

Article

An End-to-End General Language Model (GLM)-4-Based Milling Cutter Fault Diagnosis Framework for Intelligent Manufacturing

Jigang He ^{1,*}, Xuan Liu ², Yuncong Lei ³ , Ao Cao ³ and Jie Xiong ¹

¹ School of Management Science and Engineering, Southwestern University of Finance and Economics, Chengdu 611130, China; xiongjie@swufe.edu.cn

² Institute of Chinese Financial Studies, Southwestern University of Finance and Economics, Chengdu 611130, China

³ School of Mechanical Engineering, Southwest Jiaotong University, Chengdu 610031, China; yuncong_lei@my.swjtu.edu.cn (Y.L.); ao@my.swjtu.edu.cn (A.C.)

* Correspondence: hjgssg@swufe.edu.cn

Abstract: CNC machine and cutting tools are an indispensable part of the cutting process. Their life default diagnosis is related to the efficiency of the entire production process, which ultimately impacts economic performance. Many methods provided by deep learning articles have been verified for use on large cutting datasets and can help in diagnosing tools' lifetime well; however, on small samples, the challenge of learning difficulties still emerges. The rise in large language models (LLMs) has brought changes to tool life diagnosis. This study proposes a fault diagnosis algorithm based on GLM-4, and the experimental validation on the PHM 2010 dataset and a proprietary milling cutter dataset demonstrates the superiority of the proposed model, achieving diagnostic accuracies of 93.8% and 93.3%, respectively, outperforming traditional models (SVM, CNN, RNN) and baseline LLMs (ChatGLM2-6B variants). Further robustness and noise-resistance analyses confirm its stability under varying SNR levels (10 dB to −10 dB) and limited training samples. This work highlights the potential of integrating domain-specific feature engineering with LLMs to advance intelligent manufacturing diagnostics.

Keywords: large language models (LLMs); cutting life default diagnosis; multi-feature fusion



Academic Editor: Manuel Pineda-Sanchez

Received: 22 February 2025

Revised: 30 March 2025

Accepted: 31 March 2025

Published: 4 April 2025

Citation: He, J.; Liu, X.; Lei, Y.; Cao, A.; Xiong, J. An End-to-End General Language Model (GLM)-4-Based Milling Cutter Fault Diagnosis Framework for Intelligent Manufacturing. *Sensors* **2025**, *25*, 2295. <https://doi.org/10.3390/s25072295>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Tool processing is an essential aspect of daily production scenarios, and accurately diagnosing tool life failures enables processing factories to make timely and informed decisions. With the rise of artificial intelligence technology, deep learning has demonstrated significant advancements in the machinery sector. These techniques possess remarkable fitting capabilities for handling complex, large, and multi-signal data, effectively extracting features hidden within the signals to provide a reliable basis for monitoring and predicting tool life accurately [1]. First, deep learning can be combined with traditional mechanical classification methods or monitoring methods to help accurately determine the tool wear stage and classification [2–4]. Not only that, but deep learning can also be used to optimize cutting parameters [5], better handle noise [6], better fit tool wear values [7], and interact with the processing environment. Among these, the most intuitive use is the accurate monitoring and prediction of wear [8,9]. Some methods use deep learning combined with sensors to apply the surface roughness of the tool to judge the wear [10]; some use integrated models to fuse features from multiple sensors [11] and enhance the generalization of tool wear prediction through deep learning algorithms [12]; and some use deep learning of

multi-channel signals to convert time domain signals into two-dimensional time-frequency domain signals, perform splicing training with the number of cutting times after the dimensionality increase, and obtain the effect of significantly reducing the prediction time through the loss function [13]. The above methods are just the tip of the iceberg of how deep learning can help improve the monitoring and prediction of tool wear. Facts have proven that the characteristics of deep learning such as automatic feature extraction, multi-source data fusion, and dynamic adaptability can indeed help to effectively monitor and predict tool wear when there are sufficient cutting processing data.

However, this robust functionality encounters challenges in the context of small sample sizes, particularly with unbalanced samples [14]. The conditions of each tool processing operation can vary significantly, influenced by factors such as temperature, noise, workpiece surface conditions, vibration, and so on. Some methods, such as the multi-dimensional hybrid intelligent diagnosis method [15], Time-Spectrum Domain Adaptation Network (TSDAN) [16], transformer network and auxiliary classifier generative adversarial network (TRA-ACGAN) [17], and hybrid data augmentation mechanism (HDAM) [18], have contributed to solving the challenges of small samples and cross-conditions and have improved the adaptability of the model to a certain extent. Nonetheless, they are generally ineffective when facing extreme situations, and in small sample scenarios, excessive variations in a single factor can disproportionately impact the model, leading to reduced fitting accuracy and complicating fault diagnosis.

In addition, another challenge of small samples is that the proposed models and methods have poor migration capabilities when facing non-specific datasets, or in other words, across datasets, which in turn affects their generalization. Many transfer models, such as deep convolutional transfer learning network (DCTLN) [19] and joint distribution adaptation-based transfer network with diverse feature aggregation (JDFA) [20], and adversarial models, such as Dynamic Multi-Adversarial Adaptation Network (DMAAN) [21] and deep adversarial subdomain adaptation network (DASAN) [22], attempt to solve the problem of cross-data. Consequently, achieving high-precision diagnostics in the face of limited fault samples continues to be a critical focus for the industry.

Recently, as the advantages of large language models (LLMs) in processing sequence data have been continuously proven [23,24], large language models have begun to be put into practice in the mechanical field [25,26]. This study proposes a fault diagnosis algorithm based on GLM-4, as shown in Figure 1.

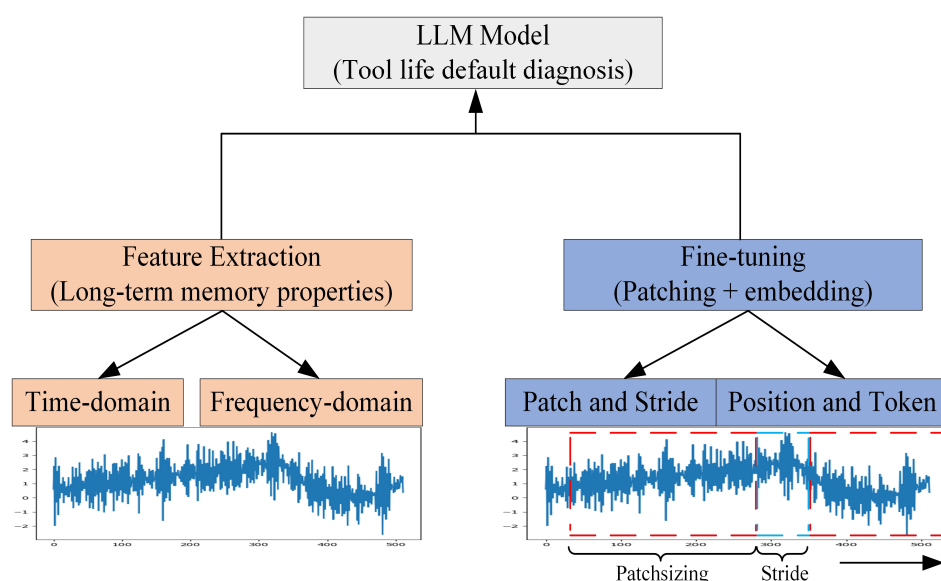


Figure 1. Model framework.

Recently, the advent of large models has marked a paradigm shift in milling cutter fault diagnosis. Traditional deep learning methods, such as CNNs and RNNs, have demonstrated effectiveness in extracting complex features and modeling temporal dependencies. However, the scale and pre-trained nature of large language models (LLMs) have opened new avenues for enhancing diagnostic performance. By leveraging massive amounts of pre-learned contextual knowledge and advanced attention mechanisms, large models not only capture global patterns across multi-sensor data but also provide robust semantic understanding even under small sample conditions. This trend is exemplified by recent studies that apply LLMs to industrial fault diagnosis tasks, where they achieve superior performance in handling noise, variability, and cross-domain generalization compared to conventional architectures. Thus, integrating large model technology into milling cutter fault diagnosis represents an exciting development that bridges state-of-the-art natural language processing techniques with practical manufacturing applications.

