Int. J. of Precis. Eng. and Manuf.-Green Tech.
https://doi.org/10.1007/s40684-026-00880-y

Online ISSN 2198-0810
Print ISSN 2288-6206

REGULAR PAPER

Efficient Input Data Strategies for LLMs (Large Language Models)-Based Bearing Fault Diagnosis

Pan-Gyun Jeong[1,6] · Yuehua Hu[2,6] · Jiyeong Kong[3,6] · Kyungwoo Song[4] · Bongyoung Yoo[1,5] · Kyung-Tae Kang[6]

Received: 22 October 2025 / Revised: 10 March 2026 / Accepted: 11 March 2026

Abstract

Early detection of bearing faults is vital for ensuring the reliability and safety of rotating machines such as automobile wheels and power plant turbines, because undetected degradation may lead to unexpected failures and operational risks. As bearings deteriorate, vibration signals exhibit characteristic changes reflecting underlying mechanical issues. Machine learning (ML) and deep learning (DL) have advanced fault detection but often rely on task-specific architectures with extensive labeled datasets and frequent retraining. Recently, large language models (LLMs) have emerged as a complementary approach, offering strong generalization through prompt-based learning. With well-designed prompts and structured inputs, LLMs can perform time-series reasoning and support diagnosis via few-shot or zero-shot learning without retraining. However, limited input capacity of LLMs challenges the handling of long-duration vibration sequences essential for modeling gradual progression. To address this, we propose a four-stage framework enabling compact representation of vibration data and efficient delivery of essential information for LLM-based analysis. First, statistical and spectral features are extracted and compressed using Principal Component Analysis (PCA) to reduce token count while preserving key information. Second, data is reformatted into natural-language-compatible input while maintaining temporal structure. Third, a Multi-Strategy Retrieval-Augmented Generation (Multi-Strategy RAG) architecture processes large-scale time-series data. Finally, task-specific prompts guide the LLM to reasoning tasks such as detection and interpretation. The proposed input data strategy leads to enhanced fault diagnosis performance across publicly available bearing benchmark datasets.

Keywords Large Language Models · Prompt Engineering · Retrieval-augmented generation · Time-series · Bearing · Fault diagnosis

✉ Kyung-Tae Kang
kangkt65@gmail.com

Pan-Gyun Jeong
panguni@hanyang.ac.kr

Yuehua Hu
leilei968@hanyang.ac.kr

Jiyeong Kong
jykong@korea.ac.kr

Kyungwoo Song
kyungwoo.song@yonsei.ac.kr

Bongyoung Yoo
byyoo@hanyang.ac.kr

[1] HYU-KITECH Joint Department, Hanyang University, Ansan 15588, Korea

[2] Department of Photonics and Nanoelectronics, Hanyang University, Ansan 15588, Korea

[3] Micro/Nano System Department, Korea University, Seoul 02841, Korea

[4] Department of Applied Statistics, Yonsei University, Seoul 03722, Korea

[5] Department of Materials Science and Chemical Engineering, Hanyang University, Ansan 15588, Korea

[6] Korea Institute of Industrial Technology, Ansan 15588, Korea

Published online: 27 April 2026

1 Introduction

In rotating machines from relatively small automobile wheels and to huge power plant turbine, bearing defects represent a critical failure mode, accounting for nearly 30% of total system malfunctions and posing a substantial risk to operational safety [1]. As rotors approach critical speeds, structural deformations can lead to increased bearing losses, elevated viscous resistance, and accelerated degradation [2]. Additionally, geometric imperfections, gyroscopic effects, and residual imbalance in bearings can trigger severe vibrations, which degrade system performance and, in the long term, threaten the integrity of the entire machine [3]. Given these risks, timely detection and monitoring of bearing faults are essential for maintaining reliability and preventing cascading failures in rotating systems.

A wide range of studies have explored bearing fault diagnosis using various machine learning (ML) and deep learning (DL) models. ML approaches include the use of signal preprocessing and handcrafted feature extraction from time, frequency, or time-frequency domains, followed by classification through traditional algorithms such as support vector machines (SVM) [4] and k-nearest neighbors (k-NN) [5]. DL models such as autoencoders (AEs) [6], restricted Boltzmann machines (RBMs) [7], and convolutional neural networks (CNNs) [8] have been applied to learn hierarchical representations directly from raw or transformed input data, eliminating the need for manual feature engineering.

The emergence of recurrent neural networks such as long short-term memory (LSTM) [9] and gated recurrent units (GRU) [10] has further facilitated time-series modeling in bearing diagnostics. These architectures have led to the development of LSTM-based end-to-end fault diagnosis frameworks [11] and GRU-based models designed for robust fault classification and degradation detection under noisy and variable operating conditions [12]. Although ML and DL techniques have significantly advanced fault diagnosis by enabling automated feature extraction and improving detection accuracy, their deployment in real-world industrial environments remains limited. These models typically require large volumes of labeled data and frequent retraining to accommodate changing operating conditions. Moreover, their limited interpretability often poses challenges for field operators who must act on model outputs in real time [13].

To address these challenges, researchers have explored the use of LLMs, which introduce a new paradigm of prompt-based learning and offer the potential for more flexible, explainable, and data-efficient solutions.

2 Related Works

Recent developments in LLMs have introduced a prompt-based paradigm that allows users to guide model behavior using natural language inputs rather than modifying model parameters or retraining on task-specific data [14]. By referencing a few examples (few-shot learning [15]) or following brief task instructions (zero-shot learning [16]), LLMs can generate flexible, context-aware responses tailored to diverse diagnostic needs. In doing so, LLMs offer a practical alternative that supports more dynamic, explainable, and low-maintenance diagnostic solutions.

