rank}_i \le k), \tag{8}$$

$$\mathrm{MRR} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 1/\mathrm{rank}_i, & \text{if } g_i \in R_i, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

For RAG-only and DT+RAG questions, answer quality is evaluated using two RAGAS metrics: faithfulness and answer relevancy [45]. Faithfulness measures whether the claims in a generated answer are supported by the retrieved contexts. Answer relevancy measures whether the answer directly addresses the question. Context precision is used as an additional retrieval-context metric and measures whether relevant information is ranked prominently among the retrieved contexts. The reported values are macro-averages over all evaluated questions.

4.3. Experimental Setting

The experiments use the IMS run-to-failure bearing vibration dataset [42] to construct the digital twin evidence source and the SKF (Svenska Kullagerfabriken) bearing maintenance handbook [43] to construct the technical-document corpus. The IMS dataset contains three run-to-failure experiments. The first experiment includes four bearings monitored by eight vibration channels, with two channels assigned to each bearing; the second and third experiments each include four bearings monitored by one channel per bearing. Figure 4 presents the bearing-level trajectory of bearing 3 from the first experiment. The SKF corpus covers Chapters 1–11 and is parsed into structured text and table units. The resulting content is segmented according to chapter titles, section headings, table boundaries, and semantic continuity. Each chunk is assigned a unique identifier and metadata, including its chapter, section, heading path, page range, and content type. The final corpus contains 608 chunks.

The digital twin database is constructed from the IMS vibration records. Time- and frequency-domain features are extracted, and degradation-sensitive features are used to calculate the health index and health state for each bearing record. The resulting feature values, temporal condition records, health indices, health states, and evidence summaries are stored in a database. No fault-diagnosis or prognostic model is trained or fine-tuned on the IMS records; instead, they are processed and stored as queryable DT condition evidence.