This study presents a novel framework for milling cutter fault diagnosis, addressing critical challenges in small-sample industrial scenarios. The key contributions are threefold:

Semantic Transformation of Numerical Signals: The GLM-4 Feature Extraction Module (GLM-4-FE) transforms raw vibration signals into semantically rich linguistic descriptions, bridging traditional numerical diagnostics with advanced language models. This transformation enhances interpretability, preserves feature significance, and improves classification accuracy and adaptability across diverse fault conditions.

Adaptive Multi-Domain Feature Fusion: The framework's adaptive feature fusion mechanism (FE) integrates time-domain, frequency-domain, and time-frequency-domain features, capturing both short-term bursts and long-range temporal dependencies. This comprehensive fusion provides robust representations capable of detecting complex failure modes, enhancing the predictive accuracy and versatility in dynamic industrial settings.

Time-Series Fine-Tuning for Robustness: By employing time-series fine-tuning (TS), the framework refines its sensitivity to temporal variations, detecting abrupt changes and non-stationary signals with precision. Combined with multi-head attention in pre-trained models, this approach ensures robustness against noise, strong adaptability to transfer learning, and stable diagnostic performance even in data-scarce or noisy environments.

The remainder of this paper is structured as follows: Section 2 details the architecture of the proposed fault diagnosis framework and its key components. Section 3 covers the experimental setup, including the datasets (PHM2010 and milling cutter data) and baseline models (SVM, CNN, RNN, ChatGLM2), followed by a comprehensive evaluation comparing the proposed model's accuracy, robustness, and generalization across various conditions. Section 4 concludes by summarizing the key contributions and potential applications in industrial fault monitoring and predictive maintenance.

2. Model Construction

2.1. Framework of Feature-Based GLM-4

2.1.1. Feature Extraction

Unlike the traditional GLM-4 framework that primarily handles semantic text, GLM-4-FE leverages time-domain vibration signals collected from bearings, represented as discrete data points by sensors. To fully exploit GLM-4's capability in processing semantic information, we extracted interpretable feature information from the raw vibration signals. Regarding feature selection, we focused on time-domain, frequency-domain, and time-frequency-domain features, as shown in Table 1. This selection is based on two main reasons: firstly, the extraction of complex features is relatively cumbersome and may require tailored designs specific to the characteristics of the vibration data; secondly, we aim to fully utilize GLM-4's robust learning, extraction, and classification capabilities by

employing easily obtainable time-domain and frequency-domain features for effective fault diagnosis. Time-domain features reflect the instantaneous variations of the signal, while frequency-domain features reveal the periodicity and frequency components, providing a more comprehensive description of the fault signal characteristics when combined.

Table 1. Feature extraction formula and description.

Feature Domain	Feature Name	Mathematical Expression	Physical Meaning
Time Domain	Mean Value (MV)	$X_1 = \frac{1}{N} \sum_{i=1}^N x_i $	The average trend of signal amplitude variation.
	Root Mean Square (RMS)	$X_2 = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$	The mean energy of the signal over a given time interval.
	Standard Deviation (SD)	$X_3 = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - X_1)^2}$	The degree of fluctuation of the signal around the mean.
	Skewness Factor (SF)	$X_4 = X_2 / X_1$	Variations in the signal waveform.
	Skewness (Ske)	$X_5 = \frac{1}{N} \sum_{i=1}^N \left(\frac{ x_i - X_1}{X_3} \right)^3$	The degree to which the signal distribution deviates from the mean symmetry line.
	Kurtosis (Kur)	$X_6 = \frac{1}{N} \sum_{i=1}^N \left(\frac{ x_i - X_1}{X_3} \right)^4$	The smoothness of the signal waveform.
Time Domain	Peak Value (PV)	$X_7 = \max x_i $	The maximum instantaneous amplitude of the signal.
	Crest Factor (CF)	$X_8 = X_7 / X_2$	The extremity of the peak in the signal waveform.
	Impact Factor (IF)	$X_9 = X_7 / X_1$	The instantaneous impact characteristics of the signal.
Frequency Domain	Mean Power Spectrum (MPS)	$X_{10} = \frac{1}{N} \sum_{i=1}^N P(f_i)$	The variation of signal power with frequency.
	Frequency Center (FC)	$X_{11} = \frac{\sum_{i=1}^N f_i \times P(f_i)}{\sum_{i=1}^N P(f_i)}$	The static portion of the spectrum.
	Mean Square Frequency (MSF)	$X_{12} = \frac{\sum_{i=1}^N f_i^2 \times P(f_i)}{\sum_{i=1}^N P(f_i)}$	The degree of fluctuation of the spectrum near the frequency centroid.
Time–Frequency Domain	Wavelet Packet Energy (WPE)	$X_{13} = \frac{1}{N} \sum_{i=1}^N \left[d_{i,k}^{(M)}(t) \right]^2$	The average energy of the signal at different scales.

2.1.2. Semantic Description

After extracting features, we proceeded with the dataset construction and fine-tuning steps to convert these numerical features into a format that GLM-4 can comprehend using linguistic descriptions instead of raw numerical values. We combined the extracted time–frequency-domain feature values with their corresponding textual descriptions and used the fault mode as labels for supervised learning, forming question–answer pairs as inputs. This approach not only preserves the physical significance of the features, avoiding information loss typically associated with feature standardization, but also enables GLM-4 to more accurately understand and process input data in new tasks. To minimize the model’s reliance on expert knowledge, the textual descriptions serve merely as illustrative examples, primarily aimed at helping the LLM comprehend the task requirements and the meaning of the input data. Through this transformation from numerical features to linguistic descriptions, GLM-4-FE’s strengths in language understanding and semantic reasoning can be effectively leveraged, achieving efficient industrial fault diagnosis.

In conclusion, GLM-4-FE facilitates a seamless transition from raw vibration data to a semantically rich representation by integrating feature extraction and linguistic fine-tuning strategies. This not only enhances the interpretability of the features but also fully utilizes GLM-4’s advanced capabilities in learning and classification, making it a powerful tool

for fault diagnosis in complex industrial environments. Future work may involve further optimizing the feature extraction methods and fine-tuning strategies to enhance the model's diagnostic accuracy and adaptability, thereby meeting the increasingly diverse and complex requirements of industrial applications.

2.2. Framework of Data-Based GLM-4

2.2.1. Patching

Industrial vibration signals often exhibit high sampling frequencies and large data sizes, which can introduce substantial redundancy and high computational costs when fed directly into the model. To address this, adjacent points are grouped into patches to capture local features and reduce the overall sequence length.

Let $X \in \mathbb{R}^T$ denote the preprocessed vibration signal. Define the patch size as P , and split X into patches of length P :

$$\text{Patch}_i = [x'_{(i-1)P+1}, x'_{(i-1)P+2}, \dots, x'_{iP}], i = 1, 2, \dots, N \quad (1)$$

where $N = T/P$. If the last segment contains fewer than P points, it is truncated. Subsequently, a one-dimensional convolution can be applied to extract feature vectors from each patch directly:

$$z_i = f_\theta(\text{Patch}_i), z_i \in \mathbb{R}^{d_{\text{patch}}} \quad (2)$$

where f_θ represents a one-dimensional convolution or any suitable neural network, and z_i is the feature representation of the i -th patch.

2.2.2. Token Embedding

To enable the GLM-4 Transformer backbone to process these numerical features, each z_i must be projected to the model's input dimension d_{model} . A linear mapping is commonly used:

$$e_i = W_E z_i + b_E \quad (3)$$

where $W_E \in \mathbb{R}^{d_{\text{model}} \times d_{\text{patch}}}$ and $b_E \in \mathbb{R}^{d_{\text{model}}}$. If the patching procedure already produces z_i with dimensionality d_{model} , this linear layer can be skipped. The resulting sequence of embedded patch vectors is as follows:

$$E = [e_1, e_2, \dots, e_N]^T \in \mathbb{R}^{N \times d_{\text{patch}}} \quad (4)$$

2.2.3. Positional Embedding

Because the Transformer architecture itself does not inherently encode the ordering of input tokens, it is necessary to inject positional information explicitly. For time-series data, a sinusoidal positional embedding is often employed, embedding the position directly into the sequence representation. For the i -th patch (indexing from 1 or 0), the position encoding can be defined as follows:

$$\begin{cases} p_{i,2k} = \sin\left(i/10000^{2k/d_{\text{model}}}\right), \\ p_{i,2k+1} = \cos\left(i/10000^{2k/d_{\text{model}}}\right), \end{cases} k = 0, \dots, \frac{d_{\text{model}}}{2} - 1 \quad (5)$$

Writing this in vector form yields $p_i \in \mathbb{R}^{d_{\text{model}}}$. The final input to the Transformer's first layer is obtained by adding each patch embedding e_i to its positional vector:

$$h_i^{(0)} = e_i + p_i, i = 1, \dots, N \quad (6)$$

2.2.4. Transformer Blocks

Consider a multi-layer Transformer with L layers. From layer 1 to L , each layer's computation can be summarized as follows. Let $\mathbf{H}^{(l-1)} \in \mathbb{R}^{N \times d_{model}}$ denote the output of the $(l-1)$ -th layer.