To enable fault diagnosis and forecasting based on numerical sensor data, it is essential to restructure time-series inputs into a form that LLMs can effectively interpret and reason over, while preserving and extracting the underlying semantic patterns embedded in the data. This has been approached by either translating time-series data into natural language using techniques such as tokenization for language modeling [17] and prompt-based encoding for sequence-to-sequence inference [18], or by retrieving the most relevant contextual segments from external memory and incorporating them into prompts through retrieval-augmented generation (RAG) [19].

Both strategies are constrained by the token limitations inherent to current LLM architectures. Textual transformation of large-scale time-series data often leads to excessive input length, which degrades model performance even within the context window [20], and makes it difficult for the model to attend to information located mid-sequence due to limitations in long-context attention [21]. RAG-based methods, while more efficient in reducing input size, still suffer from performance degradation when the total number of retrieved tokens approaches the model's capacity [22], and are prone to redundant retrieval and context dilution, which reduce instruction adherence [23].

We propose a structured pipeline that enables LLMs to effectively perform fault diagnosis in rotating machinery systems. The pipeline begins with principal component analysis (PCA), which compresses high-dimensional features into a lower-dimensional space while preserving the core degradation patterns. After extracting a large set of features from the original time-series vibration data, PCA is applied to compactly capture bearing-related fault characteristics while simultaneously reducing dimensionality. The resulting PCA scores are then converted into textual representations, transforming them into a format more suitable for LLM-based reasoning. Compared to using raw vibration features directly, this process significantly reduces the number of tokens required per file, allowing a broader time range of the original time-series data to be represented within each LLM input. Subsequently, a Multi-Strategy

RAG architecture is employed to retrieve relevant content from large-scale time-series data based on semantic similarity, recency, and randomized data provision.

This Multi-Strategy retrieval is designed not only to enhance contextual relevance but also to mitigate data bias and reduce the risk of hallucination in LLM responses. Finally, prompt engineering techniques aligned with the structure and semantics of the retrieved PCA-guided content are employed to guide the LLM's interpretation, thereby supporting accurate and timely fault identification.

3 Methodology

3.1 Dataset and Signal Preprocessing

This study employs bearing vibration signals from the IMS, CWRU, and Paderborn University (PU) bearing datasets, covering diverse operating and degradation conditions. The dataset includes measurements from multiple bearings (indexed as Bearing1-Bearing4 in this study), with vibration signals sampled at a dataset-dependent frequency f_s. Each vibration signal is segmented into fixed-length intervals, where the number of samples in each segment is defined as:

$$L = f_s \cdot T \tag{1}$$

where Tdenotes the segment duration, which varies across datasets.

For each bearing $k \in \{1,2,3,4\}$and each acquisition timestamp j, the corresponding vibration signal segment is expressed as:

$$x_{j,k}(n), \quad n = 0,1,\ldots,L-1 \tag{2}$$

To construct meaningful input representations, we extract statistical features from both time and frequency domains [24]. The combined feature vector for the $j-\text{th}$ segment of bearing k is defined as:

$$F_k^{(j)} = \left[F_{k,time}^{(j)}, F_{k,freq}^{(j)}\right] \in R^{15} \tag{3}$$

The time-domain vector $F_{k,time}^{(j)} \in R^{10}$ includes ten statistical descriptors:

$$F_{k,time}^{(j)} = \begin{bmatrix} \mu_k^{(j)}, RMS_k^{(j)}, \sigma_k^{(j)}, CF_k^{(j)}, Skew_k^{(j)}, \\ SF_k^{(j)}, Kurt_k^{(j)}, P2P_k^{(j)}, EF_k^{(j)}, IF_k^{(j)} \end{bmatrix} \tag{4}$$

The frequency-domain vector $F_{k,freq}^{(j)} \in R^5$ consists of spectral characteristics derived from the Fourier-transformed signal:

$$F_{k,freq}^{(j)} = \begin{bmatrix} PeakFreq_k^{(j)}, P2PFreq_k^{(j)}, \\ SpecKurt_k^{(j)}, SpecBW_k^{(j)}, \\ SpecSkew_k^{(j)} \end{bmatrix} \tag{5}$$

3.2 Augmentation and Normalization

To better capture the temporal dynamics in vibration signals, we extend the input representation by incorporating first- and second-order differences of the extracted features prior to dimensionality reduction. This delta augmentation strategy has been shown to enrich feature representations and enhance degradation pattern discovery [25]. The first and second-order delta components are defined as:

$$d \in \{time, freq\} \tag{6}$$

$$\Delta F_{k,d}^{(j)} = F_{k,d}^{(j)} - F_{k,d}^{(j-1)} \tag{7}$$

$$\Delta^2 F_{k,d}^{(j)} = \Delta F_{k,d}^{(j)} - \Delta F_{k,d}^{(j-1)} \tag{8}$$

The augmented feature vector becomes:

$$m_{time} = 30, \quad m_{freq} = 15 \tag{9}$$

$$\widehat{F}_{k,d}^{(j)} \in R^{m_d} \tag{10}$$

To mitigate the influence of noise or outliers, we apply a robust normalization technique using the Median Absolute Deviation (MAD) [26]. Each delta-augmented feature vector is standardized as:

$$\widetilde{F}_{k,d}^{(j)} = \frac{\widehat{F}_{k,d}^{(j)} - \mu_d}{MAD\left(\left|\widehat{F}_{k,d}^{(j)} \mu\right|\right) + \epsilon_d} \tag{11}$$

where, μ denotes the median and ε is a small constant to prevent division by zero. Although originally proposed for robust outlier detection this MAD-based formulation is applied here for feature normalization, being adapted to ensure consistent scaling across samples.

3.3 PCA Score Computation

To reduce dimensionality while preserving degradation-relevant information, PCA is applied separately to the time-domain and frequency-domain feature blocks after normalization and delta augmentation [27, 28].

