applied sciences

MDPI

Article

MechaForge: A Multi-Strategy Time-Series Synthesis Framework for Intelligent Fault Diagnosis

Xiyang Zhang [1,2,3], Xia Liu [1,2,*], Feiyang Li [1,2], Yi Hu [2,3], Dong Yu [2,3] and Yongze Ma [1,2]

[1] Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China; lifeiyang23@mails.ucas.ac.cn (F.L.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Shenyang CASNC Technology Co., Ltd., Shenyang 110168, China
* Correspondence: liuxia24@mails.ucas.ac.cn

Abstract

Intelligent fault diagnosis of rotating machinery is essential for manufacturing reliability and predictive maintenance, yet deployment of deep learning models is limited by data scarcity: fault samples are rare, costly, and hazardous to obtain. Conventional synthetic data methods such as Generative Adversarial Networks and Variational Autoencoders often exhibit mode collapse, spectral distortion, and limited physical interpretability. This work presents MechaForge, a multi-strategy framework that employs Large Language Models (LLMs) as physics-guided generators for bearing fault time-series data. The approach is grounded in bearing kinematics, Motor Current Signature Analysis (MCSA), and the interpretation of in-context learning as implicit Bayesian inference. Within MechaForge, four progressively constrained tracks are defined: a real-data baseline, few-shot LLM mimicry, multi-stage semantic reasoning, and physics-guided generation with constraints on root mean square, kurtosis, and fault-band spectral energy. For direct benchmarking, conventional VAE- and GAN-based augmentation baselines are additionally evaluated under the same dataset split, synthetic-data budget, downstream CNN architecture, and evaluation metrics. Experiments on the Paderborn bearing dataset show that the Basic LLM track achieves the strongest performance under the present protocol (0.7862 accuracy, 0.7648 macro-F1), exceeding the added VAE and GAN baselines (both 0.7428 accuracy; 0.7202 and 0.7257 macro-F1, respectively), while a control experiment confirms that synthetic data provides discriminative structure rather than labeled noise. These results indicate the promise of LLM-based diagnostic augmentation under data scarcity in the present Paderborn setting, rather than a definitive demonstration of broad transferability across fault-diagnosis scenarios.

Keywords: fault diagnosis; rolling bearings; time-series synthesis; Large Language Models; physics-guided generation

Check for updates

Academic Editor: Degang Xu

Received: 26 March 2026
Revised: 24 April 2026
Accepted: 28 April 2026
Published: 6 May 2026

Copyright: © 2026 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.

1. Introduction

The global manufacturing paradigm, increasingly driven by Industry 4.0 and the emerging cyber-physical integration of Industry 5.0, relies heavily on the continuous, optimal performance of rotating machinery [1]. In these complex electromechanical systems, rolling element bearings constitute the most critical points of failure, accounting for approximately 40–50% of all machine breakdowns [2]. Unplanned downtime resulting from bearing degradation incurs substantial economic losses, jeopardizes operational

safety, and degrades product quality across critical sectors including aerospace, wind energy, and automotive manufacturing [3]. Consequently, intelligent fault diagnosis (IFD)—leveraging advanced signal processing, vibration analysis, current signature analysis, and deep learning—has become the cornerstone of modern predictive maintenance, enabling critical capabilities ranging from early fault detection to remaining useful life (RUL) estimation in complex industrial environments [4,5]. Recent deep learning studies have addressed not only bearing fault detection but also remaining useful life (RUL) estimation, further highlighting the broader importance of data-driven monitoring for predictive maintenance. Complementing this PHM perspective, vibration-based anomaly detection using deep learning models and latent-space representations also supports practical condition-monitoring pipelines for industrial machines [6].

A fundamental obstacle to deploying data-driven IFD at scale is data scarcity [7]. In operational production environments, machinery operates under healthy conditions for the vast majority of its lifecycle. Faults, particularly in incipient micro-spalling stages, are temporally rare and evolutionarily unpredictable [8,9]. Deliberate fault injection on production lines is prohibitively expensive and poses safety risks; natural degradation requires costly expert annotation via teardown inspections [7,10]. As a result, anomalous fault data frequently constitutes less than 1% of industrial datasets, creating severe class imbalance that skews decision boundaries, induces overfitting, and degrades generalization in the field [11,12].

The prognostic and health management community has sought to address this data gap using generative models, predominantly Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [13–17]. However, applying these models to high-frequency, non-stationary mechanical signals reveals critical theoretical and practical limitations. First, adversarial training is notoriously unstable and highly susceptible to mode collapse, routinely failing to capture the full distribution of diverse fault severities and operating conditions [18]. Second, and more critically, these architectures lack an intrinsic understanding of the physical laws governing rotating machinery; they operate purely in the data domain, frequently generating signals with severe spectral distortion where critical kinematic frequencies—such as the Ball Pass Frequency of the Outer race (BPFO)—are obfuscated, aliased, or entirely lost due to inadequate frequency-domain constraints [19–21]. Finally, GANs suffer from a fundamental lack of interpretability, offering no mechanism to inject domain knowledge, enforce energy conservation, or constrain generation based on known kinematic relationships.

Recent years have witnessed rapid adoption of Large Language Models (LLMs) for time-series analysis, forecasting, and synthesis [22–25]. LLMs combine large-scale pre-training with self-attention to provide strong sequence modeling, cross-modal knowledge, and zero-shot generalization [26,27]. The modality gap between discrete language tokens and continuous sensor signals is being systematically addressed: for example, ChatTS treats multivariate time series as a primary modality with value-preserving normalization for conversational reasoning and structural synthesis [28]. In mechanical fault diagnosis, BearGen fuses LLM guidance with diffusion models to synthesize bearing fault signals while preserving frequency structure that GANs often distort [29]; SDForger shows that tabular embeddings enable autoregressive LLMs to generate high-fidelity synthetic series with minimal fine-tuning [30]; and Thoth-style mid-training bridges general-purpose LLMs to time-series understanding for text-to-time-series generation [31].