Multi-head attention

$$\mathbf{Q} = \mathbf{H}^{(l-1)} \mathbf{W}_Q, \mathbf{K} = \mathbf{H}^{(l-1)} \mathbf{W}_K, \mathbf{V} = \mathbf{H}^{(l-1)} \mathbf{W}_V \quad (7)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{model} \times d_{model}}$. The attention mechanism is given by the following:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} \quad (8)$$

where d_k is the dimensionality per attention head. If there are h heads, their outputs are concatenated and then linearly projected to produce the final attention output $\mathbf{M}$.

Residual Connection and Feed-Forward Network (FFN)

$$\mathbf{H}'^{(l)} = \mathbf{H}^{(l-1)} + \mathbf{M} \quad (9)$$

$$\mathbf{H}^{(l)} = \mathbf{H}'^{(l)} + \text{FFN}(\mathbf{H}'^{(l)}) \quad (10)$$

Here, the FFN typically consists of two linear layers with an activation function in between.

This process is repeated through layers 1 to L , finally yielding the following:

$$\mathbf{H}^{(L)} = [\mathbf{H}_1^{(L)}, \mathbf{H}_2^{(L)}, \dots, \mathbf{H}_N^{(L)}]^T \quad (11)$$

which serves as the Transformer's top-layer output.

2.2.5. Classification Head

For fault diagnosis, the model must predict a discrete fault category. A classification head can be added on top of $\mathbf{H}^{(L)}$. Average or max-pool all patch outputs to form a global representation:

$$\mathbf{h}_{pool} = \frac{1}{N} \sum_{i=1}^N \mathbf{H}_i^{(L)} \quad (12)$$

which is then fed into a classification layer. Introduce a learnable token at the start of the sequence. After propagation through the Transformer, $\mathbf{H}_{CLS}^{(L)}$ is taken as the global feature vector. Let $\mathbf{h}_{agg} \in \mathbb{R}^{d_{model}}$ denote the aggregated representation. A linear mapping followed by softmax is applied:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_c \mathbf{h}_{pool} + \mathbf{b}_c), \mathbf{W}_c \in \mathbb{R}^{C \times d_{model}}, \mathbf{b}_c \in \mathbb{R}^C \quad (13)$$

where C denotes the number of fault categories.

2.2.6. Training and Inference

Suppose the training set is $\left\{ \left(X^{(m)}, y^{(m)} \right) \right\}_{m=1}^M$, where $y^{(m)} \in \{1, \dots, C\}$ indicates the ground-truth fault category. A standard cross-entropy loss is employed:

$$\mathcal{L} = - \sum_{m=1}^M \sum_{c=1}^C \mathbb{I}[y^{(m)} = c] \ln(\hat{y}_c^{(m)}) \quad (14)$$

in which $\hat{y}_c^{(m)}$ is the predicted probability for category c on sample m , and $\mathbb{I}$ is the indicator function. Given the vast number of parameters in GLM-4, low-rank adaptation (LoRA) can be applied to reduce memory and computational overhead. Only a subset of attention projection or feed-forward parameters is trained, while the majority of model weights remain frozen—beneficial for large-scale industrial fault diagnosis scenarios.

Overall, this workflow—encompassing patching, embedding, positional encoding, and a Transformer-based backbone—enables efficient and accurate fault diagnosis by leveraging the expressive power of large-scale language models while addressing the unique challenges of high-frequency industrial vibration data.

3. Experiment and Results

3.1. Case 1: PHM 2010

3.1.1. Dataset Description

For validation purposes, the performance of the proposed model is first tested using the PHM 2010 public dataset [27], where the experimental parameters are shown in Table 2, and the cutting conditions of the dataset remain unchanged.

Table 2. PHM 2010 competition experiment parameters.

Parameter	Category	Parameter	Category
Model	Roders Tech RFM 760	Radial cutting depth	0.125 mm
Workpiece material	Nickel-based superalloy 718	Axial cutting depth	0.2 mm
Cutter/Tool	3-Tooth ball nose milling cutter	Number of sensors	3
Spindle speed	10,400 RPM	Sensing channels	7
Feed rate	1555 mm/min	Sampling frequency	50 HZ
Cutting speed	5000–20,000 rpm	Tool diameter	6–12 mm

3.1.2. Comprehensive Evaluation

The diagnostic accuracy of each model in Case 1 is shown in Figure 2 and Table 3. In experiments on the PHM2010 dataset, SVM, CNN, and RNN represent more traditional machine learning and early deep learning approaches. Their accuracies are relatively stable but not outstanding, with SVM mostly fluctuating between 0.79 and 0.87, CNN showing a slight improvement, and RNN benefiting from temporal dependencies yet still limited in its feature extraction capabilities. As large models such as ChatGLM2-6B and GLM-4 enter the comparison, there is a marked increase in accuracy. In particular, ChatGLM2-6B-TS and GLM-4-FE exhibit strong adaptability in handling temporal information and feature representation, typically achieving scores above 0.88–0.90. Notably, when advanced feature extraction and time-series fine-tuning strategies are integrated, GLM-4 and its variants consistently reach 0.92–0.93, demonstrating their potential in multi-scale temporal capture and feature fusion.

Table 3. Average diagnostic accuracy of each model in Case 1.

Model	Diagnostic Accuracy
SVM	0.832 ^{+0.325} _{−0.275}
CNN	0.858 ^{+0.157} _{−0.098}
RNN	0.854 ^{+0.267} _{−0.195}
ChatGLM2-6B-FE	0.883 ^{+0.215} _{−0.116}
ChatGLM2-6B-TS	0.884 ^{+0.135} _{−0.076}
GLM-4-FE	0.898 ^{+0.093} _{−0.048}
GLM-4-TS	0.918 ^{+0.069} _{−0.052}

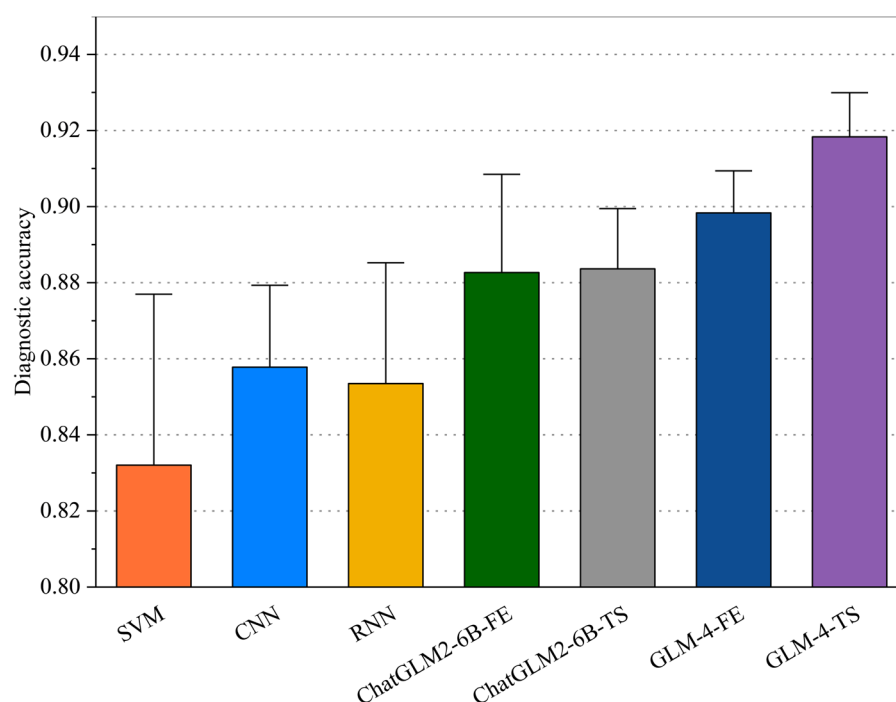


Figure 2. Diagnostic accuracy of each model in Case 1.