The time-domain block $\tilde{F}^{(j)}_{k,time} \in R^{30}$ and the frequency-domain block $\tilde{F}^{(j)}_{k,freq} \in R^{15}$ are collected across N segments to form:

$$Z_d \in R^{N \times m_d} \tag{12}$$

The covariance matrices are computed as:

$$C_d = \frac{1}{N-1} Z_d^{\mathrm{T}} Z_d \tag{13}$$

The first principal components $w_{1,time}$ and $w_{1,freq}$ are then obtained, and each segment is projected onto these directions to yield the PCA scores:

$$S^{(j)}_{k,d} = \left(\tilde{F}^{(j)}_{k,d}\right)^{\mathrm{T}} w_{1,d} \in R \tag{14}$$

Finally, the time-domain and frequency-domain PCA scores are concatenated:

$$S^{(j)}_{k} = \left[S^{(j)}_{k,time}, S^{(j)}_{k,freq}\right] \in R^2 \tag{15}$$

This representation preserves the dominant degradation trends of both time and frequency domains in a concise and interpretable form and is further used to generate text files for bearing fault detection.

As shown in Fig. 1(a), from a conventional PCA perspective, retaining multiple principal components is required for the time-domain feature block to exceed the commonly used 80% cumulative variance threshold. This reflects the multi-factor nature of time-domain statistical features and is acknowledged in this study.

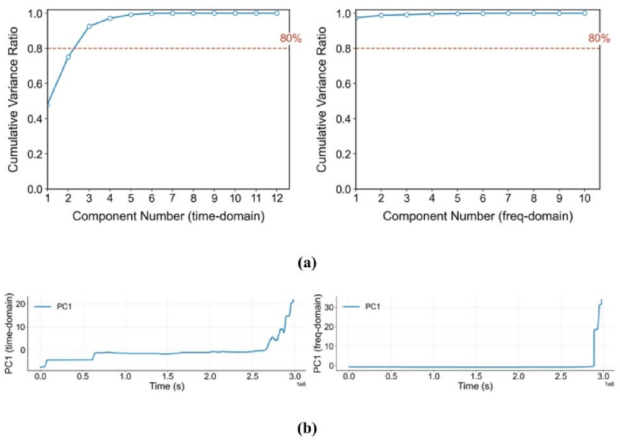


Fig. 1 Example PCA results for time-domain and frequency-domain features, (**a**) Cumulative explained variance ratios of principal components for the time-domain features (left) and the frequency-domain features (right), (**b**) Temporal evolution of the first principal component (PC1) scores for the time-domain features (left) and the frequency-domain features (right)

However, the objective of the proposed framework is not variance maximization or signal reconstruction, but dimensionality reduction under the constraints of LLM-based few-shot inference. In this setting, feature values are serialized into text tokens and processed within a fixed context window. Retaining additional principal components increases token length proportionally, which reduces the temporal span that can be provided to the LLM.

Figure 1(b) shows that the first principal component in both the time and frequency domains preserves the dominant temporal degradation pattern, including the late-stage monotonic increase associated with bearing degradation. This indicates that the primary degradation-related temporal information remains captured despite the reduced dimensionality.

Accordingly, only the first principal component from each domain is retained to control token growth while maintaining the long-term temporal structure required for few-shot LLM-based fault reasoning.

3.4 Score-to-Text Template

To enable LLMs to effectively process the compressed vibration information, we convert the PCA-derived time- and frequency-domain scores into structured natural language templates. At each timestamp, the compressed scores from all four bearings are reformatted into a unified textual representation that includes the measurement timestamp, extracted features from each domain, and the corresponding PCA scores for each bearing.

By presenting degradation trends as semantically aligned text, this representation improves interpretability and facilitates effective document retrieval in the subsequent prompt-based forecasting process. The overall workflow from data acquisition to PCA-score-based textual conversion is illustrated in Fig. 2, and an example of the prompt template is provided in **Appendix A.**

3.5 Multi-Strategy RAG for Large-Scale Time-Series

To support efficient and scalable inference over long-term vibration monitoring data, we implement a Multi-Strategy RAG that selects a subset of relevant PCA score summaries prior to language model inference.

The retrieval strategy combines cosine similarity-based document selection to ensure semantic alignment with the query [29], recency-aware filtering to reflect the latest degradation states and maintain temporal relevance [30], and random sampling of temporally contiguous blocks to

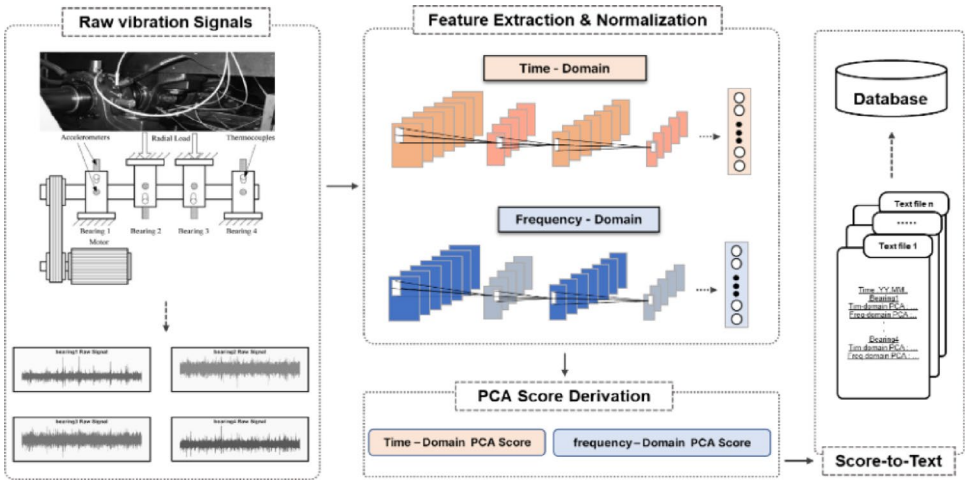


Fig. 2 Workflow for transforming vibration signals into structured textual templates using PCA

introduce distributional diversity and capture the overall temporal progression of the signal [31].

