10.37965/jdmd.2026.1587

A Task-Based Few-Shot Learning Framework for Wind Turbine Gearbox Fault Diagnosis via Large Language Models

Junhong Chen[2], Shuai Yang[1,*], Xianguang Li[3]

[1]International School, Chongqing Technology and Business University, Chongqing 400067, People's Republic of China

[2]School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, People's Republic of China

[3]School of Mechanical Engineering, Chongqing Technology and Business University, Chongqing 400067, People's Republic of China

*Corresponding author: jerryyang@ctbu.edu.cn

Received 30 June 2026; Revised 28 July 2026; Accepted 18 August 2026; Published online 20 August 2026

Abstract:

Wind turbine gearbox fault diagnosis is essential for ensuring the safe and stable operation of wind energy systems. However, in industrial scenarios, limited fault data and highly complex, non-stationary operating conditions lead to significant distribution shifts, which reduce the generalization ability of traditional methods. To address this issue, we propose a task-oriented fault diagnosis framework based on large language models. Time-domain and frequency-domain features are extracted from vibration signals to construct structured representations. Inspired by few-shot learning, a support–query task construction strategy is introduced, reformulating classification as conditional task reasoning. Low-Rank Adaptation is adopted for parameter-efficient fine-tuning of the pre-trained model. In addition, a global coverage and local balance task sampling strategy is designed to enhance task diversity and mitigate sample imbalance. Experiments on a wind turbine gearbox dataset show consistent performance across multiple backbone models, including LLaMA2-7B, LLaMA3-8B, Qwen-7B, and Baichuan-7B. LLaMA3-8B achieves the best performance, with 98.00% accuracy and an F1-score of 0.9798. These results demonstrate strong robustness and cross-model generalization under few-shot and complex operating conditions.

Keywords: wind Turbine; fault diagnosis; large language model; few-shot learning

1. Introduction

Wind energy is an important renewable energy source, and wind turbines have been widely deployed worldwide[1][2]. However, due to their operation in complex and harsh environments, wind turbines are subject to a relatively high failure rate[3]. These faults often lead to turbine shutdowns, thereby reducing power generation efficiency and causing considerable economic losses[4]. Therefore, the development of efficient and accurate fault diagnosis methods for wind turbines is of great practical significance.

At present, wind turbine fault diagnosis methods can be broadly categorized into experience-based methods, traditional machine learning methods, and deep learning methods[5][6]. Traditional fault diagnosis approaches typically rely on manual feature extraction and domain-specific expert knowledge. Such methods not only involve a labor-intensive feature engineering process but also exhibit limited capability in capturing fault characteristics under complex operating conditions[7]. In recent years, owing to its powerful automatic feature learning capability, deep learning has been widely applied to wind turbine fault diagnosis and has achieved remarkable diagnostic performance[8][9][10]. He[11] et al. proposed a novel Spatio-Temporal Incremental Broad Network with Pseudo-Label Driven Adaptation for fault diagnosis of wind turbine gearboxes under multi-condition, non-stationary operating processes. The effectiveness of the proposed method was further validated on two industrial wind turbine gearbox datasets, demonstrating superior performance and strong potential for practical applications. Du[12] et al. proposed a Dual-Branch Dynamic Convolutional Temporal Attention Network for few-shot fault diagnosis. The method constructs a time–frequency dual-branch architecture incorporating dynamic convolution modules. Experimental results demonstrate the superiority of the proposed model. Ma[13] et al. proposed a novel Spatio-Temporal Graph Convolutional Neural Network (STGCN) based on deep learning. First, a Gaussian kernel is employed to construct an adjacency matrix, thereby graph-structuring Supervisory Control and Data Acquisition (SCADA) data to enhance spatial feature representation. Subsequently, a GCN is used to extract spatial features, while a one-dimensional Convolutional Neural Network is adopted to extract temporal features. Finally, a spatio-temporal feature fusion module is designed to integrate the extracted information, forming a complete STGCN framework.

Deep learning models primarily rely on data-driven pattern matching mechanisms and lack an understanding of fault mechanisms and domain knowledge. In contrast, large language models(LLM) possess strong capabilities in knowledge representation and reasoning[14]. They can integrate fault mechanism knowledge with data-driven features for comprehensive analysis, thereby improving diagnostic reliability under complex operating conditions[15]. Some researchers have begun to explore the application of LLM in the field of fault diagnosis. Tao[16] et al. proposed a large language model-based framework for bearing fault diagnosis, in which LoRA- and QLoRA-based fine-tuning strategies were adopted to enhance the generalization capability of the LLM in analyzing vibration signal features. Experimental results further validated the effectiveness and rationality of the proposed framework. Lin[17] et al. proposed a fault diagnosis method based on multimodal LLM. First, modality alignment training is performed between engineering data description text and operational state features to enhance the model's ability to understand time-series modalities. Second, fuzzy semantic embeddings are introduced to mitigate the issue of pattern overlap in the feature space. Meanwhile, learnable prompt embeddings are employed to inject domain-specific knowledge for fault diagnosis. Finally, a LoRA-based fine-tuning strategy is adopted to optimize model performance. Zhou[18] et al. proposed a novel multimodal fault diagnosis framework that integrates time-series vibration signals with textual descriptions. A BERT-based LLM is employed to enhance feature representations and capture semantic relationships among fault categories. Extensive experiments on industrial datasets validate the effectiveness of the proposed framework.

Despite recent progress in related research, several limitations still remain to be addressed. First, when mechanical equipment fails, it typically requires immediate shutdown for maintenance, resulting in a limited number of available fault samples[19]. However, fine-tuning high-quality LLM generally relies on large volumes of high-quality labeled data, which is difficult to satisfy in real industrial scenarios.Second, real-world industrial equipment often operates under dynamically varying conditions over long periods, and data distributions differ significantly across time, load, and operating environments, leading to pronounced non-stationarity in fault data. LLM may suffer from degraded generalization performance or even model collapse when confronted with fine-grained fault categories and cross-condition distribution shifts.Therefore, how to effectively improve model stability and robustness under few-shot and multi-condition settings remains a critical issue that requires further investigation.

To address the aforementioned issues, this study proposes a LLM-based framework tailored for fault diagnosis tasks. The framework integrates the idea of few-shot learning by constructing task-specific datasets and designing corresponding fine-tuning strategies to enable effective task adaptation of the large model. On this basis, parameter-efficient fine-tuning is performed on the pre-trained LLM using the LoRA method, allowing the model to better learn fault diagnosis-related knowledge and thereby improving its diagnostic capability and generalization performance.