From a deeper perspective, large models leverage multi-head attention mechanisms and extensive pretrained parameters to effectively capture multi-dimensional correlations in fault signals across different time spans, covering both short-term fine-grained features and long-range dependencies. At the same time, GLM-4-FE further reduces redundancy in raw signals, enhancing discriminative power across various fault modes; optimizing for GLM-4-TS strengthens the model's ability to detect non-stationarity and abrupt changes. These factors collectively give the GLM-4 series a strong advantage in robustness and predictive accuracy, making them well suited to the stringent demands of industrial fault diagnosis.

3.2. Case 2: Milling Cutter Experiment

3.2.1. Dataset Description

A high-quality cutter's lifetime cycle can generate a small error (ϵ) in the economic output of the subsequent time series. The experimental data source of the milling cutter used in this study comes from real processing data. The milling cutter under consideration is the APMT1135 carbide cutter, a product of Duracarb. Its fundamental parameters include a tool tip angle of 85 degrees, a blade relief angle of 11 degrees, a blade length of 11 mm, a thickness of 3.5 mm, an inscribed circle diameter of 6.35 mm, and a maximum cutting depth of 9 mm.

3.2.2. Comprehensive Evaluation

The diagnostic accuracy of each model in Case 2 is shown in Figure 3 and Table 4. In this experiment, SVM, CNN, and RNN—representing traditional and early deep learning approaches—exhibit relatively stable yet limited performance. In contrast, ChatGLM2-6B-FE and ChatGLM2-6B-TS more effectively capture multi-dimensional features in fault signals, typically reaching accuracies around 0.90. Notably, ChatGLM2-6B-TS integrates refined TS fine-tuning, enhancing the model's ability to detect subtle temporal variations. Meanwhile, GLM-4-FE demonstrates superior generalization in multi-scale FE, balancing both local bursts and overall trends.

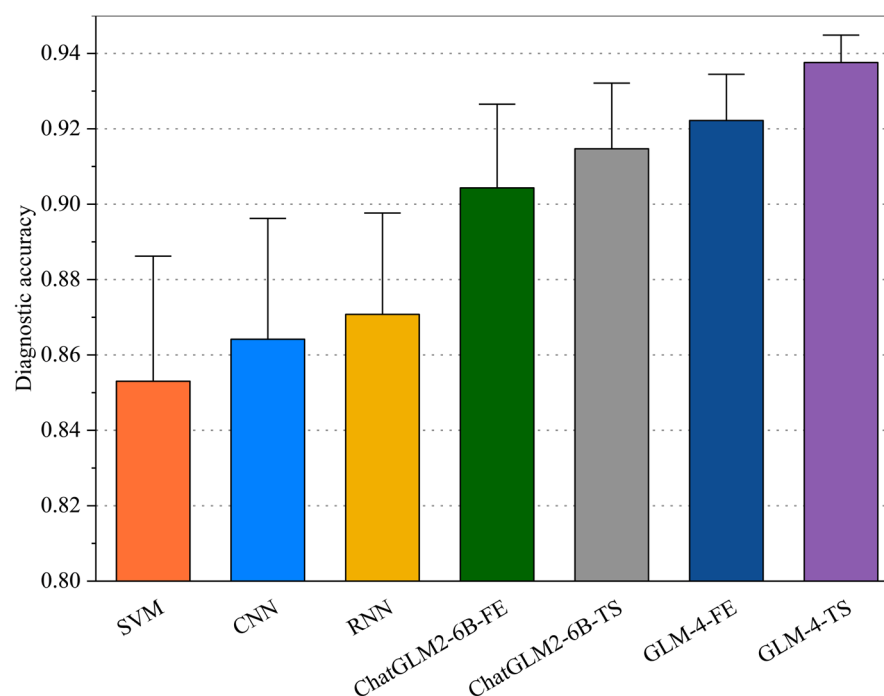


Figure 3. Diagnostic accuracy of each model in Case 2.

Table 4. Average diagnostic accuracy of each model in Case 2.

Model	Diagnostic Accuracy
SVM	0.853 ^{+0.382} _{−0.295}
CNN	0.864 ^{+0.329} _{−0.265}
RNN	0.871 ^{+0.275} _{−0.235}
ChatGLM2-6B-FE	0.904 ^{+0.215} _{−0.235}
ChatGLM2-6B-TS	0.915 ^{+0.175} _{−0.153}
GLM-4-FE	0.922 ^{+0.135} _{−0.116}
GLM-4-TS	0.938 ^{+0.079} _{−0.061}

Building upon this foundation, GLM-4-TS leverages the global attention of large pre-trained models and time-domain adaptation in unison, pushing the average accuracy beyond 0.93. The underlying reasons are twofold: on the one hand, multi-head attention uncovers intricate dependencies in fault data; on the other hand, the synergy of FE and TS fine-tuning reduces noise and redundancy, allowing flexible transitions between short- and long-term features. This multi-module strategy grants GLM-4-TS heightened robustness and sensitivity to varied failure patterns, offering a highly promising solution for industrial applications.

3.3. Performance Analysis

3.3.1. Robustness Analysis

In Figure 4, we compare SVM, CNN, and RNN at various test set proportions (from 10% to 40%) to assess each model's robustness and adaptability under different data splits. Overall, as the test set ratio increases and the training set shrinks, the average accuracy tends to decrease. SVM exhibits relatively large fluctuations—sometimes maintaining around 0.88–0.89 accuracy at lower test ratios but becoming prone to overfitting or underfitting when the data are limited. CNN shows certain advantages in capturing local features, especially under 15% or 20% test splits, thanks to its convolutional structure for extracting signal patterns. RNN, leveraging temporal dependency modeling, achieves

higher accuracy in certain splits but experiences performance swings when the training data become insufficient.

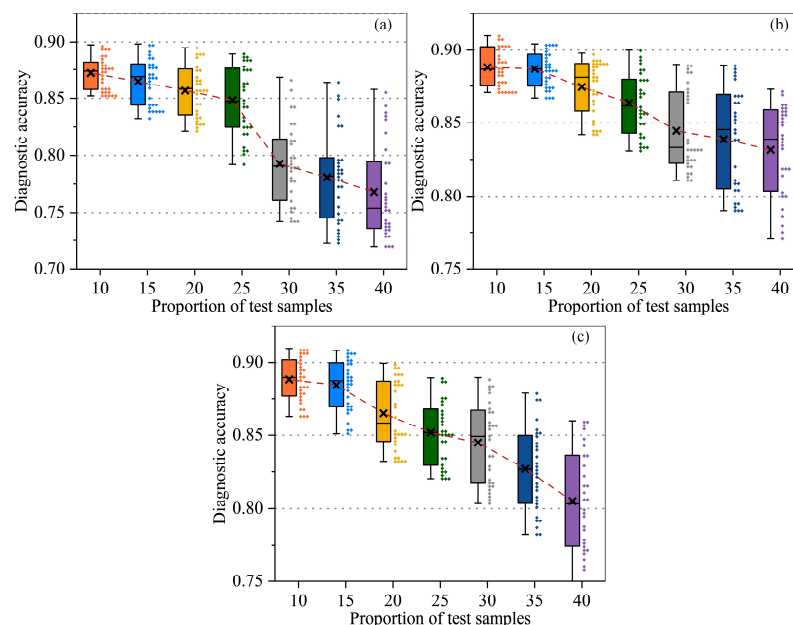


Figure 4. Diagnosis accuracy of deep learning models in different sample proportions (a–c).

In Figure 5, we examined ChatGLM2-6B-FE, ChatGLM2-6B-TS, GLM-4-FE, and GLM-4-TS under various test set ratios to evaluate their fault diagnosis performance. Overall, higher accuracies are observed when the training set is larger (i.e., a smaller test ratio). Models that incorporate FE tend to exhibit stable performance across diverse signal scenarios, while those applying TS fine-tuning excel at capturing temporal dependencies and short-term fluctuations, enhancing the detection of impulsive faults and nonstationary patterns. As the test proportion increases, the available training data decrease, generally leading to a drop in accuracy; however, performance differences among the models become more pronounced.