3.5.1 Document and Query Embedding

We embed each document $d_i \in D$ and the query q into dense vector representations:

$$v_i = embedding\,(d_i) \in R^n \tag{16}$$

$$v_q = embedding\,(q) \in R^n \tag{17}$$

3.5.2 Similarity-Based Retrieval

We compute dense embeddings for each document using a pretrained sentence transformer, and select the top- k documents that are semantically most relevant to the user query based on cosine similarity:

$$sim\,(q, d_i) = \frac{v_q \cdot v_i}{|v_q|_2\,|v_i|_2} \tag{18}$$

$$TopK_{sim}\,(q) = Top - k\,(\{d_i \in D | sim\,(q, d_i)\}) \tag{19}$$

where v_q and v_i are the vector embeddings of the query and document d_i, respectively.

3.5.3 Recency-Based Retrieval

To ensure inclusion of the most recent degradation signals, we select the latest k documents based on timestamp metadata:

$$TopK_{recent} = sort_{desc}\,(\{d_i \in D | timestamp\,(d_i)\})\,[:k] \tag{20}$$

3.5.4 Random Contiguous Block Sampling

To preserve temporal locality and avoid overfitting to highly similar patterns, we randomly sample a continuous block of k documents:

$$s \sim U\,(0, N - k) \tag{21}$$

$$RandomBlock_k = \{d_s, d_{s+1}, \ldots, d_{s+k-1}\} \tag{22}$$

The final retrieved set is the union of all three sources, with duplicates removed:

$$D_{retrieved} = TopK_{sim} \cup TopK_{recent} \cup RandomBlock_k \tag{23}$$

These documents are then sorted in ascending order of timestamp to preserve chronological progression:

$$D_{input} = sort_{desc}\,(D_{retrieved},\ by\ timestamp) \tag{24}$$

By leveraging a Multi-Strategy RAG framework that integrates query-relevant similarity retrieval, recency-based selection reflecting the latest system status and random sampling to preserve distributional diversity, the proposed approach ensures contextual relevance, captures recent degradation signals, and mitigates data bias in large-scale time-series analysis.

3.5.5 Text-Template-Aligned Prompt

To generate interpretable fault diagnosis from PCA-compressed vibration data, we design a prompt that aligns with the score-to-text template described in Sect. 3.4.

This Text-Template-Aligned Prompt provides the LLM with structured domain guidance for analyzing degradation patterns and identifying faulty bearings. It includes:

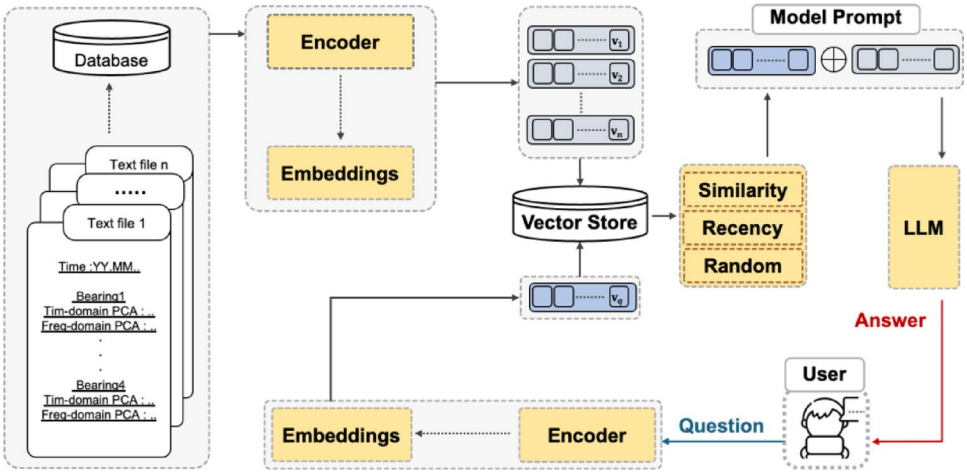


Fig. 3 Integrated pipeline for PCA-to-text conversion, prompt alignment, and Multi-Strategy RAG

- A system role definition instructing the LLM to act as a bearing degradation analysis assistant.
- Interpretation rules for assessing the direction and magnitude of PCA score changes.
- Reasoning guidelines for identifying abnormal degradation patterns linked to faults.
- Definition of the desired output format.

By aligning the input structure with the prompt's reasoning format, this design reduces common few-shot inference biases such as recency and label-frequency effects and improves prediction stability and consistency [32].

As illustrated in Fig. 3, the proposed framework integrates PCA-to-text conversion (Sect. 3.1–3.4), the Multi-Strategy RAG framework (Sect. 3.5), and text-template-aligned prompt engineering (Sect. 3.6) into a unified workflow for reliable degradation analysis. Additional implementation details and prompt examples are provided in **Appendix B**.

4 Experimental Setup

4.1 LLM backbone used in the experiment

In this study, we compare the performance of four LLMs: Llama 3.1, Gemma 3, Qwen 3, and Mistral-small 3.2, each serving as the core engine for semantic inference. Detailed specifications of these models are provided in **Appendix C**.

4.1.1 Datasets Used in the Experiments

In this study, two widely used bearing benchmarking datasets, namely the IMS and CWRU datasets, are employed for performance evaluation and comparative analysis.

The IMS datasets consist of long-term vibration measurements collected under accelerated degradation conditions (NASA/IMS, 2004) [33]. Acceleration signals were recorded from four bearings (Bearing1–Bearing4) at approximately 20 kHz using an NI DAQ Card 6062E, with 1-second recordings acquired every 10 min, yielding segments of 20,480 samples. Dataset configuration and failure details are summarized in Table 1.