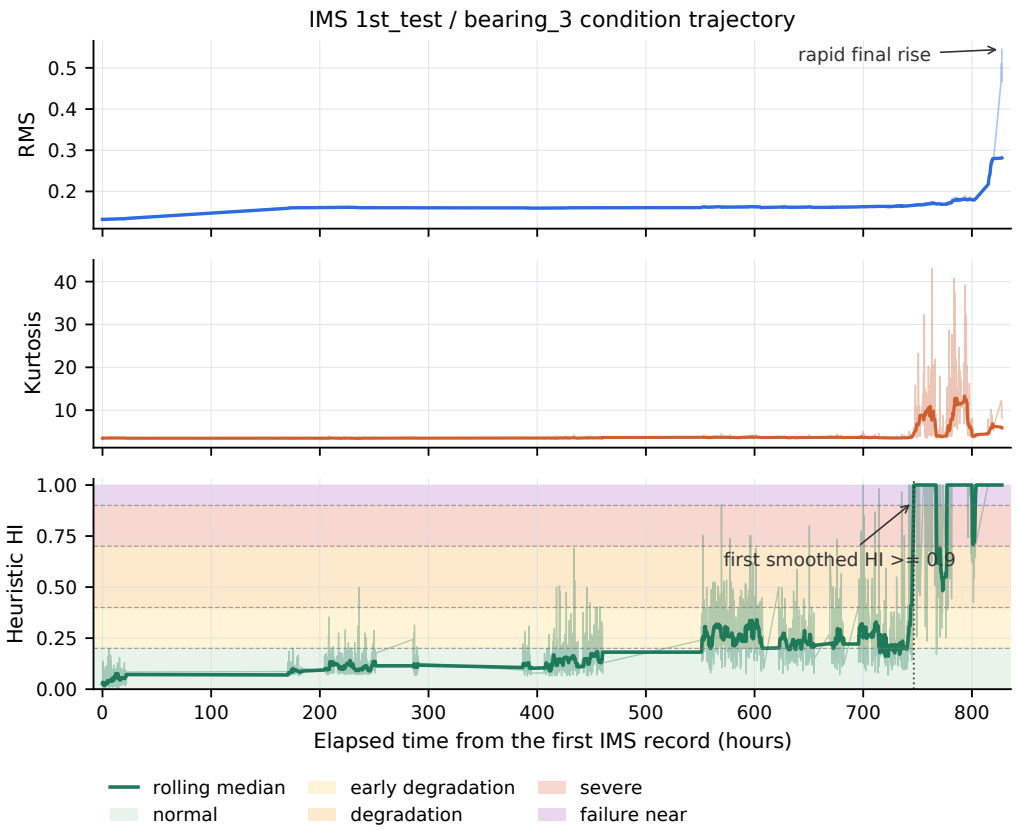

Figure 4. Representative feature and heuristic health index trajectory for bearing 3 in the IMS 1st test. The panels show RMS, kurtosis and the health index over the run-to-failure sequence.

As an illustration of the resulting condition evidence, Figure 4 shows a representative trajectory from the IMS instantiation used in the experiments. The RMS remains relatively stable through most of the sequence and rises rapidly near the final stage, while kurtosis exhibits intermittent impulsive peaks during the later degradation period. The corresponding heuristic HI increases from the normal region toward the severe-degradation and failure-near regions, although short-term fluctuations remain visible. This example shows how the extracted features are converted into a time-dependent condition trajectory that can be queried by the DT-only route and interpreted by the DT+RAG route.

The experiments follow the four predefined routes: unrelated, DT-only, RAG-only, and DT+RAG. The router is evaluated on the complete benchmark, whereas the expected route is fixed for each route-specific downstream experiment. DT-only questions are processed using the digital twin query module, RAG-only questions using the technical-document retrieval module, and DT+RAG questions using the digital twin query module followed by technical-document retrieval.

For technical-document retrieval, BM25, dense, hybrid, and hybrid + reranker are compared. Dense retrieval uses `bge-m3` embeddings, and reranking uses `bge-reranker-v2-m3`. Hybrid retrieval combines lexical and dense candidates using reciprocal rank fusion with $k = 60$. For hybrid and hybrid + reranker, the initial candidate set contains the top 20 BM25 and dense retrieval results. The final number of contexts provided to the answer model is fixed to `top-k = 5`, and the reranker returns the top five contexts.

For DT-only and DT+RAG questions, the DT planner uses Llama 3.3 70B to generate structured query parameters before SQL execution. The final answer-generation model is also Llama 3.3 70B. The main implementation settings are summarized in Table 6.

Table 6. Main implementation settings used in the experiments.

Component	Setting
Answer model	Llama 3.3 70B using the local identifier `llama3.3:70b`; final answer temperature 0.2.
Router models	Rule router, LLM router, and hybrid router; LLM/hybrid router model set to Llama 3.3 70B in the full 140-question router experiment.
DT planner	LLM-based planner with structured JSON query output; DT+RAG planner model set to Llama 3.3 70B.
Embedding model	`bge-m3`, 1024-dimensional embeddings, served through an HTTP embedding service.
Reranker	`bge-reranker-v2-m3`; final reranked context count controlled by `reranker-top-k`.
Judge model	`gpt-4o-mini` for RAGAS-style grounding-related answer indicators in the reported RAG-only and DT+RAG experiments.
RAG corpus	608 SKF manual chunks from Chapters 1–11.

4.4. Experimental Results

4.4.1. Router Performance

The router is evaluated on the full 140-question benchmark, including 30 unrelated questions, 30 DT-only questions, 50 RAG-only questions, and 30 DT+RAG questions. Table 7 reports the overall routing accuracy defined in Equation (3) and the per-route accuracies defined in Equation (4). The hybrid router achieves the best overall accuracy of 0.979. The rule router performs strongly on unrelated and DT-related questions but is less reliable for RAG-only manual-knowledge questions. The LLM router correctly identifies all the unrelated, RAG-only, and DT+RAG questions, but it makes several errors on DT-only questions. The hybrid strategy combines these two behaviors and provides the most stable route selection.