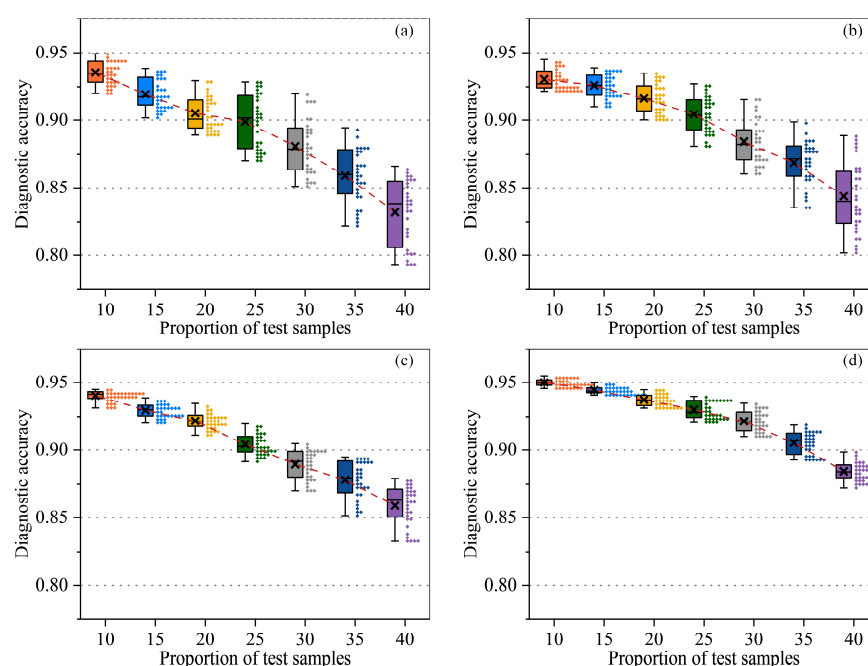


Figure 5. The diagnostic accuracy of LLMs under different sample proportions (a–d).

Notably, GLM-4-TS consistently maintains average accuracies above 0.93 across most test ratios, surpassing ChatGLM2-6B-TS and GLM-4-FE. This advantage likely stems from the synergy between the large-scale pretrained model's ability to capture global context via multi-head attention and the TS-driven refinement of time-domain features. Specifically, the former uncovers multi-scale dependencies in fault signals, while the latter focuses on local dynamics and temporal evolution. Even at a 40% test ratio, when all models face a significant accuracy drop, GLM-4-TS demonstrates superior robustness through integrated adaptations and deeper feature representation, making it well suited for industrial requirements that prioritize both limited data and high diagnostic precision.

3.3.2. Noise Resistance Analysis

In Figure 6, we compare GLM-4-FE and GLM-4-TS under varying signal-to-noise ratios (SNR ranging from 10 dB down to -10 dB). As noise levels increase and the SNR decreases, the overall accuracy declines: while both models can maintain around 0.90 accuracy at a higher SNR, the performance tends to drop significantly in extremely noisy conditions such as -10 dB. Nonetheless, GLM-4-TS demonstrates greater robustness in most noisy scenarios, indicating that TS fine-tuning enhances the model's sensitivity to subtle signal distortions and random interference.

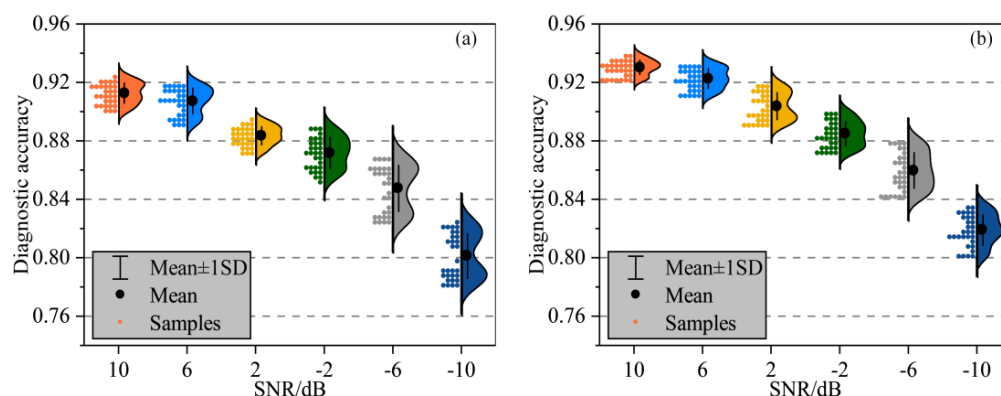


Figure 6. Diagnostic accuracy of GLM-4-TS under different SNRs (a,b).

The superior performance of our approach can be attributed to the synergistic integration of FE and TS. Specifically, GLM-4-TS exploits the large model's multi-head attention to capture global contextual patterns, while its TS component effectively adapts to local signal variations and nonstationary behaviors. In low SNR conditions—where critical information is increasingly masked by noise—traditional features alone struggle to maintain discriminative power. By dynamically refining time-domain details, the TS mechanism produces more distinct representations, leveraging the model's extensive parameter capacity and contextual learning to enhance resilience and adaptability.

3.3.3. Hyperparameter Analysis

From Figure 7, we observe a distinct pattern in GLM-4-TS's diagnostic accuracy across different batch sizes and learning rates. When the learning rate is relatively high (e.g., 10^{-1}), increasing the batch size gradually improves the accuracy, rising from 0.8888 to 0.9209. This suggests that a larger batch size can stabilize parameter updates even under a high learning rate, preventing overly turbulent gradients. However, when the learning rate is moderate (such as 10^{-2} or 10^{-3}), although increasing the batch size often leads to a higher accuracy, certain configurations—such as batch size = 256 and learning rate = 10^{-2} —show declining performance, implying that “bigger batch” does not always guarantee better results. Notably, at a learning rate of 10^{-3} and batch size = 128, the model achieves an

accuracy of 0.9337, indicating a potential sweet spot balancing stable training and efficient gradient updates.

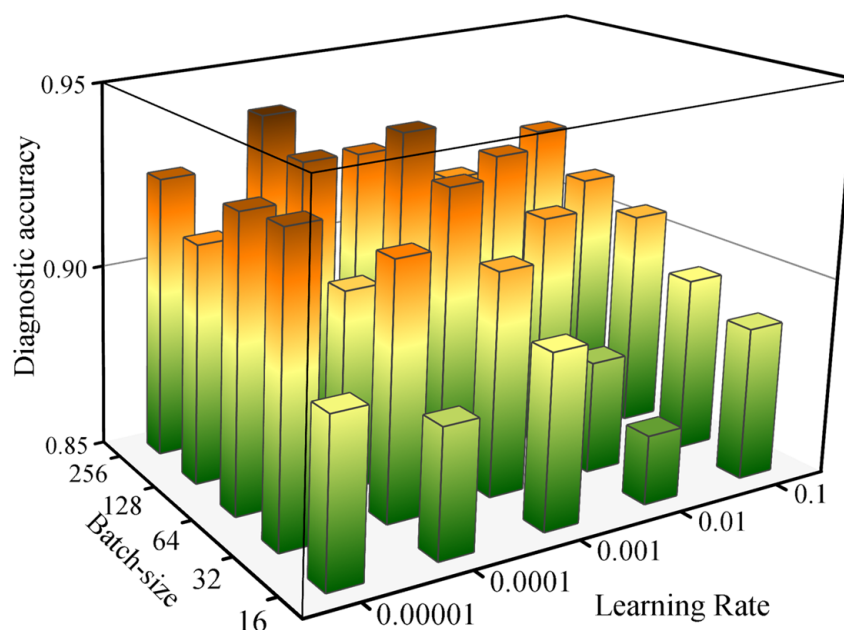


Figure 7. The influence of hyperparameters on the diagnostic accuracy of GLM-4-TS.