Recent studies have also highlighted that bearing fault diagnosis under changing operating conditions remains a central challenge, and that transferability across varying speeds and loads may be more practically important than in-condition classification alone. In parallel, recent work has emphasized the value of interpretable diagnosis under domain

shift, for example through prototype-attention domain adaptation frameworks designed for varying operating conditions [32]. This line of work helps position the present manuscript more clearly with respect to domain robustness and explainable diagnosis.

Despite recent progress in GAN-, diffusion-, and LLM-based time-series synthesis, important challenges remain unresolved for industrial fault signals, including spectral distortion, limited interpretability, and weak mechanisms for injecting domain-specific physical priors into the generation process. In this context, MechaForge is positioned as a physics-guided prompting framework for LLM-based synthesis, in which bearing kinematics and MCSA-related priors are translated into prompt-level generation guidance rather than treating the LLM as a purely data-driven sequence imitator. The main contributions of this paper are summarized as follows: (1) a multi-strategy LLM synthesis framework for bearing fault diagnosis under data scarcity; (2) a physics-guided prompting mechanism that introduces statistical and kinematic priors into generation; (3) a segmented concurrent generation strategy for constructing longer sequences under context-length limits; and (4) an experimental evaluation that analyzes both downstream diagnostic performance and the statistical characteristics of generated signals.

This paper presents MechaForge, a physics-guided LLM synthesis framework for bearing fault diagnosis. In contrast to purely empirical treatments of LLM synthesis, the present work grounds the generative process in mechanical and electromagnetic theory: the kinematic equations of bearing impacts and the principles of stator current modulation are translated into natural-language constraints, so that the LLM acts as a controllable, physics-guided synthesizer rather than an unconstrained sequence predictor. Within MechaForge, the framework compares four progressively constrained synthesis tracks, introduces a segmented concurrent generation algorithm to respect LLM context limits without spectral leakage, and provides both theoretical justification and experimental evidence that the generated sequences contain diagnostically useful, non-random structure rather than acting as mere labeled noise. In addition, the experimental study includes direct VAE- and GAN-based augmentation baselines under the same evaluation protocol so that the proposed LLM-based tracks can be interpreted against conventional generative benchmarks rather than against the real-only setting alone. The remainder of the paper is organized as follows. Section 2 develops the theoretical foundations (bearing kinematics, MCSA, and the mechanistic link to LLM in-context learning). Section 3 describes the MechaForge methodology and the four experimental tracks. Section 4 details the experimental setup; Section 5 presents and interprets the results in light of the theory; Section 6 discusses implications and limitations; and Section 7 concludes.

2. Theoretical Foundations and Phenomenological Mechanisms

A scientifically rigorous synthesis framework must be rooted in the physical phenomena governing the electromechanical system. Faithful generation of bearing fault signatures requires a precise mathematical description of contact kinematics, structural vibration decay, and electromagnetic coupling, as well as a principled link between these continuous processes and the discrete token-generation mechanism of an LLM. This section provides that foundation.

2.1. Kinematics of Rolling Element Bearings and Impulsive Vibration

The dynamic vibration response of a defective rolling element bearing is deterministic given geometry and shaft speed [33,34]. A rolling element bearing comprises an inner raceway, an outer raceway, rolling elements (balls or rollers), and a cage. A localized defect (e.g., fatigue spalling, pitting, or a micro-crack) causes the rolling elements to generate a sequence of periodic, high-energy impacts as they pass over the defect.

The repetition rates of these impacts are given by characteristic defect frequencies, which are the primary diagnostic signatures in the frequency domain. For shaft rotational frequency f_r (Hz), pitch diameter D, rolling element diameter d, contact angle θ, and n rolling elements, and assuming pure rolling without slip, the kinematic frequencies are:

Ball Pass Frequency of the Outer Race (BPFO), representing the frequency at which rolling elements pass a localized defect on the stationary outer raceway:

$$\text{BPFO} = \frac{nf_r}{2}\left(1 - \frac{d}{D}\cos\theta\right) \tag{1}$$

Ball Pass Frequency of the Inner Race (BPFI), representing the frequency at which a defect on the rotating inner raceway strikes the rolling elements:

$$\text{BPFI} = \frac{nf_r}{2}\left(1 + \frac{d}{D}\cos\theta\right) \tag{2}$$

Fundamental Train Frequency (FTF), representing the rotational frequency of the cage assembly:

$$\text{FTF} = \frac{f_r}{2}\left(1 - \frac{d}{D}\cos\theta\right) \tag{3}$$

Ball Spin Frequency (BSF), representing the frequency at which a rolling element rotates about its own central axis:

$$\text{BSF} = \frac{Df_r}{2d}\left(1 - \left(\frac{d}{D}\cos\theta\right)^2\right) \tag{4}$$

From a modeling perspective, a localized defect acts as a sudden displacement excitation: mechanical contact produces an instantaneous impulse that excites the high-frequency resonant modes of the bearing and housing. This is classically described by a multi-degree-of-freedom (MDOF) spring–mass–damper system [33]. Denoting the inner raceway by mass m_1 and the outer by m_2, the governing equations take the form:

$$m_1\ddot{x}_1 + c_1\dot{x}_1 - c_1\dot{x}_2 + k_{1x}x_1 - k_{1x}x_2 = F_x \tag{5}$$

$$m_1\ddot{y}_1 + c_1\dot{y}_1 - c_1\dot{y}_2 + k_{1y}y_1 - k_{1y}y_2 = F_y - W_s \tag{6}$$

$$m_2\ddot{x}_2 + (c_1 + c_{2x})\dot{x}_2 - c_1\dot{x}_1 + k_{2x}x_2 = F_x \tag{7}$$

$$m_2\ddot{y}_2 + (c_1 + c_{2y})\dot{y}_2 - c_1\dot{y}_1 + k_{2y}y_2 = F_y \tag{8}$$

where c and k are damping and stiffness parameters and F represents nonlinear contact forces from Hertzian theory. The underdamped response yields a time-domain signal that is not a pure sinusoid but an amplitude-modulated carrier: an abrupt shock followed by an exponential decay. The envelope spectrum exposes the impulse repetition rate, which must align with BPFO or BPFI for physical plausibility.