Table 7. Router evaluation results on the full 140-question benchmark.

Router	Overall	Unrelated	DT-Only	RAG-Only	DT+RAG
Rule	0.929	1.000	0.933	0.880	0.933
LLM	0.971	1.000	0.867	1.000	1.000
Hybrid	0.979	1.000	0.933	1.000	0.967

4.4.2. DT-Only Question Answering

The DT-only results are shown in Table 8. The overall query accuracy reaches 0.967, and execution accuracy reaches 1.000. This indicates that, once the query parameters are correctly generated, the SQL execution layer reliably returns the expected DT evidence. The overall answer accuracy is 0.933, while the joint accuracy is 0.900.

Table 8. DT-only question-answering results.

Category	Num.	Query	Exec.	Answer	Joint
Feature	10	1.000	1.000	0.900	0.900
State	10	0.900	1.000	1.000	0.900
Trend	10	1.000	1.000	0.900	0.900
Overall	30	0.967	1.000	0.933	0.900

Note: Num. denotes the number of questions. Query, Exec., and Answer denote query accuracy, execution accuracy, and answer accuracy, respectively. Joint denotes joint accuracy.

The feature and trend categories both obtain perfect query accuracy but slightly lower answer accuracy, suggesting that most errors occur during final response formulation rather than database access. The state category obtains perfect answer accuracy but one query-parameter error. Overall, the results show that the planner–executor–answer separation is useful for identifying the source of DT-only errors.

4.4.3. RAG-Only Question Answering

Table 9 reports the RAG-only results on 50 SKF manual questions. Hybrid + reranker achieves the highest Hit@5, MRR, and context precision, while dense retrieval obtains the highest faithfulness and answer relevancy indicators. Hybrid and hybrid + reranker obtain the same Hit@1.

Table 9. RAG-only question-answering results.

Method	H@1	H@5	MRR	Faith.	Relev.	Prec.
BM25	0.540	0.740	0.627	0.892	0.902	0.888
Dense	0.520	0.840	0.639	0.950	0.948	0.915
Hybrid	0.600	0.880	0.706	0.915	0.931	0.924
Hybrid + Reranker	0.600	0.920	0.718	0.926	0.939	0.966

Note: H@1 and H@5 denote Hit@1 and Hit@5, respectively. MRR denotes mean reciprocal rank. Faith., Relev., and Prec. denote faithfulness, answer relevancy, and context precision, respectively.

BM25 obtains competitive Hit@1 because many SKF manual questions contain explicit technical terms that match manual headings or chunk text. Dense retrieval obtains the highest faithfulness and answer relevancy indicators, suggesting that semantic matching can provide contexts that support more relevant generated responses even when the reviewed gold chunk is not always ranked first. Hybrid retrieval achieves the best Hit@1 among non-reranked methods and improves MRR by combining lexical and semantic signals. Hybrid + reranker maintains the same Hit@1 as hybrid, further increases Hit@5 and MRR, and obtains the highest context precision, showing that reranking is useful for ordering SKF manual evidence in ordinary RAG-only QA. These grounding-related indicators should be interpreted together with the retrieval metrics rather than as direct proof of engineering answer correctness.

The remaining RAG-only retrieval misses mainly occur when several neighboring SKF chunks describe closely related procedures. For example, questions about radial shaft seal installation may retrieve chunks on installing a seal over a shaft or into a housing while the reviewed gold chunk is a preparation-oriented section. Such cases are not necessarily irrelevant retrievals; rather, they reveal that manual QA over procedural documents may require evidence sets spanning adjacent sections instead of a single gold chunk.