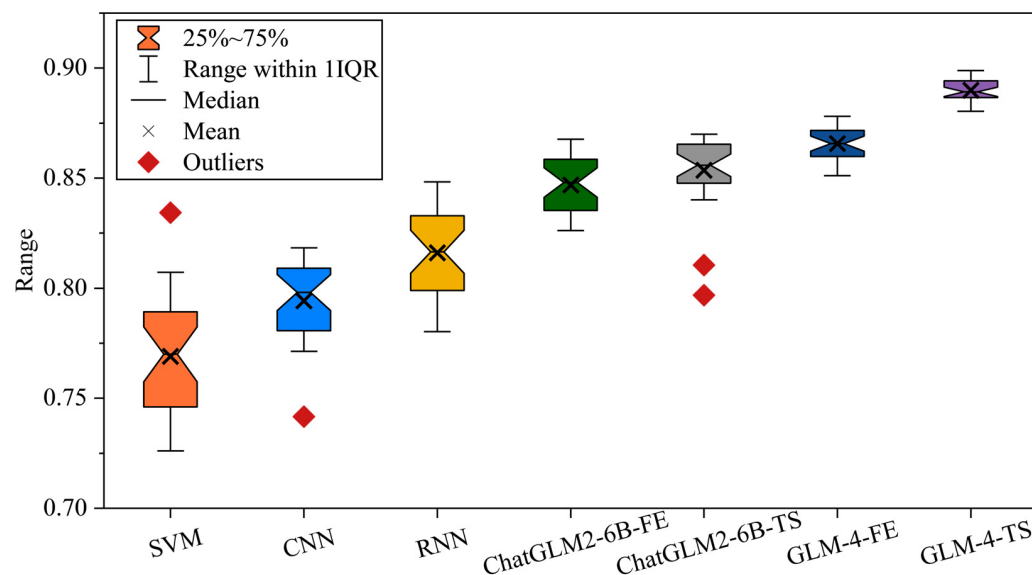
At lower learning rates (10^{-4} and 10^{-5}), the situation becomes more nuanced. In general, a larger batch size still tends to yield better accuracy; for instance, with a learning rate of 10^{-4} , batch size = 256 reaches 0.9382—over five percentage points higher than batch size = 16. Nonetheless, once the learning rate decreases to 10^{-5} , batch size = 32 achieves 0.9295 accuracy, whereas batch size = 128 lags at 0.9137. This outcome can be attributed partly to the slower convergence under lower learning rates: an excessively large batch size may further dilute gradient updates, inhibiting rapid learning. In summary, GLM-4-TS demonstrates strong diagnostic capability in most settings, but optimal hyperparameter choices depend heavily on task requirements, computational resources, and stability considerations. Striking the right balance between the batch size and learning rate is key to achieving both high accuracy and robust convergence.

3.3.4. Cross Verification

From this cross-validation experiment, we observe that traditional approaches (SVM, CNN, RNN), trained on Case 1 and fine-tuned with only 30% of the data from Case 2, generally achieve average accuracies around 0.78–0.84 on the remaining 70% test portion, as shown in Figure 8. While CNN and RNN capitalize on convolutional or recurrent structures to model limited spatiotemporal dependencies and occasionally attain slightly higher performance under simpler fault patterns, their generalization still falls short when faced with complex or noisy operating conditions. In contrast, ChatGLM2-6B-FE and ChatGLM2-6B-TS leverage large-model multi-head attention, with the former accelerating feature aggregation via FE and the latter capturing time-domain bursts through TS fine-tuning. Nonetheless, their accuracies typically hover between 0.83 and 0.86, indicating that discrepancies in task transition and data distribution may constrain their cross-scenario generalization potential. See Table 5.

Table 5. Division of training and test dataset.

No.	Train	Transfer Dataset	Test Dataset
1	Case 1	30% Case 2	70% Case 2
2	Case 2	30% Case 1	70% Case 2

**Figure 8.** Experimental results based on Case 1 with limited cross-dataset.

Within the GLM-4 series, our experiments reveal that GLM-4-FE achieves stable diagnostic accuracies of approximately 0.86–0.88, demonstrating robust feature fusion and cross-domain adaptability. In contrast, GLM-4-TS consistently surpasses accuracies of 0.89–0.90, emerging as the most effective variant. Two principal factors underlie these results: first, the pre-trained model's rich contextual representation facilitates the discovery of hidden fault patterns across diverse data sources; second, the TS strategy enables flexible adjustment for varying sampling rates and nonstationary signal characteristics, effectively mitigating domain shift. Notably, even with fine-tuning on only 30% of Case 2's data, GLM-4-TS exhibits strong small-sample transferability and multi-condition synthesis, underscoring its broad applicability in industrial fault diagnosis.

In Figure 9, we compared traditional machine learning models, early deep learning models (CNN, RNN), and advanced language models (ChatGLM2-6B-FE, ChatGLM2-6B-TS, GLM-4-FE, GLM-4-TS) in terms of diagnostic accuracy under different data splitting strategies. The experimental setup involved initial training on Case 2, followed by fine-tuning with only 30% of the data from Case 1 and testing on the remaining 70%. The results demonstrate that GLM-4-TS outperforms all other models, achieving an average accuracy of 0.9115, significantly higher than its counterparts. Specifically, the superiority of GLM-4-TS lies in its integration of FE and TS fine-tuning optimization strategies, which enhance its capabilities in feature fusion and temporal information capture. In contrast, traditional models such as SVM, CNN, and RNN, while performing adequately in certain scenarios, generally exhibit lower accuracies and greater variability across diverse and complex fault patterns. The ChatGLM2-6B series models, although improved by incorporating FE and TS, still fall short of GLM-4-TS's performance, indicating room for enhancement in their feature extraction and temporal adaptation processes.

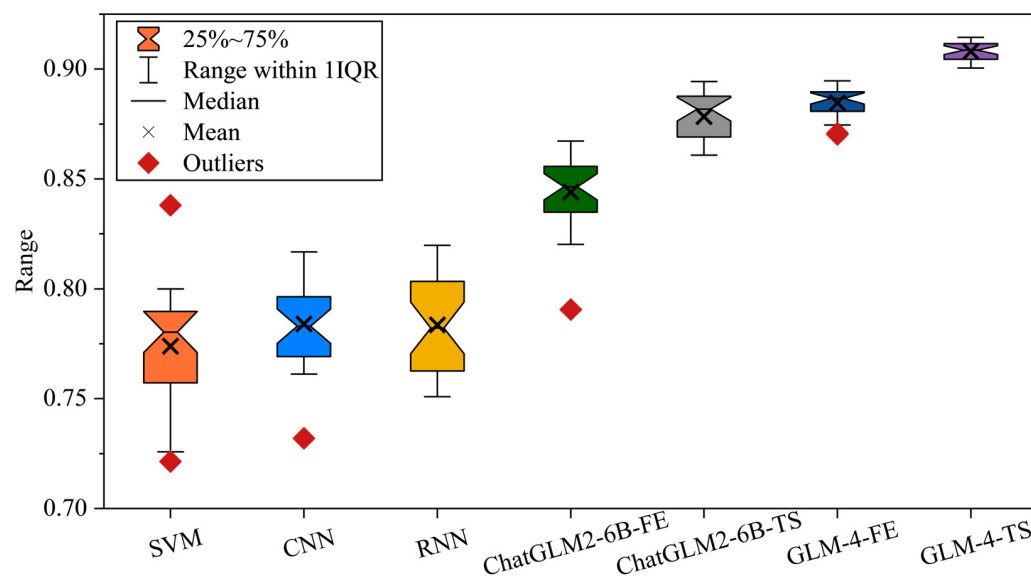


Figure 9. Experimental results based on Case 2 with limited cross-dataset.

3.3.5. Training Loss Visualization

From Figure 10, we can observe distinct convergence patterns for each model. SVM, CNN, and RNN all begin with initial losses around 0.9 and exhibit relatively stable or slow declines in the early epochs. Notably, SVM demonstrates greater fluctuations between Epochs 2 and 3, suggesting that traditional machine learning and early deep learning approaches often require more careful parameter tuning and feature selection when dealing with complex tasks. While CNN and RNN reduce their losses to around 0.70–0.74 by about the 10th epoch, they do not continue dropping quickly thereafter, indicating inherent limitations in their ability to capture nonstationary temporal features and a tendency toward local minima or sensitivity to high-frequency noise. Additionally, CNN experiences a temporary bump at Epochs 7–8, possibly due to instability in the learning rate, kernel configuration, or regularization settings.