Inner-race faults add kinematic complexity: the defect is fixed to the rotating shaft and moves in and out of the radial load zone. Impulse amplitude therefore varies with angular position relative to the load vector, so the high-frequency impulse train is subject to secondary amplitude modulation at the shaft frequency f_r. Synthesizing inner-race faults thus requires capturing both the primary resonance and this low-frequency envelope [33].

2.2. Physical Mechanisms of Motor Current Signature Analysis

Vibration analysis measures mechanical displacement and acceleration; Motor Current Signature Analysis (MCSA) provides a non-invasive alternative via the stator current

waveform of the drive motor [35,36]. MCSA rests on electromechanical coupling between mechanical load and air-gap magnetic flux.

Bearing faults induce microscopic, periodic radial displacements of the rotor (dynamic air-gap eccentricity), altering the permeance of the rotating flux. In addition, impacts at the defect introduce torque micro-oscillations; the rotor's inertia acts as a low-pass filter, so these appear as small fluctuations in instantaneous angular velocity. These effects modulate the supply frequency f_e (50 Hz or 60 Hz). Phase modulation theory (Bessel expansion of the electromechanical equations) predicts that the stator current spectrum exhibits sidebands around f_e [35]. The fault-related sideband frequencies are:

$$f_{\text{current}} = |f_e \pm m \cdot f_{\text{fault}}|. \tag{9}$$

with m a positive harmonic index and f_{fault} the mechanical defect frequency (e.g., BPFO, BPFI). For synthetic multimodal data to be physically coherent, the vibration channel must exhibit impulsive, exponentially decaying shocks with high kurtosis, while the phase current channels must exhibit smooth, sinusoidal waveforms with sidebands obeying Equation (9). Random or spectrally misaligned current synthesis would fail to support MCSA-based diagnosis [35,37].

2.3. Mechanistic Interpretability: LLMs as Implicit Bayesian Synthesizers

Using language models—designed for discrete token prediction—to generate high-frequency continuous time series requires a theoretical link between LLM behavior and physical signal generation. Mechanistic interpretability and in-context learning (ICL) research suggest that LLM behavior can often be interpreted through a Bayesian lens, rather than through pure memorization alone [26,38]. In this view, pre-training maps a broad hypothesis space of generation functions, while few-shot demonstrations in the prompt condition this prior toward a task-relevant region. When the prompt includes numerical examples, the time series is represented explicitly as line-wise serialized multivariate sensor records rather than abstract tuples. Specifically, each time step is written as one text line in a fixed-order, comma-separated numerical format, and the full prompt is constructed by concatenating these lines as plain numerical text. The model output is required to follow the same format and is then parsed line by line back into a multichannel time-series array. Under this interpretation, the predictive behavior of ICL can be discussed in terms of a Bayes gap (deviation from the Bayes-optimal predictor) and posterior variance (inherent task uncertainty), although this remains an interpretive perspective rather than a formally proven mechanism in the present study.

For industrial synthesis, an unconstrained prompt leaves the prior relatively unguided and can produce physically implausible waveforms. Encoding physics in the prompt—including RMS ranges, kurtosis-related targets, cross-correlation bounds, and BPFO/BPFI-related spectral bands—can be interpreted as introducing prompt-level priors that shrink the hypothesis space and reduce the Bayes gap in the current setting. In the present study, these quantities are used as prompt-level priors rather than as learned or universally calibrated physical parameters. Accordingly, the Physics-Guided track should be interpreted as a dataset- and task-specific prompting configuration designed for the current experimental setting. The present study does not include a sensitivity analysis over these hyperparameters, and the effectiveness of the selected thresholds should therefore not be extrapolated beyond the current dataset and operating conditions.

3. Methodology: The MechaForge Framework

Building on the theory of Section 2, MechaForge implements a multi-strategy methodology for synthesizing bearing fault time series, organized into four progressively con-

strained tracks shown in Figure 1: Baseline, Basic LLM, Multi-Stage LLM, and Physics-Guided LLM. Within this architecture, real data is used not only for baseline training but also as a reference for all synthesis tracks. A single real dataset feeds the baseline CNN and all synthesis strategies: Basic and Multi-Stage LLM use it as few-shot or analysis input, while Physics-Guided LLM additionally imposes kinematic and MCSA-derived constraints. Synthetic and real data are merged only for training and then evaluated with the same CNN on the same held-out test set, which keeps the test set uncontaminated and ensures that performance gains are attributable to the synthesis strategy.

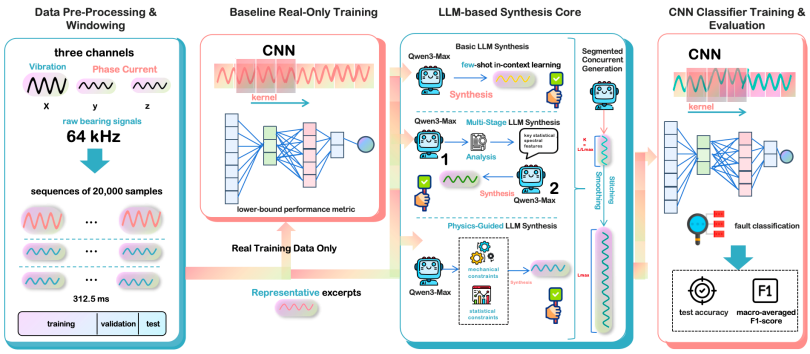

Figure 1. Overview of the MechaForge pipeline.

3.1. Progressive Constraint Integration Strategies

To isolate the effect of domain knowledge on synthesis quality and downstream diagnosis, the framework is organized into four experimental tracks that progressively inject constraints: from a real-data-only baseline and unconstrained few-shot LLM generation, through semantic reasoning, to physics-guided generation with explicit kinematic and MCSA-derived targets. This design allows a comparison of how prompt engineering may reduce the effective Bayes gap in the current setting, under the Bayesian interpretation of Section 2.3, while also illustrating the trade-off between structural fidelity and diversity. The four tracks are as follows.

Track 1—Real-Only Baseline. A 1D CNN is trained only on the scarce real dataset. This gives a lower bound on performance under data scarcity.