4.4.4. DT+RAG Question Answering

The DT+RAG results are shown in Table 10. This setting is more challenging than RAG-only QA because the system must first obtain DT evidence and then use it to retrieve document knowledge for question answering. Since the purpose of this experiment is to compare retrieval strategies under the DT-enhanced query condition, route selection is not included in this table and is evaluated separately in the router experiment. Hybrid retrieval achieves the highest Hit@1, MRR, faithfulness, and context precision among the evaluated DT+RAG retrieval strategies, while its Hit@5 is tied with dense retrieval.

Table 10. DT+RAG diagnostic reasoning results.

Method	H@1	H@5	MRR	Faith.	Relev.	Prec.
BM25	0.567	0.600	0.573	0.771	0.701	0.952
Dense	0.367	0.800	0.543	0.825	0.717	0.981
Hybrid	0.600	0.800	0.648	0.851	0.692	0.983
Hybrid + Reranker	0.467	0.767	0.591	0.807	0.684	0.971

Note: H@1 and H@5 denote Hit@1 and Hit@5, respectively. MRR denotes mean reciprocal rank. Faith., Relev., and Prec. denote faithfulness, answer relevancy, and context precision, respectively. The values are macro-averages over the evaluated DT+RAG questions.

BM25 achieves a relatively strong Hit@1 when the DT-enhanced query contains explicit engineering keywords that match the SKF manual. Dense retrieval obtains higher Hit@5 than BM25, indicating that semantic retrieval is useful for recalling related evidence even when the top-ranked chunk is not always the gold reference. Hybrid retrieval performs best in this controlled DT+RAG comparison because it combines exact term matching and semantic similarity, which is appropriate for DT-enhanced queries containing both feature names and engineering interpretations. The hybrid + reranker setting does not outperform hybrid in the DT+RAG evaluation, suggesting that the reranker may reorder some useful hybrid candidates away from the reviewed offline gold chunk when the reranking query is more intent-focused.

Answer relevancy in DT+RAG is lower than in RAG-only QA. This is expected because DT+RAG questions require the model to integrate numerical condition evidence, engineering semantics, and manual knowledge while avoiding unsupported claims about exact fault locations or maintenance actions. The high context precision values across DT+RAG settings indicate that the retrieved contexts are generally relevant, but the final answer must still balance specificity and diagnostic caution. These metrics assess grounding behavior and relevance under retrieved contexts; expert validation would still be required to establish complete maintenance-answer correctness.

5. Discussion and Limitations

The proposed four-route DT–RAG framework provides a unified architecture for heterogeneous bearing diagnosis questions by selecting unrelated, DT-only, RAG-only, or DT+RAG processing paths. However, the current framework mainly supports single-turn question answering and limited sequential tool use, so it does not yet provide a fully agentic reasoning process. Future work will focus on improving the RAG architecture through follow-up questioning, uncertainty-aware clarification, iterative evidence verification, and maintenance-history memory. Structured representations, including knowledge graphs and GraphRAG, can further connect digital twin states, engineering concepts, fault mechanisms, and maintenance actions to support more explicit multi-step reasoning and evidence tracing.

The benchmark provides a unified basis for evaluating the router and the four query-handling routes, including DT querying, technical-document retrieval, and evidence-grounded answer generation. However, because it is grounded in a single IMS dataset covering one bearing type and controlled operating conditions, as well as one SKF maintenance handbook, its scale and coverage of question formulations, bearing conditions, maintenance scenarios, and valid evidence combinations remain limited. Future benchmark development will focus on expanding linguistic and diagnostic diversity, adding implicit and ambiguous questions, including multiple valid evidence passages, covering broader bearing conditions and maintenance tasks, and strengthening expert review and inter-reviewer agreement.