Table 1 Summary of IMS bearing datasets and failure information

Dataset	Motor speed (rpm)	Radial Load (*N*)	Healthy Bearings	Faulty Bearings
Set 1	2000	6000	Bearing1 Bearing2	Bearing3 Bearing4
Set 2	2000	6000	Bearing2 Bearing3 Bearing4	Bearing1
Set 3	2000	6000	Bearing1 Bearing2 Bearing4	Bearing3

Table 2 Summary of CWRU bearing datasets and failure information

Sample ID	Motor speed (rpm)	Load (HP)	Condition
Normal_1	1730	3	Healthy
Normal_2	1750	2	Healthy
B_7_F2	1750	2	Faulty
B_21_DE12	1750	2	Faulty

The CWRU dataset [34] provides vibration measurements for bearing fault diagnosis under steady-state operating conditions. The data are distributed in MATLAB (.mat) format and organized into four motor speed groups (1797, 1772, 1750, and 1730 rpm), each comprising Normal, Drive-End (DE), and Fan-End (FE) fault conditions. DE fault signals were sampled at 12 kHz and 48 kHz, while FE fault signals were collected at 12 kHz.

To ensure consistency with the IMS-based analysis, four representative conditions were selected: two normal bearings at 1730 rpm and 1750 rpm, and two faulty bearings at 1750 rpm with ball defects (B_7_F2 and B_21_DE12). Details of the selected conditions and fault types are summarized in Table 2.

Additional experiments were performed on the Paderborn University (PU) bearing dataset [35], which contains 64 kHz vibration and motor current signals collected under diverse operating conditions using 6203 deep-groove ball bearings. Two healthy bearings (K001, K002) and two faulty bearings (KA01, KI01) with inner- and outer-ring damage were selected from accelerated lifetime tests. Dataset details are summarized in Table 3.

4.2 LLM Response Format and Evaluation Metrics

To assess the performance of the LLM, we designed a prompt-based evaluation protocol consisting of 100 repeated trials using a unified input format.:

> Q1. Identify the faulty bearing based on the PCA-compressed degradation trend.

Given the generative nature of LLMs and their tendency to produce free-form outputs, we strictly defined the expected output structure to ensure consistency and enable automated evaluation. All models were required to respond using the following constrained format:

> Predicted faulty bearing: bearing X[, bearing Y].

This format accommodates both single-fault and multi-fault scenarios. In single-fault cases, the model is expected to output a single bearing identifier. In multi-fault cases, the model should list all predicted faulty bearings as a comma-separated sequence. This design ensures consistency in output formatting while supporting evaluation in both single-label and multi-label fault settings.

4.2.1 Accuracy

Accuracy measures the proportion of fully correct predictions:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ trials} \tag{25}$$

Single-Fault: Prediction must exactly match the ground truth to be correct.

Table 3 Summary of Paderborn University (PU) bearing datasets and failure information

Sample ID	Motor speed (rpm)	Radial Load (*N*)	Condition
K001	1500	1000–3000	Healthy
K002	1500	3000	Healthy
KA01	1500	1000	Faulty
KI01	1500	1000	Faulty

Multi-Fault: Only predictions that include all ground-truth bearings are scored as correct. To evaluate partial correctness in multi-fault cases, we computed as follows:

$$PartialAccuracy_i = \frac{|Prediction_i \cap GroundTruth_i|}{|GroundTruth_i|} \tag{26}$$

This yields 1.0 if all faults are identified, 0.5 for one out of two, and 0.0 otherwise.

4.2.2 F1 Score

The weighted $F1$ score reflects both precision and recall across classes, weighted by class support:

$$F1_w = \frac{1}{N}\sum_{c=1}^{C} support_c \cdot F1_c \tag{27}$$

Where:

$$F1_c = \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c} \tag{28}$$

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \tag{29}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{30}$$

In single-fault scenarios, only one class appears in the ground truth. As a result, the weighted $F1$ score reduces to the $F1$ score of class X:

$$F1_w = F1_X \tag{31}$$

Multi-Fault: Each bearing is treated as an independent label, and the final $F1$ score is computed using binary scores for each class, where X and Y refer to the respective faulty bearing classes involved in the multi-label evaluation, as follows:

$$F1_w = \frac{support_X \cdot F1_X + support_Y \cdot F1_Y}{support_X + support_Y} \tag{32}$$

4.3 Contribution of Retrieval Strategies in RAG

Before the main experiments, we conduct a prior validation by comparing recency-based, random block-based, similarity-based, and Multi-Strategy RAG configurations under identical conditions using the same PCA-compressed vibration summaries, as shown in Fig. 4.

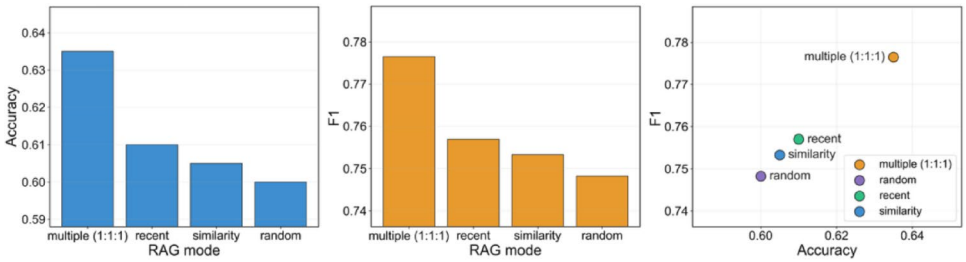


Fig. 4 Comparison between single-strategy RAG (Similarity-only, Recency-only, Random-only) and the proposed Multi-Strategy RAG in time-series fault diagnosis (IMS Dataset 1, averaged over four LLM backbones)

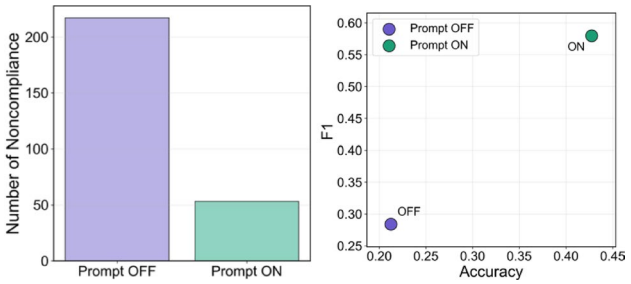


Fig. 5 Impact of prompt engineering on response format compliance and fault-diagnosis performance, comparing Prompt ON and Prompt OFF configurations (IMS Dataset 1, averaged over four LLM backbones)

The results demonstrate that the Multi-Strategy configuration consistently outperforms all single-strategy variants in terms of both accuracy and F1 score.