Track 2—Basic LLM Synthesis (Few-Shot Mimicry). The prompt contains two to three short real fault snippets (100–200 samples) as line-wise serialized multivariate sensor records in a fixed-order, comma-separated numerical format. The LLM is asked to generate a new sequence matching the energy, noise level, and statistical character of the examples, without explicit physical constraints. This probes unconstrained in-context learning and the model's ability to infer kinematic structure from raw numbers.

Track 3—Multi-Stage LLM (Semantic Reasoning). A two-stage, chain-of-thought design is used. Stage 1: the LLM analyzes a real snippet and outputs a short natural-language summary of its statistical and spectral topology (periodicity, dominant frequencies, amplitude distribution). Stage 2: that summary is used as the conditioning context for numerical synthesis. This mirrors semantic-alignment ideas in Thoth and ChatTS [28,31]. Figure 2 illustrates the flow. Stage 1 converts a short real snippet into a compact textual description (e.g., periodicity, RMS, dominant frequencies), which then acts as the sole conditioning for Stage 2 numerical generation and subsequent downstream training/evaluation; the quality of the synthetic signal therefore depends critically on the accuracy of Stage 1. A risk is error propagation: if Stage 1 misidentifies impulse frequency or RMS, Stage 2 amplifies the error; in the present experiments, Multi-Stage underperformed Basic LLM in part because analysis errors propagated into synthesis.

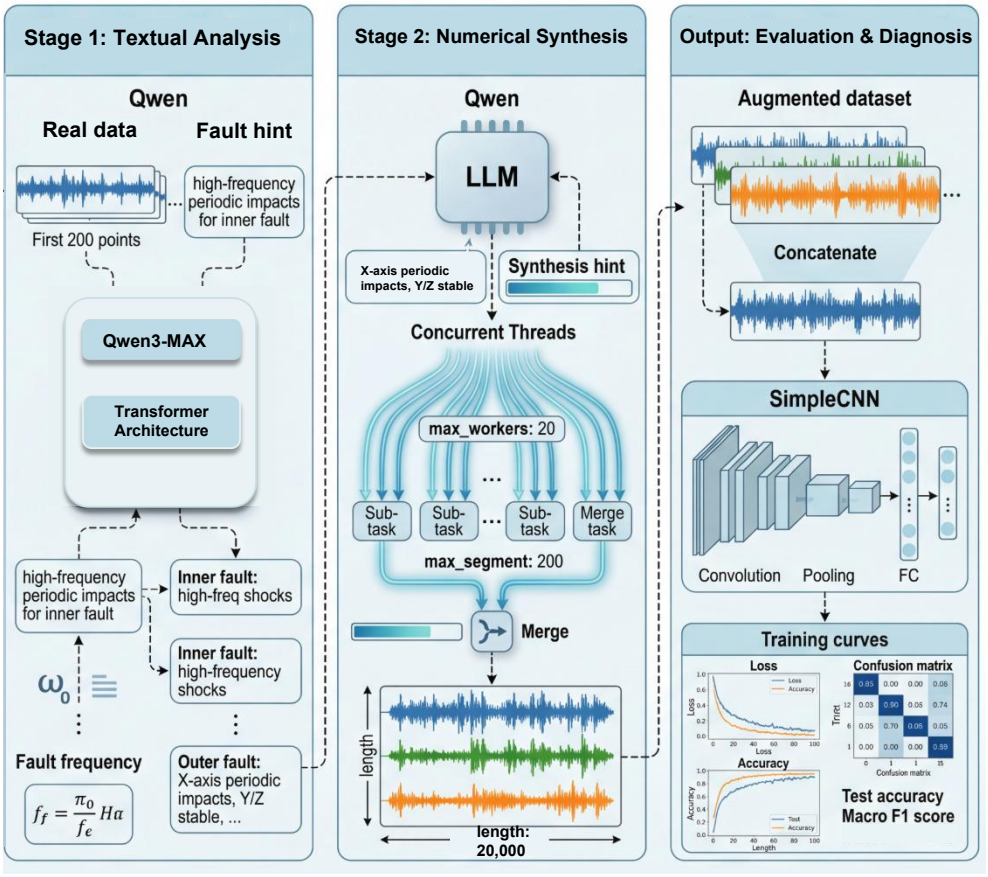

Figure 2. Multi-stage LLM pipeline. Ellipses indicate omitted repeated worker branches and abbreviated schematic content for visual clarity.

Track 4—Physics-Guided LLM. The prompt encodes the mechanical and electromagnetic constraints of Section 2: explicit targets for statistical descriptors and spectral bands derived from signal processing and kinematics. This corresponds to supplying prompt-level physical priors that are intended to reduce the Bayes gap in the current experimental setting (Section 2.3).

3.2. Mathematical Formulation of Physical Constraints

The Physics-Guided prompt limits the LLM's generative hypothesis space by demanding adherence to the following explicitly calculated mathematical targets:

Root Mean Square (RMS): Dictates the total continuous kinetic energy of the mechanical signal, ensuring the synthetic severity matches the target fault stage.

$$\mathrm{RMS}(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{10}$$

Peak Amplitude: Governs the maximum severity of the transient impact forces generated during the Hertzian contact phase.

$$\mathrm{Peak}(x) = \max_i |x_i| \tag{11}$$

Kurtosis: Defined as the fourth standardized statistical moment, kurtosis serves as a hyper-sensitive indicator of signal impulsivity. While healthy bearings approximate a standard Gaussian distribution with a kurtosis near 3, severe localized mechanical faults yield highly impulsive, heavy-tailed distributions with kurtosis values frequently exceeding

5 to 10. The LLM is prompted to target elevated kurtosis for the vibration channel as a form of generation guidance rather than as a hard numerical guarantee.

$$K(x) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^2} \tag{12}$$

Pearson cross-correlation. To reflect the MCSA decoupling of Section 2.2, the prompt encourages low cross-correlation between the vibration and current channels, with $|\rho(X,Y)|$ and $|\rho(X,Z)| < 0.3$, together with moderate correlation between the two phase-current channels, $|\rho(Y,Z)| \approx 0.5$–0.7 [35]:

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{13}$$

Spectral band targets. The prompt requires periodic transients at kinematically consistent BPFO/BPFI bands (e.g., 80–120 Hz) and elevated energy in structural resonance bands (1–10 kHz), avoiding spectral hallucination typical of unconstrained generation [33,35]. Figure 3 summarizes how these constraints—together with explicit peak-amplitude limits and cross-correlation bounds—narrow the hypothesis space and improve spectral fidelity and physical plausibility, at the cost of some diversity. Encoding the mathematical targets of Section 2 (RMS, peak amplitude, kurtosis, correlation bounds via ρ, and BPFO/BPFI bands) directly in the prompt restricts the LLM's generative distribution toward physically plausible regions in the present setting and can reduce the Bayes gap, though the resulting diversity is lower than in the unconstrained Basic LLM setting. In the present study, these quantities are used as prompt-level priors rather than as learned or universally calibrated physical parameters. Accordingly, the Physics-Guided track should be interpreted as a dataset- and task-specific prompting configuration. The current study does not include a sensitivity analysis over these hyperparameters, and the effectiveness of the selected thresholds should therefore not be extrapolated beyond the current dataset and operating conditions.

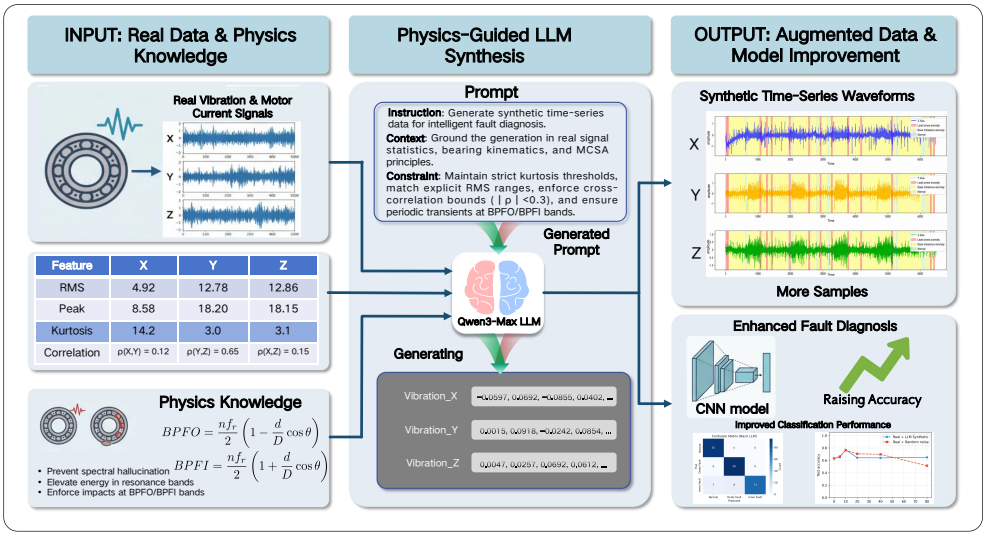

Figure 3. Physics-guided LLM synthesis. Ellipses denote abbreviated numeric sequences or repeated content for schematic presentation.

3.3. Algorithmic Resolution of Spectral Leakage: Segmented Concurrent Generation

A fundamental architectural constraint of autoregressive transformer models is the finite context window limit. A single inference call can typically generate only a short

numerical segment once formatting and delimiter tokens are accounted for. At the sampling rate used here, that duration is insufficient for capturing the longer mechanical modulations relevant to downstream diagnosis.

The segmented concurrent generation strategy with overlap-averaging is introduced here as a practical response to transformer context-length constraints and segment-boundary discontinuities. Its role in the present framework is to construct longer sequences while reducing obvious boundary artifacts through linear cross-fading in the overlap region. However, the current study does not include a direct comparison against alternative stitching, padding, or sequence-extension strategies, and the method should therefore be understood as an engineering-motivated design choice under the present implementation setting rather than as a generally optimal reconstruction scheme.

To obtain long 20,000-sample sequences, MechaForge generates shorter segments and then blends them in the overlap region. Naive concatenation would introduce discontinuities at boundaries; in the frequency domain this corresponds to convolution with a Dirichlet (sinc) kernel and causes spectral leakage, smearing energy across bins and distorting fault signatures and MCSA sidebands. To preserve C^0 continuity and reduce leakage, overlapping segments are blended with a linear cross-fade. The final signal in the overlap region is

$$\hat{s}(i) = \alpha(i)\, s_k(i) + (1 - \alpha(i))\, s_{k+1}(i), \tag{14}$$

with $\alpha(i)$ decreasing linearly from 1 to 0 across the boundary.

Algorithm 1 summarizes the segmented concurrent generation and overlap-averaging procedure used in the present framework.

Algorithm 1 Segmented Concurrent Generation with Overlap-Averaging

Input: target length N, segment length L_{seg}, overlap length L_{ov}, prompt template P, maximum concurrency C
Output: synthesized sequence $\hat{s}[1:N]$

1. Set $\Delta = L_{\text{seg}} - L_{\text{ov}}$.
2. Compute $K = \lceil (N - L_{\text{ov}})/\Delta \rceil$.
3. For $k = 1, 2, \ldots, K$ in parallel with at most C workers, build a segment-specific prompt P_k from P, generate segment s_k with target length L_{seg}, parse the textual output into a numeric multichannel sequence, and reject/regenerate the segment if parsing fails or length is invalid.
4. Initialize $\hat{s} = s_1$.
5. For $k = 2, 3, \ldots, K$, blend the overlap with coefficients $\alpha_i = 1 - (i-1)/(L_{\text{ov}} - 1)$ for $i = 1, 2, \ldots, L_{\text{ov}}$, then append the non-overlap part $s_k[L_{\text{ov}} + 1 : L_{\text{seg}}]$ to $\hat{s}$.
6. Return the first N samples of $\hat{s}$.