2. Methodology

2.1 Data Preprocessing

Traditional deep learning models can automatically learn and extract abstract fault features from raw data through an end-to-end training paradigm[20]. However, for LLM, it is relatively challenging to directly perform feature modeling from high-frequency raw vibration signals. Given the strong advantages of LLM in semantic information processing, this study extracts interpretable time-domain and frequency-domain features from raw vibration signals and uses them as structured inputs. These features are further organized into large-scale training datasets suitable for downstream fine-

Table 1. Time-Frequency Calculation Formula

$x(n)$: Original signal	
Mean	$\mu = \frac{1}{N}\sum_{n=1}^{N} x(n)$
root mean square	$x_{rms} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x^2(n)}$
standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N} (x(n)-\mu)^2}$
Peak Value	$x_{max} = \max(\lvert x(n)\rvert)$
Peak-to-Peak Value	$x_{p-p} = \max(x(n)) - \min(x(n))$
Crest Factor	$CF = \frac{x_{max}}{x_{rms}}$
Skewness	$x_{sk} = \frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\mu)^3}{\sigma^3}$
Shape Factor	$x_{sf} = \frac{x_{rms}}{\frac{1}{N}\sum_{n=1}^{N}\lvert x(n)\rvert}$
Kurtosis	$x_{ku} = \frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\mu)^4}{\sigma^4}$

Impulse Factor	$x_{IF} = \frac{x_{max}}{\frac{1}{N}\sum_{n=1}^{N}\lvert x(n)\rvert}$

Table 2. Frequency Domain Calculation Formula

$x(n)$: Original signal; $X(k) = FFT\big(x(n)\big)$; f_k: Frequency axis; $A(k) = \lvert X(k)\rvert$; K: Total number of sampling points after the spectrum is discretized	
Peak Frequency	$f_{peak} = f_k \;\; k = argmaxA(k)$
Spectral Peak-to-Peak	$x_{spp} = \max\big(A(k)\big) - \min\,(A(k))$
Spectral Bandwidth	$x_{sb} = \sqrt{\frac{\sum_{k=1}^{K}(f_k - f_c)^2 A(k)}{\sum_{k=1}^{K} A(k)}} \qquad f_c = \frac{\sum_{k=1}^{K} f_k A(k)}{\sum_{k=1}^{K} A(k)}$
Spectral Skewness	$x_{ssk} = \frac{\frac{1}{K}\sum_{k=1}^{K}(A(k) - \bar{A})^3}{(\sqrt{\frac{1}{K}\sum_{k=1}^{K}(A(k) - \bar{A})^2})^3} \qquad \bar{A} = \frac{1}{K}\sum_{k=1}^{K} A(k)$
Spectral Kurtosis	$x_{sk} = \frac{\frac{1}{K}\sum_{k=1}^{K}(A(k) - \bar{A})^4}{(\sqrt{\frac{1}{K}\sum_{k=1}^{K}(A(k) - \bar{A})^2})^4}$

tuning, thereby providing effective support for fault diagnosis tasks based on LLM.

To achieve a comprehensive representation of fault information in vibration signals, this study selects 10 time-domain features and 5 frequency-domain features as descriptive indicators. These features characterize the statistical distribution properties, amplitude fluctuation characteristics, and spectral structure information of the signals from different perspectives. As a result, they can effectively capture the temporal and spectral variation patterns of the signals, thereby providing a solid feature foundation for subsequent fault diagnosis. Time-domain features are primarily used to characterize the statistical distribution properties and impact response characteristics of vibration signals. The ten time-domain features selected in this study include the mean, root mean square, standard deviation, Peak Value, peak-to-peak value, crest factor, skewness, shape factor, kurtosis and impulse factor. These features can capture the energy distribution patterns and transient impact characteristics of the signal from multiple perspectives and are highly sensitive to mechanical fault conditions. The specific formulas are shown in Table 1. In the frequency-domain analysis, the windowed signals are first processed using a Hanning window to reduce spectral leakage effects. Subsequently, the Fast Fourier Transform is applied to obtain the spectral information of the signal. Based on this, several frequency-domain features are further extracted, including peak frequency, spectral peak-to-peak value, spectral bandwidth, spectral kurtosis and spectral skewness. The specific formulas are shown in Table 2. These features are used to describe the distribution, concentration, and variation patterns of signal energy in the frequency domain.

2.2 Task-oriented fine-tuning dataset construction.

Within the meta-learning framework, the quality and distribution of training tasks directly determine the generalization performance of the inductive bias learned by the model[21]-[22]. Although traditional random sampling strategies can ensure an unbiased task distribution, they tend to suffer from task redundancy, class imbalance, and excessive reuse of support samples in data-limited scenarios[21]. These issues may further lead to representation collapse and overfitting in the learned feature space. To address this issue, this study proposes a task-oriented dataset construction method. The core idea is inspired by few-shot learning (FSL), where limited samples are randomly combined and restructured into diverse training tasks. In this way, the model is exposed to a wide range of task distributions even under low-data conditions, enabling it to learn task-level representations and thereby acquire modeling and reasoning capabilities for specific fault diagnosis tasks.

As shown in Fig. 1, each task consists of two components: a support set and a query set. The support set provides labeled samples along with their corresponding category information, serving as prior knowledge for class reference and pattern learning. The query set contains unlabeled samples to be identified, with their labels hidden during the inference stage. The model learns class-specific representations from the support set and performs similarity or relational matching between the support set and query set samples. Based on this comparison, the model is able to infer and identify the fault category of each query sample. Its mathematical formulation is given as follows:

Support set:

$$S = \{(x_i, y_i)\}_{i=1}^{N \cdot K} \tag{1}$$

Query set:

$$Q = (x_q, y_q) \tag{2}$$

The core significance of this task structure lies in transforming the learning objective from conventional “class boundary function” modeling to a “conditional task reasoning function” paradigm:

$$f_\theta : (S, x_q) \rightarrow y_q \tag{3}$$

From a theoretical perspective, the proposed task reasoning paradigm differs from traditional metric learning methods in terms of its generalization mechanism. Conventional metric learning methods typically construct a fixed feature embedding space and perform classification based on predefined class boundaries, where their generalization capability largely depends on the representation capacity of the learned feature space. However, under complex operating conditions and few-shot scenarios, fixed class boundaries are vulnerable to variations in data distributions, resulting in degraded generalization performance. In contrast, the conditional task reasoning framework proposed in this study shifts the source of generalization capability from the feature space level to the task level. Instead of relying on fixed global classification rules, the model dynamically adjusts its decision criteria according to the contextual information provided by the support set in each task. As the number and diversity of training tasks increase, the model is able to learn richer task distribution characteristics, thereby improving its adaptability to unseen operating conditions and few-shot scenarios. This conditional reasoning-based learning mechanism provides theoretical support for the stability and generalization capability of the proposed method in complex industrial fault diagnosis applications.

Specifically, the model no longer learns a fixed mapping from data to class labels. Instead, it dynamically constructs class decision criteria based on the contextual information provided in the support set of each task, and performs inference and classification on query samples accordingly. Therefore, the LLM is required to achieve an adaptive decision-making process conditioned on the given support context, rather than relying on a fixed classifier architecture for mapping.

During the training phase, if the task generation process relies entirely on random sampling, it is likely to produce a large number of duplicate tasks. In addition, certain classes may frequently appear as query sets, leading to class imbalance and even representation collapse, which in turn degrades overall training efficiency and