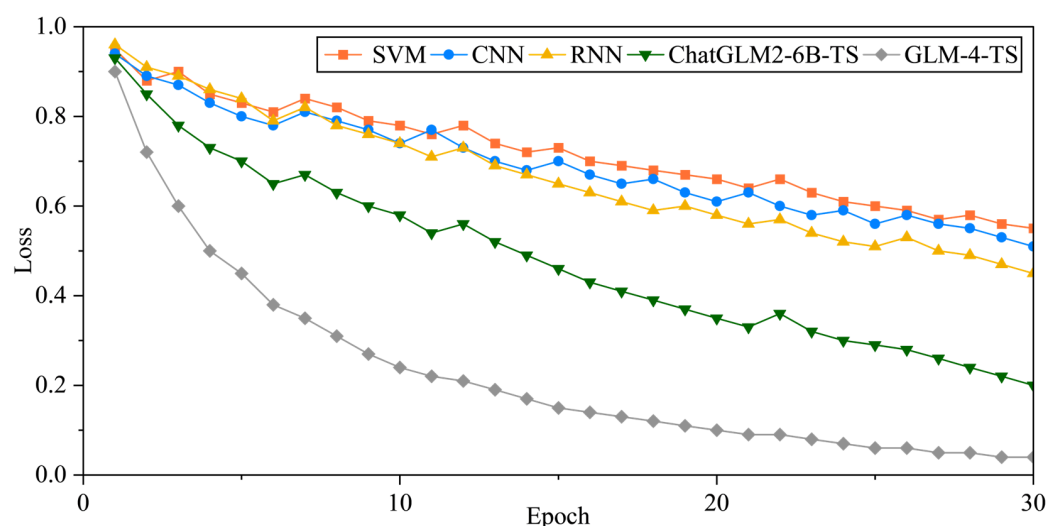


Figure 10. Training losses for each model.

In stark contrast, ChatGLM2-6B-TS and GLM-4-TS exhibit significantly faster convergence. ChatGLM2-6B-TS begins with an initial loss of 0.93 in Epoch 1, noticeably lower than the 0.95–0.96 range of the other methods. GLM-4-TS goes even further, starting at 0.90 and dropping its loss to around 0.60 after only three to four epochs—far quicker than

CNN and RNN. Such efficiency stems from the capacity of large-scale pretrained models to perform robust feature extraction and precisely characterize time-series bursts, aided by multi-head attention mechanisms. Over subsequent training rounds, GLM-4-TS leverages TS fine-tuning to an even greater extent, driving the loss down to 0.10 by Epoch 20 and eventually stabilizing at approximately 0.04, thus demonstrating superior resilience against noise and dynamic variability. Although ChatGLM2-6B-TS also maintains a steady descent in later epochs, it converges somewhat more slowly than GLM-4-TS. Overall, the results illustrate that advanced large models can achieve fine-grained learning of complex fault patterns in fewer epochs, converging to significantly lower loss levels, thereby underscoring their potent adaptability to high-dimensional, nonlinear time-series data.

From Figure 11, we observe distinct convergence behaviors among three parameter-efficient fine-tuning approaches (BitFit, QLoRA, and LoRA) applied to GLM-4-TS. BitFit begins with a loss of 0.9342 at Epoch 1, which is lower than QLoRA's 0.9757 but slightly higher than LoRA's 0.90. In subsequent epochs (e.g., Epochs 2–3 and 7–8), BitFit undergoes relatively large fluctuations yet retains a decreasing tendency, reaching 0.5318 by Epoch 10. Compared to QLoRA's 0.2586 at a similar stage, this is considerably higher, suggesting that BitFit may be less sensitive to certain crucial parameters in a high-dimensional model. Furthermore, BitFit exhibits multiple upward spikes in the later training phase (notably around Epoch 22, rebounding to 0.3837), implying that its long-term convergence may be more susceptible to hyperparameters such as the learning rate and gradient accumulation, ultimately leading to less stability than the other two methods.

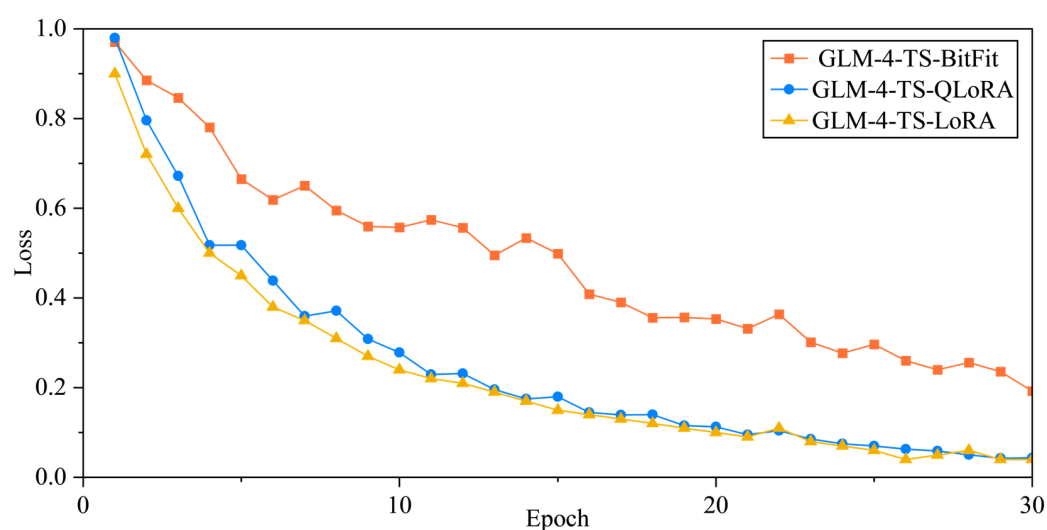


Figure 11. Influence of different fine-tuning methods on training loss.

In contrast, QLoRA starts at a relatively high loss of 0.9757 during Epoch 1 but rapidly converges in the following epochs, dropping to about 0.4577 by Epoch 5. It continues to steadily reduce its loss beyond Epoch 10, hovering near 0.25 and eventually dipping to approximately 0.0855 by Epoch 23. This reveals QLoRA's ability to effectively blend fine-grained parameter control with quantization techniques. LoRA, on the other hand, shows a comparatively favorable starting point (0.90 at Epoch 1), swiftly descending to 0.72, 0.60, and 0.50 over Epochs 2–4. Thereafter, it maintains a stable downward trend, attaining 0.24 at Epoch 10 and leveling off near 0.04 by Epoch 30, comparable to QLoRA's final range. Notably, LoRA's mid-to-late training exhibits minimal rebounds or oscillations, highlighting its stable adaptation in low-rank updates to multi-head attention and feed-forward layers, enabling the continuous extraction of model potential and suppression of redundant updates. Overall, both QLoRA and LoRA demonstrate robust convergence speed and stability for deep time-series fine-tuning tasks, whereas BitFit may require more

rigorous hyperparameter tuning and regularization strategies to unleash its full potential in complex fault diagnosis scenarios.

4. Conclusions

In summary, this study introduces a novel end-to-end fault diagnosis framework for milling cutters in intelligent manufacturing by integrating a state-of-the-art large language model (GLM-4) with advanced data preprocessing, feature extraction, and fine-tuning strategies. The framework transforms raw vibration signals into semantically rich representations, thereby bridging the gap between numerical sensor data and linguistic models. The principal contributions and findings of this work are as follows:

Innovative Semantic Transformation: The proposed GLM-4-FE module converts complex vibration signal features into interpretable textual descriptions. This semantic transformation enables the model to capture subtle fault characteristics that are often missed by traditional numerical approaches.

Robust Time-Series Fine-Tuning: The GLM-4-TS approach leverages multi-head attention and temporal adaptation to extract both local and long-range dependencies from time-series data. The experimental results on both the PHM 2010 dataset and our proprietary milling cutter dataset demonstrate diagnostic accuracies exceeding 93%, even under high-noise conditions (SNR as low as −10 dB) and with limited training samples.

Comprehensive Comparative Analysis: Compared with traditional methods (e.g., SVM, CNN, and RNN) and baseline large language models, our framework consistently outperforms in terms of accuracy, robustness, and generalization. This improvement underscores the advantage of integrating domain-specific feature engineering with the semantic reasoning capabilities of large language models.

Despite these promising results, certain limitations remain. The current framework has been validated primarily in single-tool scenarios; therefore, extending the approach to handle multi-tool systems and cross-scenario generalization remains a challenge. Additionally, while the fine-tuning process is efficient, the further optimization of hyperparameters and exploration of data augmentation techniques could enhance performance under even more challenging conditions. Future research will focus on the following:

Multi-Tool System Extension: Expanding the framework to accommodate multiple tool types and varying operational conditions, potentially through federated learning strategies to support distributed industrial deployments.

Advanced Data Augmentation: Investigating the integration of generative adversarial networks (GANs) and other augmentation methods to further address the issues of small sample sizes and extreme noise environments.

Hybrid Model Development: Exploring models that combine both feature-based and raw time-series inputs to fully exploit complementary information and further boost diagnostic accuracy.

In conclusion, the proposed GLM-4-based fault diagnosis framework provides a scalable, interpretable, and high-performance solution for real-time tool health monitoring. By bridging the gap between the theoretical advancements in large language models and practical manufacturing demands, this work lays a strong foundation for next-generation intelligent fault diagnosis systems in industrial applications.