This performance advantage reflects the inherent characteristics of time-series degradation data, which require simultaneous consideration of the current system condition, historical evolution, and semantic relevance.

Recency-based retrieval highlights recent degradation signals that are indicative of the present state, similarity-based retrieval selects historically observed degradation patterns that are most relevant to the given diagnostic query, and random block retrieval preserves long-term temporal structure across the degradation process.

By integrating these complementary retrieval mechanisms, the Multi-Strategy RAG constructs a more balanced and informative context, enabling more reliable fault diagnosis than any single retrieval strategy alone.

4.4 Contribution of Prompt Engineering

We evaluate the effect of prompt engineering through a controlled comparison between Prompt ON and Prompt OFF configurations, considering both response-format compliance and fault-diagnosis performance. In this evaluation, predictions that do not follow the required output format, *"Predicted faulty bearing: bearingX[, bearingY]"*, are treated as invalid and assigned a score of zero, thereby directly penalizing noncompliant responses.

As shown in Fig. 5, the Prompt ON configuration substantially reduces the number of format-violating outputs compared to Prompt OFF, indicating a clear improvement in the model's ability to follow explicit user instructions. This improvement in format compliance is accompanied by consistent gains in Accuracy and F1 score, demonstrating that prompt engineering contributes not only to syntactic correctness but also to more reliable fault-diagnosis performance.

These improvements can be directly attributed to the Text-Template Aligned Prompt described in Sect. 3.4, which provides structured domain guidance by explicitly linking PCA patterns to bearing degradation interpretation. By constraining how score changes are analyzed and enforcing a fixed output format, the prompt guides the LLM toward task-relevant reasoning.

As a result, the LLM produces more interpretable, consistent, and evaluable fault-diagnosis outputs from PCA score.

5 Results

Tables 4 and 5 present a benchmark evaluation of LLM-based bearing fault diagnosis across the IMS, CWRU, and PU datasets, analyzing the effects of PCA-based compression and Multi-Strategy RAG configurations in terms of accuracy and F1-score.

Across the IMS and CWRU datasets, PCA-RAG (5%) consistently achieves the highest diagnostic performance for all evaluated LLM backbones, outperforming both raw-input and No-RAG configurations in terms of accuracy and F1-score. This result demonstrates that, for large-scale and long-duration vibration time series, effective fault diagnosis with LLMs critically depends on the joint use of data compression and selective retrieval.

PCA-based compression reduces token redundancy while preserving dominant degradation trends, allowing long temporal spans to be represented within the LLM's context window. When combined with a moderate retrieval ratio, RAG (5%) supplies diagnostically relevant historical context without saturating the input space. In contrast, both higher retrieval ratios (RAG 50%) and the absence of retrieval (No-RAG) lead to degraded performance,

Table 4 Accuracy comparison of LLM-based bearing fault diagnosis on the IMS, CWRU, and PU datasets under different PCA/RAG settings

	Llama 3.1		Gemma 3		Qwen 3		Mistral-small 3.2	
	PCA	Raw	PCA	Raw	PCA	Raw	PCA	Raw
IMS Bearing dataset (average over datasets 1–3)								
Rag(5%)	**0.6583**	*0.4916*	**0.6450**	*0.5183*	**0.5300**	*0.4500*	**0.7600**	*0.7333*
Rag(50%)	0.4178	0.5777	0.2700	0.2933	0.3516	0.3316	0.4167	0.5871
No-Rag	0.3733	0.3600	0.2866	0.3750	0.3383	0.3850	0.4783	0.6338
CWRU Bearing dataset								
Rag(5%)	**1.0000**	*0.9100*	**0.9900**	*0.9550*	**0.9900**	0.9500	**1.0000**	*0.9900*
Rag(50%)	0.9100	0.7250	0.5950	0.6000	*0.9550*	0.7400	0.9750	0.8600
No-Rag	0.8900	0.6900	0.6050	0.4900	0.9500	0.5200	0.9900	0.6800
PU Bearing dataset								
Rag(5%)	**0.7750**	*0.7350*	**0.5950**	*0.5100*	*0.6750*	0.5550	*0.6400*	0.5950
Rag(50%)	0.5350	0.5700	0.5050	0.4800	0.5750	0.5750	0.5500	0.5950
No-Rag	0.5450	0.7550	0.5000	0.4800	0.5850	**0.6850**	0.5250	**0.7650**

Table 5 F1-score comparison of LLM-based bearing fault diagnosis on the IMS, CWRU, and PU datasets under different PCA/RAG settings

	Llama 3.1		Gemma 3		Qwen 3		Mistral-small 3.2	
	PCA	Raw	PCA	Raw	PCA	Raw	PCA	Raw
IMS Bearing dataset (average over datasets 1–3)								
Rag(5%)	**0.7924**	*0.6047*	**0.7538**	*0.6212*	**0.6384**	*0.5804*	**0.8225**	*0.7933*
Rag(50%)	0.5777	0.5173	0.3867	0.4009	0.4700	0.4764	0.5806	0.5871
No-Rag	0.5300	0.5055	0.4021	0.4997	0.4619	0.5253	0.6329	0.6338
CWRU Bearing dataset								
Rag(5%)	**1.0000**	*0.9526*	**0.9949**	*0.9766*	**0.9949**	0.9743	**1.0000**	*0.9949*
Rag(50%)	0.9528	0.8403	0.6951	0.7389	*0.9769*	0.8496	0.9872	0.9232
No-Rag	0.9416	0.8124	0.7289	0.6489	0.9741	0.6839	0.9949	0.8034
PU Bearing dataset								
Rag(5%)	**0.8564**	0.8470	**0.6727**	*0.6346*	*0.7621*	0.6431	*0.7222*	0.7460
Rag(50%)	0.6969	*0.8538*	0.5499	0.5755	0.6754	0.6494	0.7062	0.7383
No-Rag	0.7024	0.7235	0.5532	0.5240	0.6736	**0.8084**	0.6868	**0.8664**

suggesting that excessive context or insufficient historical grounding can dilute fault-relevant information and hinder LLM reasoning.