4. Experimental Setup

4.1. Dataset and Pre-Processing

Experiments use the Paderborn Bearing Dataset [35]: 64 kHz vibration (X) and dual-phase motor current (Y, Z) from naturally degraded bearings. Three diagnostic classes are used: Normal (K001), Outer-race (KA01–KA08), and Inner-race (KI01–KI08). Raw signals are segmented into 20,000-sample full-length sequences (≈312.5 ms), and the subsequent CNN training uses 512-sample windows with stride 128. The held-out real test set contains 276 windows. The dataset exhibits realistic noise, variable speed (900–2100 RPM), class imbalance, and intra-class severity spread. Although the Paderborn dataset spans operating speeds of approximately 900–2100 RPM, the present study does not adopt an explicitly speed-stratified train/test protocol. The current experiments therefore evaluate the framework under the existing split setting, rather than under a dedicated cross-speed generalization setup.

4.2. Implementation Details

Synthesis is performed with Qwen3-Max [39] through track-specific prompt templates (API; temperature 0.5–0.7; up to 20 concurrent threads; 4096 tokens per segment). Each generated response is expected to follow a fixed-order, comma-separated multichannel format, where one text line corresponds to one time step. During post-processing, the raw model output is parsed line by line with a regular-expression-based parser; empty or malformed lines are discarded, and parsing stops once the target sequence length is reached. Group-based data splitting is performed before augmentation to reduce leakage between training, validation, and testing, and the data split uses a fixed random seed, whereas the API-based LLM generation stage remains stochastic under the present endpoint configuration. The reported synthetic data are generated for the current experimental runs under the stated configuration, rather than averaged over repeated regeneration cycles. Full prompt templates, output-format rules, parsing logic, and track-specific post-processing criteria are provided in Supplementary File S1. For direct benchmarking, the VAE and GAN baselines are trained on the same real training split and use the same synthetic augmentation budget as the LLM-based tracks; the resulting synthetic windows are then incorporated under the same downstream CNN protocol.

The diagnostic model is a standard 1D CNN (three blocks, 32/64/128 filters, kernel size 3), trained with Adam (learning rate 5×10^{-4}), batch size 64, and early stopping (patience 8) on validation accuracy [4,40]. Using a single fixed architecture ensures that performance differences across tracks reflect the value of the synthetic data rather than model capacity.

4.3. Evaluation Metrics

Performance is measured by test accuracy and macro-averaged F1-score. Macro-F1 is the unweighted mean of per-class F1, so it is sensitive to minority classes and appropriate for imbalanced fault diagnosis. Given that the current evaluation is conducted on a held-out real test set of 276 windows, the reported accuracy and macro-F1 values in this study are presented as descriptive performance measures under the present experimental configuration. Because both LLM-based synthesis and CNN training involve stochastic components, the current revision does not include repeated-seed uncertainty estimates, confidence intervals, or formal significance testing, and the corresponding performance differences should therefore be interpreted with appropriate caution.

5. Results and Analysis

All synthesis strategies are merged with the real training set and evaluated on a held-out real test set of 276 windows. The results are interpreted in light of the theoretical framework of Section 2 and the design of Section 3.

5.1. Overall Diagnostic Performance

Table 1 summarizes the direct benchmark under the unified evaluation protocol. Because the held-out test set contains 276 windows and the present revision does not include a repeated-trial significance study, the performance differences in Table 1 should be interpreted as empirical observations under the current setup rather than as statistically established robustness estimates. The real-only baseline attains 52.54% accuracy, confirming that data scarcity severely limits the CNN's ability to learn generalizable features. Among the evaluated augmentation methods, the Basic LLM track performs best, reaching 78.62% accuracy and 0.7648 macro-F1. The newly added VAE and GAN baselines both reach 74.28% accuracy, with macro-F1 values of 0.7202 and 0.7257, respectively, indicating that conventional generative augmentation is also beneficial under the same protocol,

but remains below the Basic LLM track in the present configuration. Physics-Guided LLM reaches 71.00% accuracy and 0.7000 macro-F1; this suggests that prompt-level physical priors can improve plausibility, but in the current setting they also narrow the support of the generated distribution and do not outperform the added VAE/GAN baselines. Multi-Stage LLM attains 61.23% accuracy and 0.6177 macro-F1, and its underperformance is interpreted in part through error propagation across the two-stage pipeline, because Stage 2 depends on an intermediate textual abstraction rather than on direct access to the original measured sequence.

Table 1. Test-set performance under the unified benchmark protocol (276 windows).

Method	Test Accuracy	Macro F1	Δ Acc (%)
Baseline (Real-Only)	0.5254	0.5031	–
Basic LLM (Few-Shot)	0.7862	0.7648	+26.08
VAE	0.7428	0.7202	+21.74
GAN	0.7428	0.7257	+21.74
Physics-Guided LLM	0.7100	0.7000	+18.46
Multi-Stage LLM	0.6123	0.6177	+8.69

5.2. Per-Class Performance and Relation to Kinematics

Per-class F1 scores (Table 2) and the Basic LLM confusion matrix (Figure 4) reflect the kinematics of Section 2.1.

Table 2. Per-class F1 scores highlighting varied sensitivities to the synthesis strategies.

Method	Normal Class	Outer-Race Fault	Inner-Race Fault
Baseline	0.48	0.55	0.47
Basic LLM	0.85	0.85	0.59
VAE	0.75	0.90	0.51
GAN	0.74	0.89	0.54
Physics-Guided LLM	0.73	0.78	0.59
Multi-Stage LLM	0.50	0.89	0.46

Outer-race faults (BPFO) achieve the highest F1 across the evaluated methods (0.78–0.90 in Table 2): the outer race is stationary, so impact mechanics are consistent and the signature is a stable periodic impulse train that both the generative models and the classifier handle well [33,35]. Inner-race faults (BPFI) cap at F1 0.59: as in Section 2.1, the defect rotates with the shaft through the load zone, causing amplitude modulation and phase shifts that complicate both synthesis and discrimination. The Normal class benefits from synthesis (e.g., Basic LLM F1 0.85), reducing false alarms. The confusion matrix for the Basic LLM model (Figure 4), evaluated on the same held-out test set of 276 windows, shows strong diagonal entries for Normal and Outer-Race and off-diagonal mass involving Inner-Race; rows denote true classes and columns denote predicted classes, consistent with these kinematic differences and the per-class F1 in Table 2.

To assess whether the classification gains correspond to physically plausible synthetic signals, we next compare per-window time-domain statistics of real and synthetic data (Figure 5).