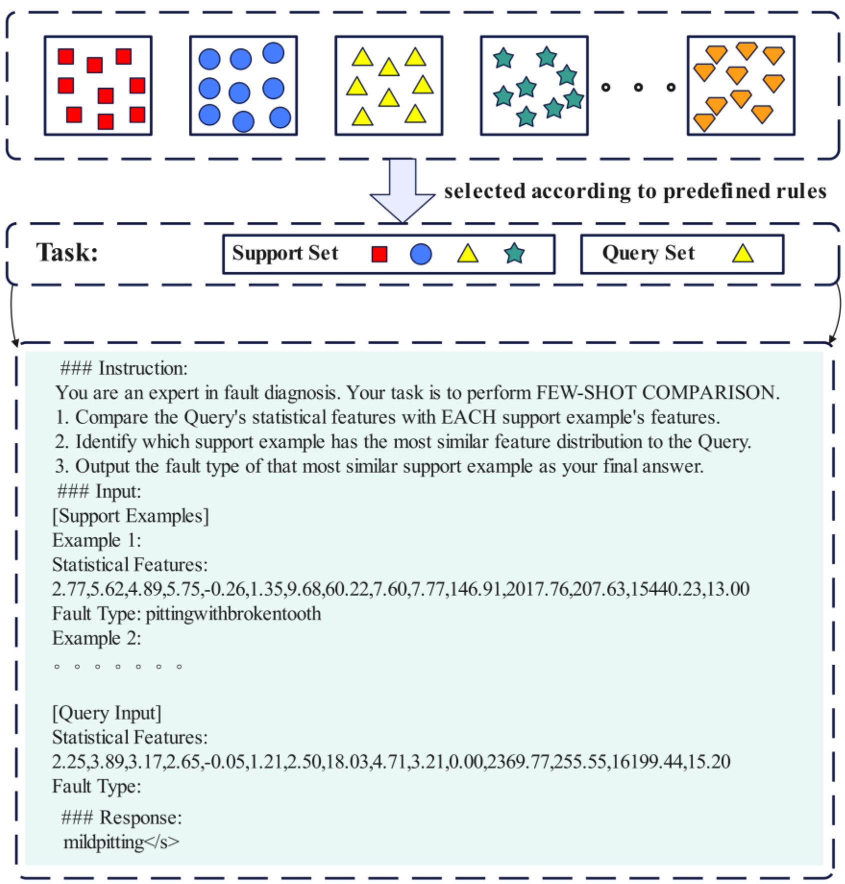

Fig. 1 Task-Based Fault Diagnosis Dataset

generalization performance.

To address these issues, this study proposes a "global coverage + local balance" task generation strategy, which consists of the following three components:

(1) Global query space coverage

All training windows are treated as potential query candidates, thereby maximizing task-space coverage and enabling the model to learn the full data distribution boundaries.

(2) Structured enumeration of class combinations

All possible N-way class subsets are generated through combinatorial enumeration, and valid combinations containing the query class are selected, ensuring explicit structural diversity in the task space.

(3) Balanced sampling mechanism for support sets

During support sample selection, a usage-frequency constraint is introduced, assigning higher sampling priority to less frequently used samples. This prevents overfitting to high-frequency samples and promotes balanced utilization of the dataset.

In addition, to further reduce structural redundancy in the task space, this study introduces a task uniqueness constraint mechanism. A structured representation (task signature) is defined for each task.

$$T = (y_q, x_q, sorted(S)) \tag{4}$$

It should be noted that Equation (4) is a conceptual representation for clarity. In the actual implementation, to avoid the numerical precision issues caused by directly comparing floating-point feature vectors, the deduplication mechanism is realized using the unique integer indices assigned to each sample during the preprocessing stage. Specifically, the uniqueness of a task is determined by the combination of the query index, the query label, and the sorted set of support indices, rather than by comparing the continuous feature values themselves. This index-based strategy ensures deterministic and exact deduplication at the combinatorial structure level.

This mechanism not only avoids redundant sampling at the sample level but also ensures task uniqueness and independence at the combinatorial structural level. As a result, it increases the diversity and information entropy of the task distribution, enabling the model to learn richer and more diverse task reasoning paths.

During the testing phase, this study further strengthens the constraints on task construction to ensure the authenticity and generalization capability of the evaluation results while preventing label leakage. The core principle is to strictly decouple the test queries from the support context.

Specifically, query samples are selected exclusively from the independent test dataset, while support samples are obtained only from the training dataset. This means that no samples from the test set are involved in constructing the support set, and the model cannot obtain any prior information related to the test samples through the support set before making predictions. If test samples were allowed to participate in support set construction, the fault category information and feature distribution associated with the test samples could be directly provided to the model through the support samples, resulting in an overestimation of model performance. By strictly constraining the sources of query and support samples, this study prevents potential label information leakage and ensures that the model performs fault reasoning solely based on the limited prior knowledge provided by the support set. Therefore, this evaluation strategy can more objectively reflect the few-shot reasoning capability and generalization performance of the proposed framework under unseen testing conditions.

In addition, at the testing stage, multiple task instances are constructed for each query sample. In other words, each test sample is repeatedly evaluated under different support-set compositions. This strategy essentially places the same query sample within diverse contextual supports for multiple rounds of inference. By aggregating the results across these repeated evaluations, the randomness introduced by a single sampling process can be effectively reduced, leading to more stable and statistically meaningful performance estimates.

This mechanism essentially establishes a "conditional perturbation evaluation framework," where the model's stability and robustness across different task contexts are assessed by varying the compositional structure of the support set.

The characteristics of industrial fault

diagnosis scenarios are considered in this study, where fault samples are typically limited and only a small number of labeled samples can be obtained. Therefore, a 4-way 1-shot task setting is adopted. Specifically, each task contains four fault categories (N=4), and only one labeled sample per category is provided in the support set (K=1). The 4-way configuration enables the model to distinguish multiple fault categories within a single task while maintaining sufficient task diversity. Furthermore, considering the prompt-based input mechanism of large language models, increasing the number of support samples or categories may introduce redundant information and increase the input sequence length, thereby leading to higher computational costs. Therefore, the 4-way 1-shot setting achieves a favorable balance among task complexity, data scarcity, and prompt length control.

The overall framework flowchart is shown in Fig. 2.

2.3 LoRA Fine-Tuning

In the context of wind turbine fault diagnosis based on LLM, the model is required to possess two key capabilities. Semantic understanding of complex industrial signals, and task reasoning and generalization ability under few-shot conditions. Although traditional full-parameter fine-tuning can improve performance on specific tasks, it suffers from significant limitations in practical industrial applications, including high computational cost, training instability, and strong sensitivity to small-scale datasets. Therefore, there is a need to introduce a more efficient and stable parameter adaptation method[23].