On the PU dataset, PCA-RAG (5%) remains competitive for Llama 3.1 and Gemma 3, while Qwen 3 and Mistral-small 3.2 show comparable or higher performance with Raw-No-RAG setting. This reflects the PU dataset's relatively short and low-noise time-series, which reduces the need for aggressive compression and retrieval. Importantly, this suggests that the optimal input strategy depends on dataset scale and signal complexity, rather than indicating a limitation of the proposed approach.

Overall, the results confirm that fault-diagnosis performance on large-scale time-series data with LLMs hinges on the complementary interplay between data compression and selective retrieval.

The most reliable gains arise when these mechanisms operate in concert, and the input is tuned to approximate the LLM's effective context length a pattern consistently reflected in both accuracy and F1-score across models and datasets. Detailed dataset-specific results and analyses of context-length variations under different PCA/RAG settings are provided in **Appendix D** and **E**, respectively.

6 Conclusion

This study demonstrates that the reliability of LLM-based bearing fault analysis fundamentally depends on compact and evidence-focused input delivery rather than raw data volume. Directly supplying long, high-dimensional vibration time series degrades reasoning due to context-length constraints and diffuse attention in conventional LLMs.

By transforming vibration signals into semantically meaningful, low-dimensional representations and combining them with multi-strategy selective retrieval, the proposed framework preserves degradation-critical information while controlling context size.

Experimental results consistently show that, for large-scale vibration time series, carefully curated and compressed inputs outperform excessive raw-context provision, confirming that more context is not necessarily better.

Instead, prioritizing evidence quality over quantity leads to more stable and accurate LLM inference.

Beyond performance gains, the proposed approach is well suited for industrial deployment, where massive sensor streams must be interpreted efficiently for real-time condition monitoring and predictive maintenance. The underlying input design principle is also broadly applicable to other complex time-series domains, including energy systems, medical diagnostics, and structural health monitoring, where scalability, redundancy, and semantic clarity remain central challenges.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40684-026-00880-y.

Acknowledgements This work was supported by the Ministry of Trade, Industry and Energy (MOTIE, Korea) [Project: Development of AI autonomous manufacturing technology based on robot system for optimizing display manufacturing process / Project Number: 00507904], and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MIST) (No. RS-2020-NR049544).

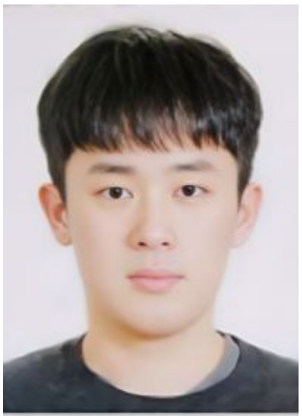
Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Li, X., Liu, J., Ding, S., Xu, Y., Zhang, Y., & Xia, M. (2023). Dynamic modeling and vibration analysis of double row cylindrical roller bearings with irregular-shaped defects. *Nonlinear Dynamics*. https://doi.org/10.1007/s11071-023-09164-5
2. Prasad, K. N. V., & Narayanan, G. (2019). Electro-magnetic bearings with power electronic control for high-speed rotating machines: A review. In 2019 National Power Electronics Conference (NPEC). https://doi.org/10.1109/npec47332.2019.9034789
3. Han, B., Huang, Z., & Le, Y. (2017). Design aspects of a large scale turbomolecular pump with active magnetic bearings. *Vacuum*, *142*, 96–105. https://doi.org/10.1016/j.vacuum.2016.12.010
4. Li, Y., Yu, W., & Huang, W. (2016). A new rolling bearing fault diagnosis method based on multiscale permutation entropy and improved support vector machine based binary tree. *Measurement*, *77*, 80–94. https://doi.org/10.1016/j.measurement.2015.08.034
5. Baraldi, P., Cannarile, F., Di Maio, F., & Zio, E. (2016). Hierarchical k-nearest neighbours classification and binary differential evolution for fault diagnostics of automotive bearings operating under variable conditions. *Engineering Applications of Artificial Intelligence*, *56*, 1–13. https://doi.org/10.1016/j.engappai.2016.08.011
6. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Nature*, *323*(6088), 696–699. https://doi.org/10.1038/323533a0
7. Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory*. Tech. Rep. CU-CS-321-86).
8. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386
9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
10. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/D14-1179
11. Hinchi, A. Z., & Tkiouat, M. (2018). Rolling element bearing remaining useful life estimation based on a convolutional long-short-term memory network. *Procedia Computer Science*, *127*, 123–132. https://doi.org/10.1016/j.procs.2018.01.106
12. Han, L., Liu, H., Zhou, J., Zheng, Y., Jiang, W., & Zhang, Y. (2018). Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders. *ISA Transactions*, *77*, 167–178. https://doi.org/10.1016/j.isatra.2018.04.005
13. Zhang, S., Wang, B., & Habetler, T. G. (2020). Deep learning algorithms for bearing fault diagnostics—A comprehensive review. *Ieee Access : Practical Innovations, Open Solutions*, *8*, 29857–29881. https://doi.org/10.1109/ACCESS.2020.2972859
14. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*. https://doi.org/10.1145/3560815
15. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan,A., Shyam, P., Sastry, G., … Amodei, D. (2020). Language models are few-shot learners.In Advances in Neural Information Processing Systems.
16. Kojima, T., Gu, S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems.
17. Gruver, N., Finzi, M., Qiu, S., & Wilson, A. G. (2023). Large language models are zero-shot time series forecasters. In Advances in Neural Information Processing Systems. https://doi.org/10.48550/arxiv.2310.07820
18. Xue, H., & Salim, F. D. (2022). PromptCast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2023.3342137
19. Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler,H., Lewis, M., … Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv: Computation and Language.
20. Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: The impact of input length on the reasoning performance of large language models. In Proceedings of the Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.48550/arxiv.2402.14848
21. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00638
22. Jin, B., Yoon, J., Han, J., & Arik, S. Ö. (2024). Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. In International Conference on Learning Representations. https://doi.org/10.48550/arxiv.2410.05983
23. Leng, Q., Portes, J., Havens, S., Zaharia, M. A., & Carbin, M. (2024). Long context RAG performance of large language models. *arXiv org*. https://doi.org/10.48550/arxiv.2411.03538
24. Qaid, H. A. A. M., Zhang, B., Li, D., Ng, S. K., & Li, W. (2024). FD-LLM: Large language model for fault diagnosis of machines. *arXiv org*. https://doi.org/10.48550/arxiv.2412.01218