Author Contributions: Conceptualization, J.H.; methodology, J.H.; software, A.C. and J.H.; validation, J.H.; formal analysis, J.H. and X.L.; data curation, J.H., X.L., Y.L. and J.X.; writing—original draft preparation, J.H.; writing—review and editing, J.H., X.L., Y.L., A.C. and J.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: PHM 2010 is available at https://phmsociety.org/phm_competition/2010-phm-society-conference-data-challenge (accessed on 30 March 2025).

Conflicts of Interest: The authors declare no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
CNC	Computerized Numerical Control
CNN	Convolutional Neural Network
ChatGLM2	Chat GLM Version 2
LLM	Large Language Model
GLM-4	General Language Model 4
RNN	Recurrent Neural Network
SVM	Support Vector Machine
SNR	Signal-to-Noise Ratio
LoRA	Low-Rank Adaptation
QLoRA	Quantized LoRA
PHM	Prognostics Health Management

References

1. He, J.; Sun, Y.; Gao, H.; Guo, L.; Cao, A.; Chen, T. On-Line Milling Tool Wear Monitoring under Practical Machining Conditions. *Measurement* **2023**, *222*, 113621. [\[CrossRef\]](#)
2. Abdeltawab, A.; Xi, Z.; Longjia, Z. Tool Wear Classification Based on Maximal Overlap Discrete Wavelet Transform and Hybrid Deep Learning Model. *Int. J. Adv. Manuf. Technol.* **2024**, *130*, 2381–2406. [\[CrossRef\]](#)
3. Zhang, Y.; Qi, X.; Wang, T.; He, Y. Tool Wear Condition Monitoring Method Based on Deep Learning with Force Signals. *Sensors* **2023**, *23*, 4595. [\[CrossRef\]](#)
4. Kale, A.P.; Wahul, R.M.; Patange, A.D.; Soman, R.; Ostachowicz, W. Development of Deep Belief Network for Tool Faults Recognition. *Sensors* **2023**, *23*, 1872. [\[CrossRef\]](#)
5. Gao, Z.; Chen, N.; Yang, Y.; Li, L. An Innovative Deep Reinforcement Learning-Driven Cutting Parameters Adaptive Optimization Method Taking Tool Wear into Account. *Measurement* **2025**, *242*, 116075. [\[CrossRef\]](#)
6. Sun, Y.; Song, H.; Gao, H.; Li, J.; Yin, S. Interpretable Tool Wear Monitoring: Architecture with Large-Scale CNN and Adaptive EMD. *J. Manuf. Syst.* **2025**, *78*, 294–307. [\[CrossRef\]](#)
7. Xu, X.; Wang, J.; Zhong, B.; Ming, W.; Chen, M. Deep Learning-Based Tool Wear Prediction and Its Application for Machining Process Using Multi-Scale Feature Fusion and Channel Attention Mechanism. *Measurement* **2021**, *177*, 109254. [\[CrossRef\]](#)
8. Cheng, M.; Jiao, L.; Yan, P.; Jiang, H.; Wang, R.; Qiu, T.; Wang, X. Intelligent Tool Wear Monitoring and Multi-Step Prediction Based on Deep Learning Model. *J. Manuf. Syst.* **2022**, *62*, 286–300. [\[CrossRef\]](#)
9. Yao, J.; Lu, B.; Zhang, J. Tool Remaining Useful Life Prediction Using Deep Transfer Reinforcement Learning Based on Long Short-Term Memory Networks. *Int. J. Adv. Manuf. Technol.* **2022**, *118*, 1077–1086. [\[CrossRef\]](#)
10. Huang, P.-M.; Lee, C.-H. Estimation of Tool Wear and Surface Roughness Development Using Deep Learning and Sensors Fusion. *Sensors* **2021**, *21*, 5338. [\[CrossRef\]](#)
11. Di, Z.; Yuan, D.; Li, D.; Liang, D.; Zhou, X.; Xin, M.; Cao, F.; Lei, T. Tool Fault Diagnosis Method Based on Multiscale-Efficient Channel Attention Network. *J. Mech. Eng.* **2024**, *60*, 82–90.
12. Liu, X.; Zhang, B.; Li, X.; Liu, S.; Yue, C.; Liang, S.Y. An Approach for Tool Wear Prediction Using Customized DenseNet and GRU Integrated Model Based on Multi-Sensor Feature Fusion. *J. Intell. Manuf.* **2023**, *34*, 885–902. [\[CrossRef\]](#)
13. Yan, B.; Zhu, L.; Dun, Y. Tool Wear Monitoring of TC4 Titanium Alloy Milling Process Based on Multi-Channel Signal and Time-Dependent Properties by Using Deep Learning. *J. Manuf. Syst.* **2021**, *61*, 495–508. [\[CrossRef\]](#)
14. Chen, H.-Y.; Lee, C.-H. Deep Learning Approach for Vibration Signals Applications. *Sensors* **2021**, *21*, 3929. [\[CrossRef\]](#)
15. Lei, Y.; Zuo, M.J.; He, Z.; Zi, Y. A Multidimensional Hybrid Intelligent Method for Gear Fault Diagnosis. *Expert Syst. Appl.* **2010**, *37*, 1419–1430. [\[CrossRef\]](#)

16. Ding, Y.; Cao, Y.; Jia, M.; Ding, P.; Zhao, X.; Lee, C.-G. Deep Temporal–Spectral Domain Adaptation for Bearing Fault Diagnosis. *Knowl.-Based Syst.* **2024**, *299*, 111999. [\[CrossRef\]](#)
17. Fu, Z.; Liu, Z.; Ping, S.; Li, W.; Liu, J. TRA-ACGAN: A Motor Bearing Fault Diagnosis Model Based on an Auxiliary Classifier Generative Adversarial Network and Transformer Network. *ISA Trans.* **2024**, *149*, 381–393. [\[CrossRef\]](#)
18. Quan, Y.; Liu, C.; Yuan, Z.; Yan, B. Hybrid Data Augmentation Combining Screening-Based MCGAN and Manual Transformation for Few-Shot Tool Wear State Recognition. *IEEE Sens. J.* **2024**, *24*, 12186–12196. [\[CrossRef\]](#)
19. Guo, L.; Lei, Y.; Xing, S.; Yan, T.; Li, N. Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines With Unlabeled Data. *IEEE Trans. Ind. Electron.* **2019**, *66*, 7316–7325. [\[CrossRef\]](#)
20. Jia, S.; Deng, Y.; Lv, J.; Du, S.; Xie, Z. Joint Distribution Adaptation with Diverse Feature Aggregation: A New Transfer Learning Framework for Bearing Diagnosis across Different Machines. *Measurement* **2022**, *187*, 110332. [\[CrossRef\]](#)
21. Tian, M.; Su, X.; Chen, C.; Luo, Y.; Sun, X. Bearing Fault Diagnosis of Wind Turbines Based on Dynamic Multi-Adversarial Adaptive Network. *J. Mech. Sci. Technol.* **2023**, *37*, 1637–1651. [\[CrossRef\]](#)
22. Liu, Y.; Wang, Y.; Chow, T.W.S.; Li, B. Deep Adversarial Subdomain Adaptation Network for Intelligent Fault Diagnosis. *IEEE Trans. Ind. Inf.* **2022**, *18*, 6038–6046. [\[CrossRef\]](#)
23. Cao, Y.; Zhao, H.; Cheng, Y.; Shu, T.; Chen, Y.; Liu, G.; Liang, G.; Zhao, J.; Yan, J.; Li, Y. Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, 1–21. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Gruver, N.; Finzi, M.; Qiu, S.; Wilson, A.G. Large Language Models Are Zero-Shot Time Series Forecasters. *arXiv* **2023**; arXiv:2310.07820.
25. Tao, L.; Liu, H.; Ning, G.; Cao, W.; Huang, B.; Lu, C. LLM-Based Framework for Bearing Fault Diagnosis. *Mech. Syst. Signal Process.* **2025**, *224*, 112127. [\[CrossRef\]](#)
26. Lialin, V.; Deshpande, V.; Yao, X.; Rumshisky, A. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning. *arXiv* **2024**, arXiv:2303.15647.
27. 2010 PHM Society Conference Data Challenge. Available online: https://phmsociety.org/phm_competition/2010-phm-society-conference-data-challenge/ (accessed on 1 April 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.