In this study, LoRA fine-tuning is adopted to train the large model[24]. Low-

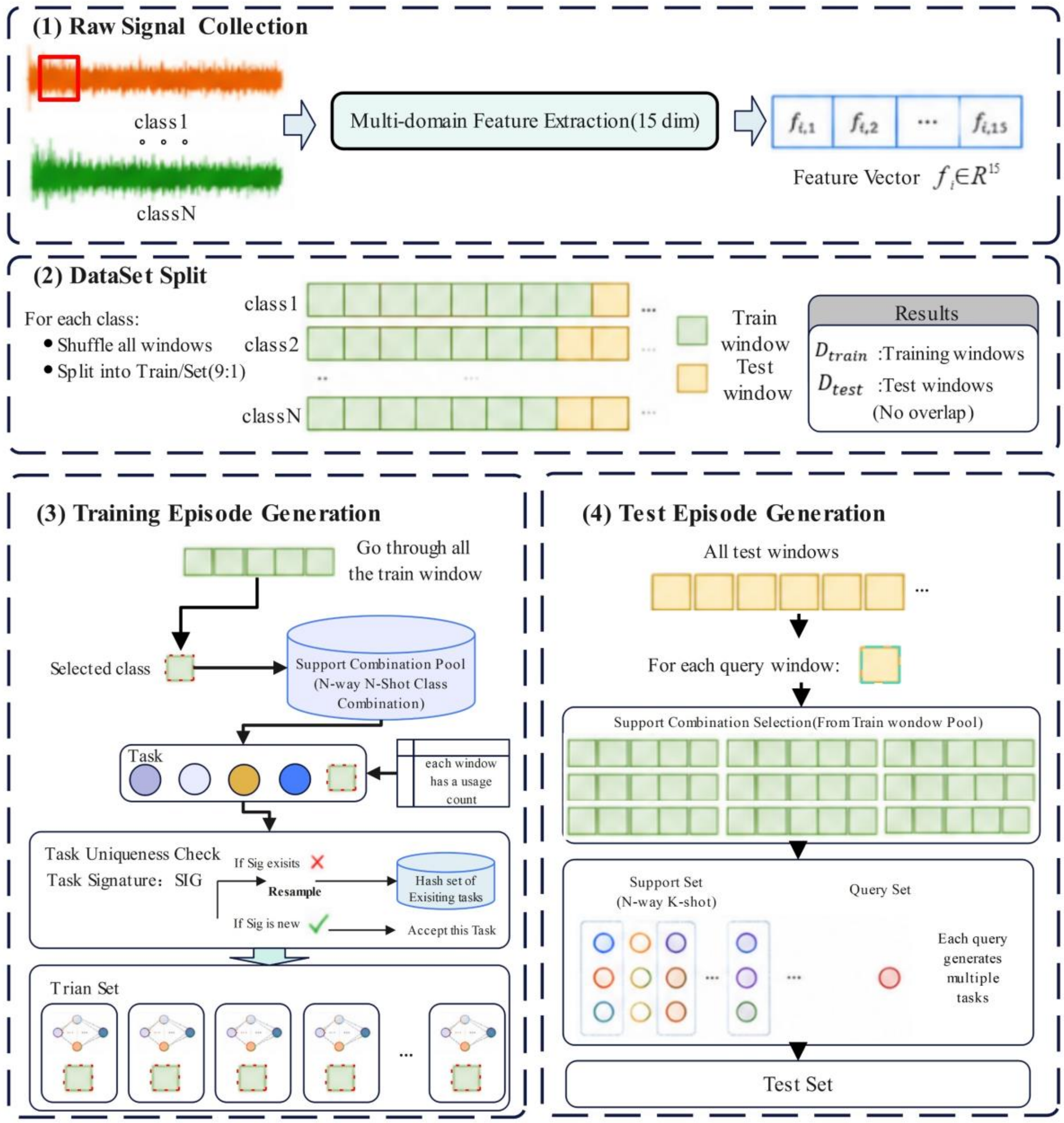

Fig. 2 Dataset Construction Flowchart

Rank Adaptation is a representative parameter-efficient fine-tuning method. As shown in Fig. 3, its core idea is to freeze the parameters of the pre-trained backbone model and model the weight updates through the introduction of low-rank matrices, enabling task adaptation without modifying the original model architecture. Unlike traditional fine-tuning methods, LoRA does not directly update the original weight matrices; instead, it constrains the weight updates within a low-dimensional subspace. This design allows the model to adapt effectively to downstream tasks by training only a small number of parameters, while preserving the general knowledge and language capabilities learned during pre-training to the greatest extent.

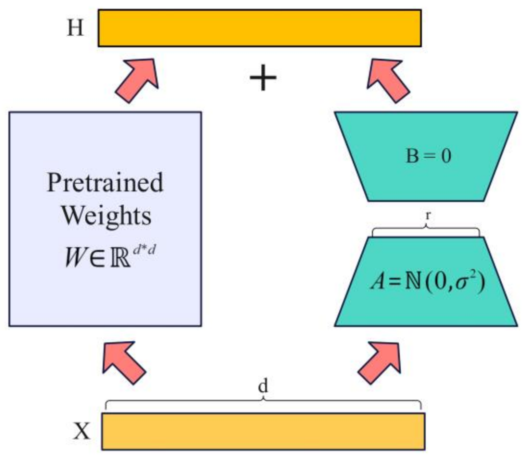

Fig. 3. LoRA Fine-Tuning Process

3. Experiment

3.1 Data

To obtain real and reliable wind turbine fault data, this study established an experimental platform based on an RCVA-3000 wind turbine system and constructed a fault data acquisition system at Chongqing Technology and Business University to conduct simulation experiments. The experimental setup mainly consists of a wind turbine, a blower, a data acquisition system, vibration sensors, and a host workstation.

As shown in Fig. 4, the airflow generated by the blower is used to simulate natural wind conditions, thereby driving the rotation of the turbine blades and further actuating the gearbox system. In all experiments, the wind turbine under different fault conditions is uniformly operated at a rotational speed of 30 Hz under minimal load conditions to ensure the comparability and consistency of the collected data.

Before each experiment, the surfaces of key components are cleaned to remove residual metal debris and other interfering factors, and the predefined fault modes are verified. After each experimental run, the collected data are preliminarily analyzed and inspected to ensure data quality, which guides the subsequent experimental design. Prior to data acquisition, newly installed components undergo a pre-operation process to eliminate initial transient instability, reduce power interference and other external noise effects, and thereby improve the reliability and stability of the experimental data.

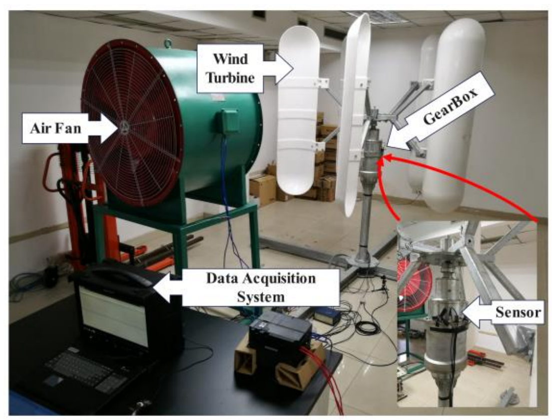

Fig. 4. Experimental Setup of The Wind Turbine System

This study takes the sun gear of a wind turbine gearbox as the primary research object. A total of five operating conditions

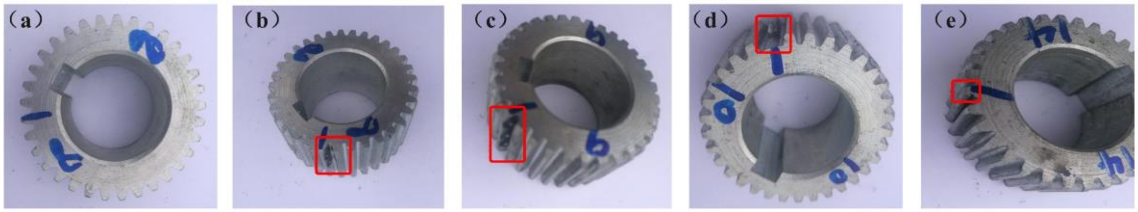

Fig. 5. Fault Details: (a) Healthy; (b) Light Pitting; (c) Medium Pitting; (d) Partial Tooth Breakage; (e) 0.3 mm × 0.5 mm Worn.

Table 3. Detailed Operating Conditions

Health		Load		
		L1	L2	L3
Speed	30Hz	C00L1S30	C00L2S30	C00L3S30
	50Hz	C00L1S50	C00L2S50	C00L3S50
Light Pitting		Load		
		L1	L2	L3
Speed	30Hz	C08L1S30	C08L2S30	C08L3S30
	50Hz	C08L1S50	C08L2S50	C08L3S50
Medium Pitting		Load		
		L1	L2	L3
Speed	30Hz	C09L1S30	C09L2S30	C09L3S30
	50Hz	C09L1S50	C09L2S50	C09L3S50
Partial Tooth Breakage		Load		
		L1	L2	L3
Speed	30Hz	C10L1S30	C10L2S30	C10L3S30
	50Hz	C10L1S50	C10L2S50	C10L3S50
0.3 mm × 0.5 mm Worn		Load		
		L1	L2	L3
Speed	30Hz	C14L1S30	C14L2S30	C14L3S30
	50Hz	C14L1S50	C14L2S50	C14L3S50

are considered, including one healthy state and four typical fault states. Specifically, the four fault types include mild pitting, moderate pitting, partial tooth breakage, and wire-cut damage of 0.3 mm × 0.5 mm. The detailed morphologies of each fault sample are shown in Fig. 5.