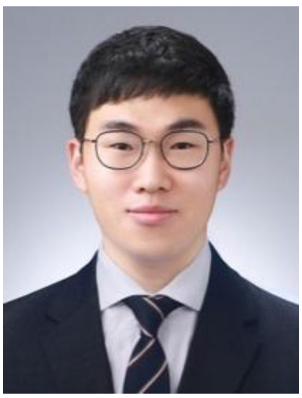
25. Susto, G. A., Schirru, A., Pampuri, S., & McLoone, S. (2016). Supervised aggregative feature extraction for big data time series regression. *IEEE Transactions on Industrial Informatics*, *12*(3), 1243–1252. https://doi.org/10.1109/TII.2015.2496231
26. Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013
27. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
28. Zhu, J., Hu, T., Jiang, B., & Yang, X. (2020). Intelligent bearing fault diagnosis using PCA–DBN framework. *Neural Computing and Applications*, *32*(14), 10773–10781. https://doi.org/10.1007/s00521-019-04612-z
29. Zhu, S., Wu, J., Xiong, H., Xia, G., & Yang, Y. (2011). Scaling up top-k cosine similarity search. *Data & Knowledge Engineering*, 70(1), 60–83. https://doi.org/10.1016/j.datak.2010.08.004.
30. Yang, S. N., Wang, D., Zheng, H., & Jin, R. (2024). TimeRAG: Boosting LLM time series forecasting via retrieval-augmented generation. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). https://doi.org/10.48550/arxiv.2412.16643
31. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., & Silvestri, F. (2024). The power of noise: Redefining retrieval for RAG systems. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. https://doi.org/10.1145/3626772.3657834
32. Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). *Calibrate before use: Improving few-shot performance of language models*. Computation and Language.
33. Lee, J., Qiu, H., Yu, G., & Lin, J. (2007). Rexnord Technical Services IMS Bearing Dataset. NASA Prognostics Data Repository. Retrieved from http://www.imscenter.net
34. Case Western Reserve University Bearing Data Center (2007). .https://engineering.case.edu/bearingdatacenter
35. Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016). Condition Monitoring of Bearing Damage in Electromechanical Drive Systems by Using Motor Current Signals of Electric Motors: A Benchmark Data Set for Data-Driven Classification. *PHM Society European Conference*, *3*(1). https://doi.org/10.36001/phme.2016.v3i1.1577



Pan-Gyun Jeong is a student researcher in the Autonomous Manufacturing & Process R&D Department at the Korea Institute of Industrial Technology (KITECH), Korea. He graduated with a B.S. in Mechanical Engineering from University of Suwon (USW), Korea in 2024 and is now pursuing an M.S. in the HYU–KITECH Joint Department at Hanyang University, Korea.

Yuehua Hu is a Ph.D. student in the joint M.S./Ph.D. program at the Department of Nano Photonics and Nanoelectronics, Hanyang University (HYU), Korea, and the Autonomous Manufacturing and Process R&D Department at the Korea Institute of Industrial Technology (KITECH), Korea. He received his B.S. from Chungnam National University (CNU), Korea in 2023.

Jiyeong Kong is a student researcher in the Autonomous Manufacturing & Process R&D Department at the Korea Institute of Industrial Technology (KITECH), Korea. She received a B.S. in Mechatronics Engineering from Konkuk University, Korea in 2025 and is currently pursuing an M.S. in the Micro/Nano System Department at Korea University, Korea.



Kyungwoo Song is an assistant Professor at Department of Applied Statistics, Yonsei University, Korea. He received BS, MS, and Ph.D. at KAIST, Korea. He worked as an assistant professor at the Department of Artificial Intelligence, University of Seoul, Korea from 2021 to 2023.

Bongyoung Yoo is a Professor at Department of Materials and Chemical Engineering, Hanyang University, Korea. He received his B.S. and M.S. in Metallurgical Engineering from Yonsei University, Korea and his Ph.D. in Materials Science from the University of California, Los Angeles (UCLA), USA.

Kyung-Tae Kang is a principal researcher at Korea Institute of Industrial Technology (KITECH) , Korea and serves as a Secretary of IEC TC119 (printed electronics). He received B.S., M.S. and Ph.D at mechanical engineering department of Seoul National University, Korea in 1988, 1990 and 1994. He worked as a visiting fellow at United Technologies Research Center, USA and Massachusetts Institute of Technology, USA. During his career, he received several prizes including IEC Thomas Edison Award, KSME Technology Award.