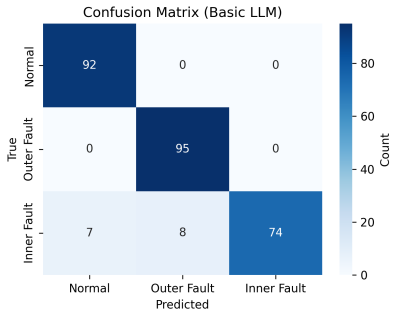

Figure 4. Basic LLM confusion matrix.

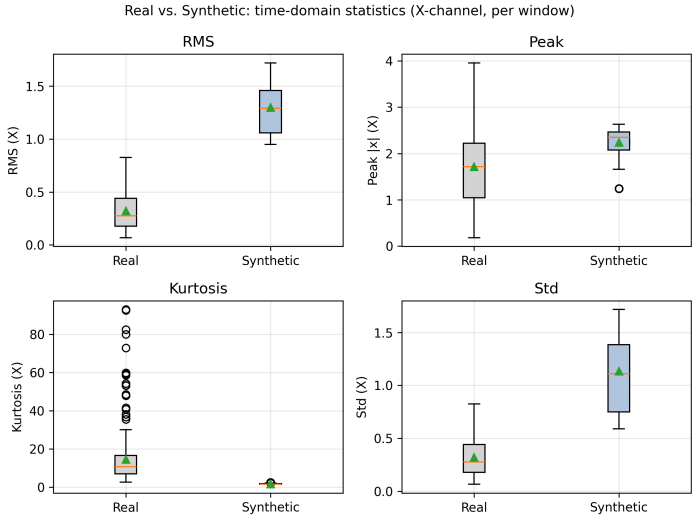

Figure 5. Per-window time-domain statistics of real training data and the Basic LLM synthetic track (X-channel only). The green triangles indicate means, the yellow central lines indicate medians, and the circles indicate outliers.

5.3. Real vs. Synthetic: Time-Domain Statistics

Per-window statistics (RMS, peak, kurtosis, std) on the X-channel are computed for real training data and the Basic LLM synthetic track only. The comparison in Figure 5 does not aggregate multiple synthesis strategies unless explicitly stated elsewhere. Synthetic windows show systematically higher RMS, peak, and std, indicating an amplitude/normalization mismatch with the real sensor range. Crucially, real fault signals exhibit high kurtosis (median in the teens, outliers to tens) corresponding to impulsive, heavy-tailed impacts, whereas synthetic data is near-Gaussian (kurtosis near 0–3). The discrepancies in RMS, peak, std, and especially kurtosis indicate that the current Basic LLM pipeline does not yet provide strict distribution matching to the real signal distribution. In particular, the present implementation relies on prompt-level guidance and basic text-to-sequence post-processing, but does not include an explicit normalization–calibration module, a higher-order moment matching objective, or an acceptance filter based on impulsivity-related statistics. Accordingly, the generated signals should be interpreted here as diagnostically useful augmentations rather than as distribution-matched surrogates of the real data. For clarity, the kurtosis target introduced in the Physics-Guided prompt should be interpreted as generation guidance rather than a hard numerical guarantee; it therefore should not be conflated with the Basic LLM-only statistics shown in Figure 5. Despite this scale and shape gap, the control experiment below indicates that the classifier still extracts discriminative structure from synthetic data rather than treating it as uninformative noise. The time-domain statistics in Figure 5 provide only partial evidence about the generated signals, and they should not be interpreted as a direct validation

of frequency-domain physical fidelity. In particular, the present study does not include dedicated evaluation of envelope-spectrum consistency, high-frequency phase alignment, or fault-band preservation. Accordingly, the current analysis supports the view that the generated sequences contain diagnostically useful structure, but it does not establish that they faithfully reproduce the full frequency-domain behavior of real fault signals.

The gap in scale and kurtosis shown in Figure 5 motivates the following control: we compare augmentation with the same number of LLM synthetic windows versus labeled random noise to test whether synthetic data contributes discriminative structure or merely acts as regularization.

5.4. Control Experiment: Synthetic Data vs. Random Noise

A critical unresolved issue pertains to the exact mechanism by which LLM-generated sequences enhance diagnostic performance: specifically, whether they genuinely map the physical fault kinematics, or merely act as a form of data augmentation that regularizes the classification model through labeled noise. Table 3 and Figures 6 and 7 report a controlled comparison: the same number of extra windows per class are added either as LLM synthetic data or as Gaussian noise (matched per-channel variance) with the same class labels. At low volume (10 extra windows per class), both improve over baseline, but synthetic improves more (0.619 vs. 0.607). At high volume (80 per class), synthetic maintains accuracy at 0.65 while the same amount of random noise reduces accuracy to 0.51. This control experiment follows a separate protocol from the main benchmark in Table 1; accordingly, the baseline values reported in Table 3 should not be interpreted as numerically identical to those in the main comparison table. The control experiment therefore supports the conclusion that the LLM-generated sequences are not equivalent to labeled random noise, because they preserve diagnostically useful structure under augmentation. However, to avoid overstating the evidence, the stronger manifold interpretation is replaced here with a more cautious formulation: the synthetic data exhibit non-random, diagnosis-relevant structure under the present experimental setting.

Table 3. Control experiment: real only vs. real + LLM synthetic vs. real + random noise (same number of extra windows per class).

Condition	Test Accuracy	Macro F1
Baseline (0 extra)	0.583	0.551
Real + LLM Synthetic (10)	0.619	0.595
Real + Random Noise (10)	0.607	0.581
Real + LLM Synthetic (80)	0.650	0.641
Real + Random Noise (80)	0.510	0.490

Figure 6 quantifies this contrast under a fixed augmentation budget of 80 extra windows per fault class. The curves show that when the classifier is trained on real data plus LLM-generated synthetic windows, test accuracy remains around 0.65—a clear gain over the baseline within this control setting. Under the same protocol but with the synthetic windows replaced by Gaussian noise (matched per-channel variance and identical class labels), accuracy drops to approximately 0.51, i.e., below the no-augmentation baseline. This divergence indicates that the benefit of augmentation is not due to extra training samples per se, but to the fact that the LLM output preserves diagnosis-relevant structure that random noise does not.