These fault types cover the typical degradation and damage modes that may occur during the actual operation of gear systems. They can comprehensively reflect the dynamic response characteristics of the gearbox sun gear under different health conditions, thereby providing a reliable data foundation for the construction and validation of subsequent fault diagnosis models.

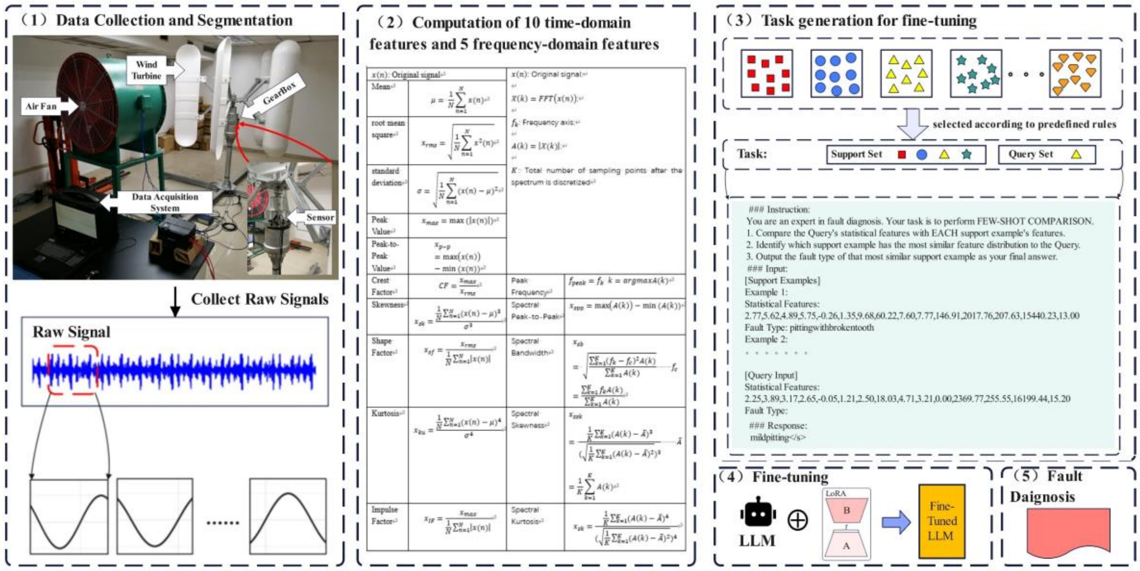

Fig. 6. Framework Flowchart

In addition, two different rotational speeds and three different load conditions are included. The specific combinations of fault categories under each operating condition are listed in Table 3

3.2 Experimental procedure

The overall framework of this study is illustrated in Fig. 6. First, raw fault vibration signals are collected using the established wind turbine experimental platform to provide fundamental data support for subsequent data-driven modeling. Then, a two-stage data partitioning strategy is applied: in the first stage, the dataset is split into training and testing subsets in chronological order to ensure strict temporal independence between them and avoid information leakage caused by temporal correlations; in the second stage, a sliding window approach is further employed to segment each subset. Both the window length and stride are set to 1024, enabling continuous time-series signals to be divided into a series of independent segments with local statistical characteristics.

In the resulting dataset, the training set contains 30 operating conditions, with 9 independent signal frames per class, yielding a total of 270 samples. The test set also includes the same 30 categories, consisting of 30 independent signal frames in total, ensuring complete and consistent class coverage. Subsequently, time–frequency features are extracted from each signal segment, and task samples are constructed according to predefined task generation rules to achieve a mapping from raw signals to task-level data.

During task construction, considering the limited availability of labeled fault samples in practical wind turbine systems, this study adopts a 4-way 1-shot few-shot learning setting. Specifically, each task contains four different fault categories, and only one labeled sample from each category is provided in the support set. The 4-way configuration increases task-level classification difficulty compared with binary tasks, while the 1-shot setting reflects the realistic condition where only limited labeled fault samples are available. Therefore, this configuration provides a reasonable trade-off between task complexity and data scarcity. 20,000 training tasks are generated from the training set to enhance the model's ability to learn diverse task compositions and improve generalization. For the test set, 5 tasks are generated for each independent signal frame, resulting in a total of 150 test tasks to evaluate the model's stability and robustness under different contextual conditions.

Finally, the constructed task-based dataset is fed into the model for training and testing, thereby completing the training and performance evaluation of the fault diagnosis model under a task-learning paradigm.

3.3 Evaluation metrics

To validate the performance of the proposed framework, the assessment metrics employed are Accuracy, Precision, Recall, and F1-Score.

The detailed formulations of each evaluation metric are provided in Table 4.

Table 4. Evaluation-Metric Formulas

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

where True Positive (TP) represents positively classified samples with correct identification, True Negative (TN) represents properly classified negative instances, False Positive (FP) represents misclassified positive samples, and False Negative (FN) represents negative with incorrect case classification.

Since this study focuses on a multi-class fault diagnosis task rather than a binary classification problem, the Precision, Recall, and F1-score metrics are calculated using the one-vs-rest strategy. Specifically, for each fault category, the corresponding category is treated as the positive class, while all other categories are regarded as the negative class. The Precision, Recall, and F1-score are then calculated separately for each category. Finally, the overall evaluation results are obtained through weighted averaging, where the contribution of each category is determined according to its number of samples.

4. Result

4.1 Generality Evaluation Across Different LLM Backbones

This study selects four LLM as backbone models for comparative experiments, including LLaMA2-7B[25], LLaMA3-8B[26], Qwen-7B[27], and Baichuan-7B[28]. Through systematic experiments on different base models, the adaptability and stability of the proposed framework across various pre-trained models are further analyzed. The corresponding parameter settings are listed in Table 5.

The experimental results are shown in Table 6 and Fig. 7. After training, the

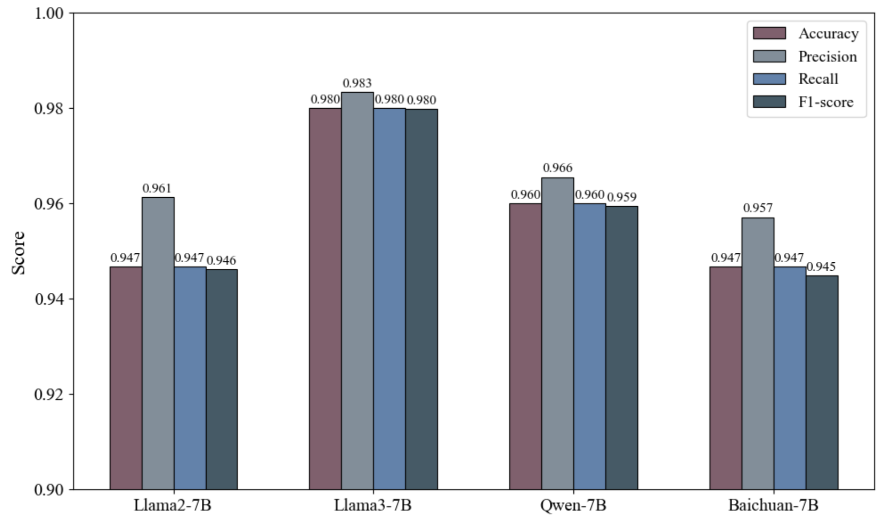

Fig. 7. Performance Metric

Table 5. Fine-Tuning Parameter Settings

Llama2-7B		Qwen-7B	
Stage	SFT	Stage	SFT
lora_rank	64	lora_rank	128
lora_target	q_proj,k_proj,v_proj,o_proj,up_proj,down_proj,gate_proj	lora_target	q_proj,k_proj,v_proj,o_proj,up_proj,down_proj
train_batch_size	32	train_batch_size	32
Learning Rate	1e-4	Learning Rate	5e-4
LR Scheduler	constant	LR Scheduler	constant
Optimizer	AdamW	Optimizer	AdamW
Epochs	3	Epochs	3
Llama3-8B		**Baichuan-7B**	
Stage	SFT	Stage	SFT
lora_rank	128	lora_rank	128
lora_target	q_proj,k_proj,v_proj,o_proj	lora_target	q_proj,k_proj,v_proj,o_proj,up_proj,down_proj,gate_proj
train_batch_size	32	train_batch_size	16
Learning Rate	1e-4	Learning Rate	1e-5
LR Scheduler	constant	LR Scheduler	constant
Optimizer	AdamW	Optimizer	AdamW
Epochs	3	Epochs	3