6. Conclusions

This paper presents a four-route DT–RAG framework for bearing diagnosis question answering. The framework routes questions to unrelated, DT-only, RAG-only, and DT+RAG paths, enabling route-specific use of structured digital twin evidence and technical-document knowledge. A route-specific benchmark is developed to evaluate the router and the four processing paths. It provides a unified basis for assessing route selection, DT querying, document retrieval, and answer grounding across heterogeneous diagnostic questions. The experimental evaluation demonstrates the effectiveness of the proposed framework. The hybrid router achieves strong route-selection performance, the DT module

provides reliable structured condition evidence, and different retrieval strategies show distinct advantages for RAG-only and DT+RAG questions.

Future work will improve the framework through follow-up questioning, uncertainty-aware clarification, iterative evidence verification, and maintenance-history memory. Knowledge graphs and GraphRAG will be investigated to support multi-step reasoning and more explicit connections among DT condition states, engineering concepts, fault mechanisms, and maintenance actions. The benchmark will be extended with more diverse and ambiguous diagnostic questions, multiple valid evidence passages, broader bearing conditions and maintenance tasks, and strengthened expert review.

Author Contributions: Conceptualization, H.H.; methodology, H.H.; software, B.Z.; data curation, B.Z.; writing—original draft preparation, B.Z.; writing—review and editing, H.H. and Z.C.; supervision, H.H. and Z.C.; funding acquisition, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Pioneer and Leading Goose Research and Development Program of Zhejiang Province [Grant No. 2025C01088].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code are available at: https://github.com/kurisuu6/bearing-dt-rag-llm (accessed on 14 September 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Randall, R.B.; Antoni, J. Rolling element bearing diagnostics—A tutorial. *Mech. Syst. Signal Process.* **2011**, *25*, 485–520. [CrossRef]
2. Rai, A.; Upadhyay, S.H. A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings. *Tribol. Int.* **2016**, *96*, 289–306. [CrossRef]
3. Smith, W.A.; Randall, R.B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64–65*, 100–131. [CrossRef]
4. Jia, F.; Lei, Y.; Lin, J.; Zhou, X.; Lu, N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.* **2016**, *72–73*, 303–315. [CrossRef]
5. Cerrada, M.; Sánchez, R.V.; Li, C.; Pacheco, F.; Cabrera, D.; de Oliveira, J.V.; Vásquez, R.E. A review on data-driven fault severity assessment in rolling bearings. *Mech. Syst. Signal Process.* **2018**, *99*, 169–196. [CrossRef]
6. Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A.K. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, 106587. [CrossRef]
7. Spirto, M.; Melluso, F.; Nicolella, A.; Malfi, P.; Cosenza, C.; Savino, S.; Niola, V. Rolling bearing fault detection using symmetrized dot pattern indices and feedforward neural network. *Struct. Health Monit.* **2026**. [CrossRef]
8. Jiang, G.; Li, D.; Feng, K.; Li, Y.; Zheng, J.; Ni, Q.; Li, H. Rolling Bearing Fault Diagnosis Based on Convolutional Capsule Network. *J. Dyn. Monit. Diagn.* **2023**, *2*, 275–289. [CrossRef]
9. Fan, J.; Zhao, L.; Li, M. Research on Digital Twin Modeling and Fault Diagnosis Methods for Rolling Bearings. *Sensors* **2025**, *25*, 2023. [CrossRef] [PubMed]
10. Qin, Y.; Liu, H.; Mao, Y. Faulty rolling bearing digital twin model and its application in fault diagnosis with imbalanced samples. *Adv. Eng. Inform.* **2024**, *61*, 102513. [CrossRef]
11. Ming, Z.; Tang, B.; Deng, L.; Yang, Q.; Li, Q. Digital twin-assisted fault diagnosis framework for rolling bearings under imbalanced data. *Appl. Soft Comput.* **2025**, *168*, 112528. [CrossRef]
12. Jiang, K.; Cai, Y.; Han, D.; Su, Y.; Feng, G. A new bearing fault diagnosis method based on digital twin-assisted domain adaptation transfer learning. *Struct. Health Monit.* **2025**, *25*, 2963–2981. [CrossRef]
13. You, K.; Liu, C.; Lin, Y.; Qiu, G.; Gu, Y. DTMPI-DIVR: A digital twins for multi-margin physical information via dynamic interaction of virtual and real sound-vibration signals for bearing fault diagnosis without real fault samples. *Expert Syst. Appl.* **2025**, *292*, 128592. [CrossRef]
14. Fang, C.; Chang, Q.; Hu, X.; Zhou, W.; Meng, X. A digital twin-enabled domain adaptation network for cross-space fault diagnosis of roller bearings. *Mech. Syst. Signal Process.* **2025**, *236*, 113053. [CrossRef]