Table 6. Performance Metrics

	Accuracy	Precision	Recall	F1-Score
Llama2-7B	0.9467	0.9613	0.9467	0.9462
Llama3-8B	0.9800	0.9833	0.9800	0.9798
Qwen-7B	0.9600	0.9655	0.9600	0.9594
Baichuan-7B	0.9467	0.9571	0.9467	0.9448

classification performance of each backbone model on the test set is as follows:

LLaMA2-7B: Accuracy: 0.9467, Precision: 0.9613, Recall: 0.9467, F1-score: 0.9462

LLaMA3-8B: Accuracy 0.9800, Precision:0.9833, Recall:0.9800, F1-score:0.9798;

Qwen-7B: Accuracy:0.9600, Precision: 0.9655, Recall: 0.9600, F1-score: 0.9594;

Baichuan-7B: Accuracy: 0.9467, Precision: 0.9571, Recall: 0.9467, F1-score: 0.9448.

The confusion matrices are shown in Fig. 8

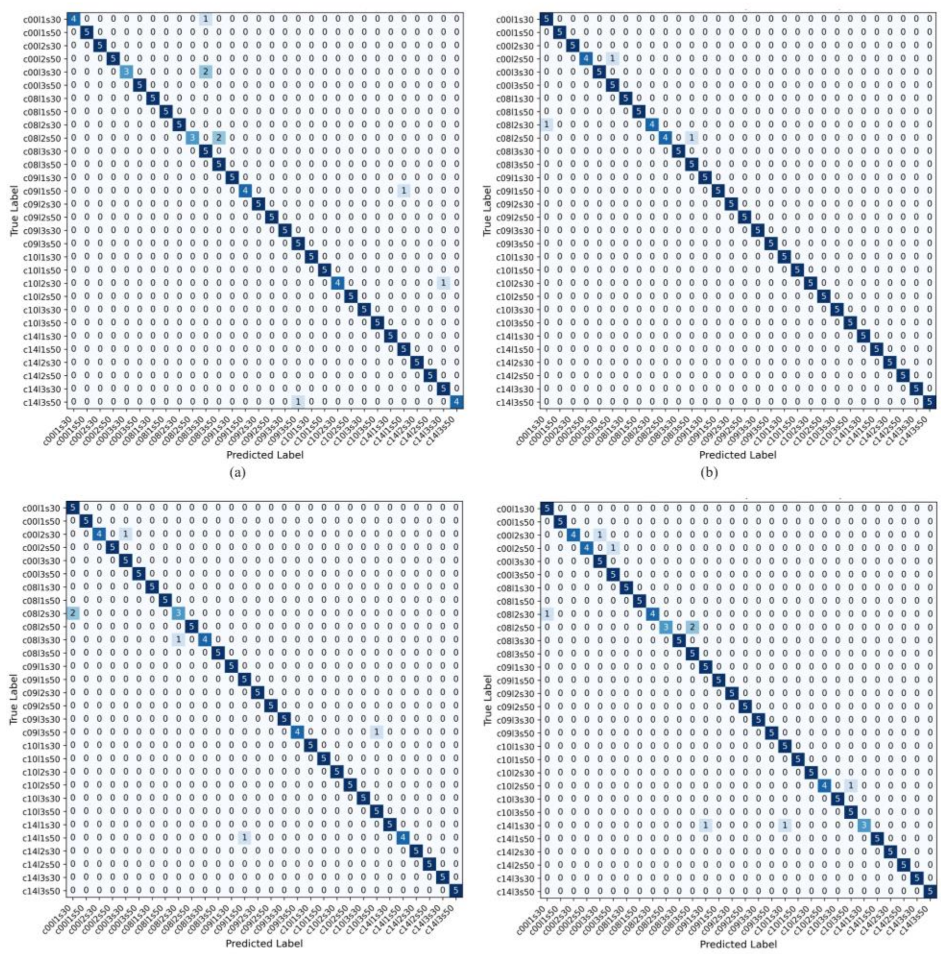

Fig. 8. Confusion Matrices of Different Models((a) Llama2-7B; (b) Llama3-8B; (c) Qwen-7B; (d) Baichuan-7B)

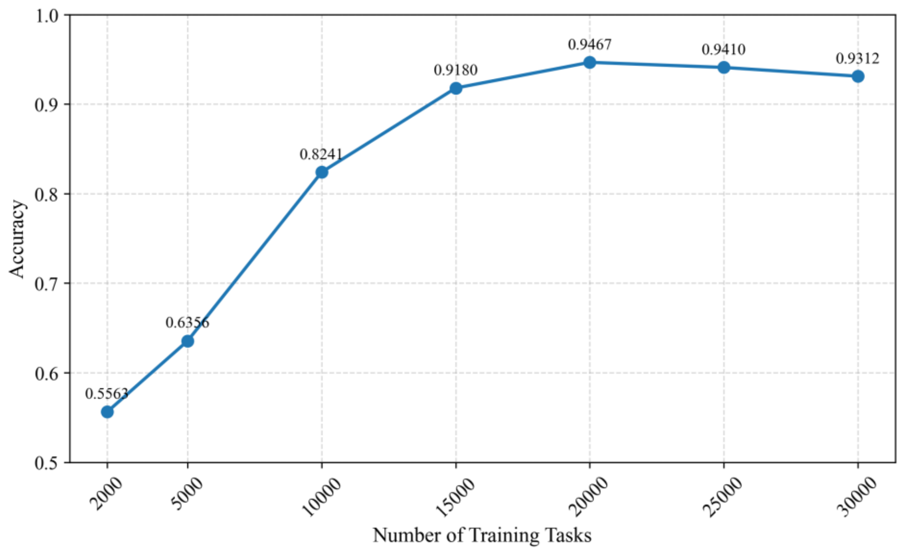

Fig. 9. Effect of Training Task Number on Diagnostic Performance

The experimental results demonstrate that the proposed framework consistently achieves stable and high performance across different backbone LLM. Among them, LLaMA3-8B achieves the best performance, with an accuracy of 0.9800 and an F1-score of 0.9798. The relatively small performance variations among different models indicate that the proposed method exhibits strong robustness and favorable cross-model adaptability to various pre-trained language models.

4.2 Effect of Training Task Number

To investigate the impact of the number of generated training tasks on model performance, an additional experiment was conducted by varying the number of training tasks while keeping all other training configurations unchanged. The numbers of training tasks were set to 2,000, 5,000, 10,000, 15,000, 20,000, 25,000, and 30,000, respectively. The model selected in this experiment was LLaMA2-7B.

Table 7. Effect of Training Task Number on Diagnostic Performance

Task number	Accuracy
2000	0.5563
5000	0.6356
10000	0.8241
15000	0.9180
20000	**0.9467**
25000	0.9410
30000	0.9312

The experimental results are presented in Table 7 and Fig. 1. The diagnostic accuracy increases significantly as the number of training tasks increases from 2,000 to 20,000. Specifically, the accuracy improves from 55.63% with 2,000 tasks to 94.67% with 20,000 tasks. This indicates that increasing the number of generated tasks provides richer task-level diversity, enabling the model to learn more comprehensive support–query relationships and improve its few-shot reasoning capability.

However, when the number of

Table 8. Comparative Experiments

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.7743	0.7854	0.7743	0.7787
Transformer	0.8067	0.8174	0.8067	0.8036
MLP	0.8317	0.8453	0.8232	0.8342
PrototypicalNet	0.8689	0.8623	0.8651	0.8549
RelationNet	0.8480	0.8494	0.8434	0.8462
Llama3-8B(ours)	**0.9800**	**0.9833**	**0.9800**	**0.9798**

training tasks is further increased to 25,000 and 30,000, the performance does not continue to improve but slightly decreases to 94.10% and 93.12%, respectively. This phenomenon suggests that when the number of tasks reaches 20,000, the task distribution has already been sufficiently explored, and additional task generation mainly introduces redundant task information rather than providing substantial new knowledge. Therefore, generating 20,000 training tasks achieves a reasonable balance between task diversity and computational efficiency.

Based on the above analysis, this study selects 20,000 training tasks to ensure sufficient task diversity while avoiding unnecessary computational overhead.

4.3 Comparison with Existing Models

To further validate the effectiveness of the proposed framework, several representative fault diagnosis methods are selected as baseline approaches for comparative experiments. All comparison methods are trained and tested using the same data partition strategy and experimental configurations to ensure the fairness and reliability of the experimental results.

The selected baseline methods mainly include convolutional neural network based methods[29], Transformer-based methods[30], multilayer perceptron based methods, and two classical few-shot learning models, namely Prototypical Network[31] and Relation Network[32]. These methods represent typical approaches in traditional deep learning, attention-based modeling, and few-shot learning, respectively, providing a comprehensive evaluation of the performance advantages of the proposed framework.

The experimental results are presented in Table 8. It can be observed that the proposed method achieves the best performance in terms of Accuracy, Precision, Recall, and F1-score, outperforming all other comparison methods. These results demonstrate that the proposed task-oriented large language model framework can effectively utilize task contextual information and achieve more accurate and stable fault identification performance under few-shot fault diagnosis scenarios, thereby verifying the effectiveness and superiority of the proposed method.

Table 9. Ablation Experimental Results

Model	Accuracy	Precision	Recall	F1-Score
Base LLM	0.3525	0.3632	0.3525	0.3516
LoRA	0.7751	0.7642	0.7742	0.7712
LoRA + Random Task	0.8871	0.8936	0.8871	0.8869
Proposed	0.9467	0.9613	0.9467	0.9462

4.4 Ablation Study of the Proposed Framework

To further validate the effectiveness of each key component in the proposed framework, an ablation study was conducted. By progressively removing or replacing the key designs in the framework, the fault diagnosis performance under different model configurations was compared. All ablation experiments were performed using the same data partition strategy, training scheme, and evaluation metrics to ensure the fairness of the experimental results. In these experiments, LLaMA2-7B was selected as the Base LLM. Specifically, four model variants were designed as follows:

(1) Base LLM: The pre-trained large language model is directly applied to the fault diagnosis task without any parameter updates. This model is used to evaluate the intrinsic capability of the original LLM in industrial fault diagnosis scenarios.

(2) LoRA: The LoRA parameter-efficient fine-tuning strategy is introduced on the basis of the Base LLM, enabling the model to adapt to the data distribution of the fault diagnosis domain. This variant is used to analyze the impact of parameter-efficient fine-tuning on performance improvement.

(3) LoRA + Random Task: Based on LoRA fine-tuning, a random task construction strategy is adopted to investigate the influence of task construction methods on the model's task-level learning capability.

(4) Proposed: The complete proposed framework, including the task-oriented dataset construction strategy, the 4-way 1-shot task setting, and the LoRA fine-tuning strategy.

As shown in Table 9. Ablation Experimental Results, when the Base LLM is directly applied to the fault diagnosis task, it achieves an accuracy of only 0.3525, indicating that the large language model without domain adaptation cannot effectively understand the specialized features involved in industrial fault diagnosis. After introducing LoRA fine-tuning, the accuracy increases to 0.7751, demonstrating that parameter-efficient fine-tuning can effectively enhance the model's domain adaptation capability.

Furthermore, after incorporating the task construction strategy, the accuracy of the LoRA + Random Task model is improved to 0.8871, verifying that task-level training can enhance the model's ability to learn support–query relationships. However, due to the lack of systematic task coverage in random task generation, its performance remains inferior to that of the complete framework. Finally, the proposed method achieves the best performance, with Accuracy, Precision, Recall, and F1-score reaching 0.9467, 0.9613, 0.9467, and 0.9462, respectively, representing a 5.96 percentage-point improvement over the LoRA + Random Task model.

These results demonstrate that LoRA fine-tuning effectively improves the model's domain adaptation capability, while the task-oriented few-shot task construction strategy further enhances the model's few-shot reasoning and generalization ability. The combination of these two components enables the proposed framework to achieve more accurate fault diagnosis under limited labeled data conditions.

5. Conclusions

This study proposes a task-oriented fault diagnosis framework for wind turbine gearboxes based on LLM, aiming to address the challenges of limited fault data and complex multi-condition operating environments in industrial scenarios. The proposed method first extracts time-domain and frequency-domain features from raw vibration signals and constructs structured input representations to adapt to the learning paradigm of LLM. Inspired by few-shot learning, a support–query task construction mechanism is designed, which reformulates fault diagnosis from a traditional fixed classification mapping problem into a conditional task reasoning process.

To improve training efficiency and model adaptability, Low-Rank Adaptation is employed for parameter-efficient fine-tuning of pre-trained LLMs. Meanwhile, a "global coverage and local balance" task generation strategy is introduced to enhance task diversity while alleviating data redundancy and class imbalance issues. In addition, a strict training–testing decoupling design is adopted to effectively prevent information leakage, thereby ensuring the objectivity and reliability of the evaluation results.

Experimental results on a wind turbine gearbox dataset involving multiple fault types, rotational speeds, and load conditions demonstrate that the proposed method achieves stable and superior performance across various backbone models, validating its effectiveness and applicability in few-shot and complex multi-condition industrial scenarios. Moreover, the consistent performance across different models further indicates strong cross-model robustness and generalization capability.

However, the proposed method still has limitations in interpretability, as the internal decision-making mechanism of the model lacks sufficient transparency and traceability. Future work will focus on analyzing the interpretability of the model reasoning process and further improving its efficiency and deployability in real-time industrial environments. In addition, more fine-grained task construction and data augmentation strategies will be explored to enhance the model's generalization ability and stability under extreme few-shot and highly non-stationary conditions, thereby facilitating its practical engineering deployment.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (51905058), New Chongqing Youth Innovative Talent Project (CSTB2024NSCQ-QCXMX0050), Chongqing Natural Science Joint Fund of Innovation and Development (Chongqing Municipal Education Commission, CSTB2024NSCQ-LZX0122), the Science and Technology Research Program of Chongqing Municipal Education Commission (KJZD-M202500802), the Venture and Innovation Support Program for Chongqing Overseas Returnees (cx2024067) and Chongqing Municipal Special Fund of Technological Innovation and Application Development (CSTB2025TIAD-STX0039, CSTB2024TIAD-KPX0088).