15. Zhang, X.; Zhen, D.; Feng, G.; Cui, Z.; Zhang, H.; Gu, F. Resonance-aware digital twin-driven data augmentation for bearing fault diagnosis under sample imbalance. *Struct. Health Monit.* **2026**. [CrossRef]
16. Shang, Z.; Wang, Z.; Pan, C.; Li, W.; Gao, M. Physics-informed auto-encoder based on digital twin for rolling bearing fault diagnosis under imbalanced sample conditions. *Eng. Appl. Artif. Intell.* **2026**, *163*, 113017. [CrossRef]
17. Ming, Z.; Tang, B.; Deng, L.; Li, Q.; Li, Z.; Zhang, X. Digital twin-enhanced framework for rolling bearings fault diagnosis under imbalanced and open-set conditions. *Mech. Syst. Signal Process.* **2026**, *257*, 114639. [CrossRef]
18. Shi, H.; Li, Y.; Zhao, C.; Wang, J.; Liu, X. A method for discontinuous data repair and precise health monitoring of rolling bearings integrating particle filtering and digital twin. *Mech. Syst. Signal Process.* **2026**, *250*, 114136. [CrossRef]
19. Zeynivand, M.; Kermansaravi, A.; Vahedi, H.; Gruosso, G. Digital Twin Synthetic Dataset for Bearing Fault Diagnosis in Industrial Spindles. *IEEE Access* **2026**, *14*, 29369–29386. [CrossRef]
20. Xiao, S.; Hou, J.; Khan, M.J.; Sun, B. Digital twin-driven zero-shot transfer diagnosis for rolling bearing faults under data-scarce conditions. *Meas. Sci. Technol.* **2025**, *36*, 106111. [CrossRef]
21. Zhang, W.; Yan, B.; Lu, S.; Du, D.; Zhou, F. Digital twin-driven few-shot fault data generation and intelligent diagnosis for rolling bearings. *J. Control Decis.* **2025**, 1–18. [CrossRef]
22. Fan, Z.; Liu, X.; Cao, Z.; Wang, H.; Zhou, Y.; Liu, Y. Digital twin-driven feature enhancement generative adversarial network for rolling bearings fault diagnosis. *Struct. Health Monit.* **2025**. [CrossRef]
23. Dai, C.; He, D.; Jin, Z.; Zhang, X.; Chen, G.; Wu, J.; Zhao, J.; Xu, Y. Digital twin-assisted graph contrastive domain adaptation for small-sample bearing fault diagnosis. *Struct. Health Monit.* **2026**. [CrossRef]
24. Li, D.; Pang, Z.; Chen, Y.; Yang, K.; Shao, J.; Luo, Y.; Zeng, Y.; He, C.; Gao, Y. FD-MVLLM: Fault diagnosis based on multimodal vibration data and large language model for bearing system. *Mech. Syst. Signal Process.* **2025**, *239*, 113226. Correction in *Mech. Syst. Signal Process.* **2026**, *251*, 114094. [CrossRef]
25. Xiao, C.; Liu, X.; Wulamu, A.; Zhang, D. KG-SR-LLM: Knowledge-Guided Semantic Representation and Large Language Model Framework for Cross-Domain Bearing Fault Diagnosis. *Sensors* **2025**, *25*, 5758. [CrossRef] [PubMed]
26. Zeng, Y.; Wang, H.; Ran, G.; Dong, Z.; Li, X.; Liu, X.; He, J. Signal LLM: Fault perception and maintenance system for dual-drive hydraulic turbine bearings based on cross-attention-LLM. *Struct. Health Monit.* **2026**. [CrossRef]
27. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
28. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Guo, Q.; Wang, M.; et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2312.10997. [CrossRef]
29. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
30. Yan, S.Q.; Gu, J.C.; Zhu, Y.; Ling, Z.H. Corrective Retrieval Augmented Generation. *arXiv* **2024**, arXiv:2401.15884. [CrossRef]
31. Lukens, S.; Bishof, M.; Siddiqui, N.; West, D. Evaluation Framework for Fault Diagnosis Using Technical Manuals in Retrieval-Augmented Large Language Models. *Annu. Conf. PHM Soc.* **2025**, *17*, 1–13 . [CrossRef]
32. Tang, J.; Peng, S.; Guo, J.; Song, D.; Gao, D.; Liu, W.; Wang, F. DT and LLM driven intelligent maintenance system for L-DED and DAG-based LLM fault diagnosis evaluation framework. *Appl. Soft Comput.* **2025**, *185*, 113942. [CrossRef]
33. Miao, R.; Wu, T.; Zhang, Z. Graph RAG-based fault diagnosis for train bogies using knowledge graphs and large language model. *Knowl.-Based Syst.* **2026**, *331*, 114855. [CrossRef]
34. Xu, J.; Xu, Z.; Jiang, Z.; Chen, Z.; Luo, H.; Wang, Y.; Gui, W. Labeling-free RAG-enhanced LLM for intelligent fault diagnosis via reinforcement learning. *Adv. Eng. Inform.* **2026**, *69*, 103864. [CrossRef]
35. Razaq, W.B.M.; Chen, H.; Machado, M.R.; Rebelo Moreira, J.L. How can the integration of AI large language models and knowledge graph enhance fault diagnosis? A systematic literature review. *Appl. Soft Comput.* **2026**, *194*, 114908. [CrossRef]
36. Tsallis, C.; Papageorgas, P.; Munteanu, R.A.; Dellagi, S. Large Language and Foundation Models for Machinery Health Monitoring: A Systematic Review. *Appl. Sci.* **2026**, *16*, 2493. [CrossRef]
37. Montgomery, D.C.; Runger, G.C. *Applied Statistics and Probability for Engineers*, 7th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2017.
38. Hong, S.; Zhou, Z.; Zio, E.; Hong, K. Condition Assessment for the Performance Degradation of Bearing Based on a Combinatorial Feature Extraction Method. *Digit. Signal Process.* **2014**, *27*, 159–166. [CrossRef]
39. Buchaiah, S.; Shakya, P. Bearing Fault Diagnosis and Prognosis Using Data Fusion Based Feature Extraction and Feature Selection. *Measurement* **2021**, *188*, 110506. [CrossRef]
40. Li, X.; Teng, W.; Peng, D.; Ma, T.; Wu, X.; Liu, Y. Feature Fusion Model Based Health Indicator Construction and Self-Constraint State-Space Estimator for Remaining Useful Life Prediction of Bearings in Wind Turbines. *Reliab. Eng. Syst. Saf.* **2023**, *233*, 109124. [CrossRef]

41. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 2318–2335. [CrossRef]
42. Qiu, H.; Lee, J.; Lin, J.; Yu, G. Wavelet Filter-Based Weak Signature Detection Method and Its Application on Rolling Element Bearing Prognostics. *J. Sound Vib.* **2006**, *289*, 1066–1090. [CrossRef]
43. SKF Group. *SKF Bearing Maintenance Handbook*; Technical Manual, PUB 10001/1 EN; SKF Group: Gothenburg, Sweden, 2011.
44. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
45. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julians, Malta, 21–22 March 2024; pp. 150–158. [CrossRef]