References

[1]. M. Li, Z. Xu, S. Li, et al., “Health prognostics and maintenance decision-making for wind energy: A comprehensive overview,” Renewable and Sustainable Energy Reviews, **Vol** 226, p. 116269, 2026.

[2]. N. Alagha, A.S.M. Khairuddin, Z.N. Haitaamar, et al., “Artificial intelligence in wind turbine fault detection and diagnosis: Advances and perspectives,” Energies, **Vol** 18, no. 7, p. 1680, 2025.

[3]. S. Zhang, Y. He, Y. Gu, et al., “UAV based defect detection and fault diagnosis for static and rotating wind turbine blade: a review,” Nondestructive Testing and Evaluation, **Vol** 40, no. 4, p. 1691-1729, 2025.

[4]. R. Jiang, F. Fang, J.J. Rodríguez-Andina, et al., “Artificial intelligence in wind turbine fault diagnosis: A systematic knowledge mapping and trend analysis,” Renewable and Sustainable Energy Reviews, **Vol**. 226, p. 116321, 2026.

[5]. Wang C, Wan S, Zhang X, et al. Research on a multi-sensor information fusion method for wind turbine gearbox fault diagnosis based on the BLSCFN model. Measurement Science and Technology, Vol. 36, no. 1, p. 015126. , 2025

[6]. S. Yang and R. Liu, “A review of graph neural networks for rolling bearing fault diagnosis,” Measurement Science and Technology, **Vol**. 37, p. 022002, 2026.

[7]. P. Jiang, Y. Wang, S. Wu, et al., “Fault diagnosis of wind turbine pitch bearings via transfer learning and an improved residual network,” Measurement, **Vol**. 254, p. 117985, 2025.

[8]. H.F. Xu, S. Yang, J.H. Chen, and C. Li, “Fault diagnosis of wind turbines using relation network and margin learning method based on acoustic signals with small-scale data,” Measurement, **Vol**. 257, p. 118741, 2026.

[9]. X. Li, B. Kang, J. Tang, et al., "A Single-Device Environment-Adaptive Mixed Reality Framework for Real-Time Industrial Fault Diagnosis," Journal of Dynamics, Monitoring and Diagnostics, **Vol**. 5, no. 1, p. 64-73, 2026.

[10]. J. Qi, Y. Kong, Y. Lv, et al., "Order spectrum-assisted sparse learning classification approach for wind turbine drivetrain fault diagnostics under variable operating conditions," Renewable Energy, **Vol**. 253, p. 123659, 2025.

[11]. Y. He, X. Li, L. Qian, et al., "A novel fault diagnosis method for wind turbine gearbox based on data-driven incremental broad network," Engineering Applications of Artificial Intelligence, **Vol**. 168, p. 114050, 2026.

[12]. Y. Du, B. Sun, B. Ma, et al., " DBANet: A Dual-Branch Dynamic Convolutional Temporal Attention Network for Few-Shot Wind Turbine Bearing Fault Diagnosis," Expert Systems with Applications, 2026, p. 131361.

[13]. J. Ma, Y. Fu, T. Cheng, et al., "SCADA data-driven spatio-temporal graph convolutional neural network for wind turbine fault diagnosis," IEEE Transactions on Instrumentation and Measurement, 2025.

[14]. X. Guo, Y. Li, S. Cui, et al., "Knowledge-guided multimodal LLM with mixture-of-experts for interpretable defect detection and fault diagnosis: An industrial AI copilot," Journal of Manufacturing Systems, **Vol**. 85, p. 662-677, 2026.

[15]. W. Du, X. Yu, Z. Guo, et al., "Channel attention residual transfer learning with LLM fine-tuning for few-shot fault diagnosis in autonomous underwater vehicle propellers," Ocean Engineering, **Vol**. 330, p. 121237, 2025.

[16]. L. Tao, H. Liu, G. Ning, et al., "LLM-based framework for bearing fault diagnosis," Mechanical Systems and Signal Processing, **Vol**. 224, p. 112127, 2025.

[17]. L. Lin, S. Zhang, S. Fu, et al., "FD-LLM: Large language model for fault diagnosis of complex equipment," Advanced Engineering Informatics, **Vol**. 65, p. 103208, 2025.

[18]. J. Zhou, Y. Guo, Z. Yang, et al., "T2MFDF: A LLM-Enhanced Multimodal Fault Diagnosis Framework Integrating Time-series and Textual Data," IEEE Transactions on Instrumentation and Measurement, 2025.

[19]. S. Yang, H. Sun, J. Chen, et al., "Fault diagnosis of wind turbine gearboxes based on an interpretable multi-source signal fusion analytical framework," Measurement, 2026, p. 121909.

[20]. Y. Zhang, C. Shen, X. Ye, et al., "A dual-component elastic adaptation network for rotating machinery incremental fault diagnosis under variable operating conditions," Journal of Dynamics, Monitoring and Diagnostics, **Vol**. 5, no. 1, p. 49-63, 2026.

[21]. J. Gupta, M. Mehrotra, A. Aggarwal, et al., "Meta-learning frameworks in lung disease detection: a survey," Archives of Computational Methods in Engineering, **Vol**. 33, no. 1, p. 1029-1059, 2026.

[22]. A. Alsaleh, E. Albalawi, S. Alsalman, et al., "Meta-Learning for Medical Image Segmentation: A Comprehensive Survey," Archives of Computational Methods in Engineering, p. 1-24, 2026.

[23]. J. Gu, J. Yuan, J. Cai, et al., "La-LoRA: Parameter-efficient fine-tuning with layer-wise adaptive low-rank adaptation," Neural Networks, p. 108095. , 2025

[24]. Gu J, Yuan J, Cai J, et al. La-LoRA: Parameter-efficient fine-tuning with layer-wise adaptive low-rank adaptation. Neural Networks, p. 108095, 2025

[25]. S. Chang, M. Wan, J. Wang, et al., "Engineering material failure analysis report generation based on QWen and Llama2," Results in Engineering, **Vol**. 26, p. 104532, 2025.

[26]. Y. Hou, C. Bert, A. Gomaa, et al., "Fine-tuning a local LLaMA-3 large language model for automated privacy-preserving physician letter generation in radiation oncology," Frontiers in Artificial Intelligence, **Vol**. 7, p. 1493716, 2025.

[27]. J. Chen, J. Tian, T. Zheng, et al., "Qwen-AEC: Instruction-tuning large language models for high-performance building code analysis," Advanced Engineering Informatics, **Vol**. 71, p. 104268, 2026.

[28]. J. Qi, Z. Luan, S. Huang, et al., "LogReader: General-Purpose Log Analysis via Open-Source Large Language Models," International Joint Conference on Neural Networks (IJCNN), 2025.

[29]. Mishra R K, Choudhary A, Fatima S, et al. Multi-fault diagnosis with wavelet assisted stacked image fusion and dual branch CNN. Applied Soft Computing, **Vol**. 176, p. 113183, 2025.

[30]. Tao W, Sun Z, Yang Z, et al. Transformer fault diagnosis technology based on AdaBoost enhanced transferred convolutional neural network. Expert Systems with Applications, Vol. 264, p. 125972, 2025.

[31]. Snell J, Swersky K, Zemel R, Prototypical networks few-shot learning. Advances in Neural Information Processing Systems,**Vol**. 30, 2017.

[32]. Sung F, Yang Y, Zhang L, et al. Learning to compare: Relation network for few-shot learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p. 1199–1208 , 2018.