To examine how this effect depends on the amount of augmented data, Figure 7 plots test accuracy against the number of extra windows added per fault class for both the synthetic and the noise conditions. At low volume (e.g., 10–20 windows per class), both

forms of augmentation can yield modest gains; as the number of augmented windows increases, however, the two curves separate, with synthetic augmentation remaining around or above 0.65 while the noise condition declines and eventually falls below the baseline. Taken together, Figures 6 and 7 support the more cautious conclusion that LLM-generated augmentation provides non-random, diagnosis-relevant structure under the present control setting, rather than behavior equivalent to labeled noise.

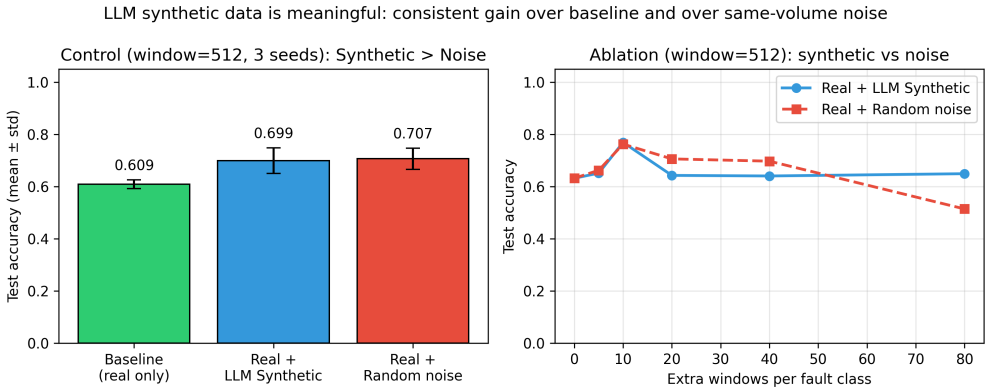

Figure 6. Control experiment: LLM synthetic vs. random noise.

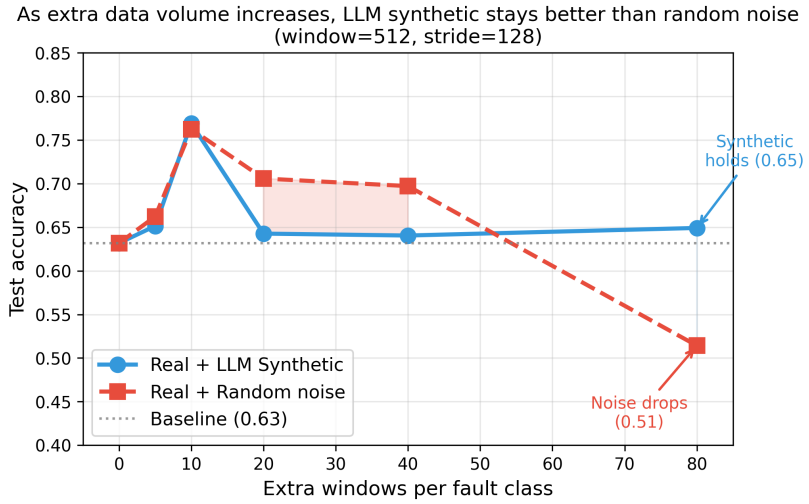

Figure 7. Test accuracy vs. augmentation volume: LLM synthetic vs. random noise.

6. Discussion

The theoretical developments and empirical findings of the MechaForge framework advance the intersection of natural language processing and physical cyber-system diagnostics. The results indicate that Large Language Models can support data-driven fault diagnosis under data scarcity, but the present evidence should be interpreted with appropriate caution in light of the single-dataset evaluation setting and the stochastic nature of the generation pipeline [1,41,42].

The time-domain statistics comparison (Figure 5) shows that synthetic data has higher RMS, peak, and std with tighter distributions and much lower kurtosis than real data—i.e., it differs in both scale and impulsivity—raising the question of whether such data is merely "noise." The control experiment (Table 3) and the volume ablation (Figures 6 and 7) address this directly: at moderate volume, adding LLM synthetic windows improves test accuracy more than the same amount of noise; at high volume (80 windows per fault class), synthetic keeps accuracy at 0.65 whereas noise reduces it to 0.51. These observations should therefore be interpreted primarily in terms of diagnostic utility rather than strict physical realism.

Under the current autoregressive text-to-numerical generation paradigm, the model can provide non-random signal structure that benefits downstream classification, but it is not yet expected to guarantee strict frequency-domain fidelity comparable to that of dedicated physics-based simulators or controlled signal-generation frameworks. The present evidence thus supports useful augmentation for diagnosis, rather than a fully validated claim of frequency-domain physical equivalence.

A central finding is the effectiveness of physics-guided prompting. By specifying generative boundaries from kinematic defect frequencies (BPFO, BPFI) and MCSA modulation principles, the framework constrains a model trained primarily on linguistic data to respect mechanical priors under prompt control. Accordingly, the present framework should be understood as combining physically motivated prompt guidance with limited post-generation screening, rather than as a fully constrained numerical generator. The current evidence therefore supports physics-informed sequence generation under prompt control, but not yet a strict constrained-generation interpretation in the optimization or decoding sense. Several refinements may improve real–synthetic alignment beyond the current implementation. Plausible directions include recalibrating amplitude ranges to the real sensor scale, adding post-generation acceptance filtering based on RMS/peak/kurtosis intervals, and introducing stronger impulsivity or spectral-band constraints to better preserve higher-order statistics. These mechanisms are not part of the present pipeline, and the remaining mismatch observed in Figure 5 should therefore be regarded as an open limitation rather than a resolved aspect of the method.

The underperformance of the Multi-Stage track is plausibly explained by error propagation across the two-stage pipeline. Because the numerical synthesis step depends on an intermediate textual abstraction rather than on direct access to the original measured sequence, inaccuracies in Stage 1 can propagate into Stage 2 and adversely affect downstream classification. Potential mitigation strategies include template-constrained summarization, explicit reinjection of measured signal statistics into the second-stage prompt, and rejection or filtering of generations that violate basic signal constraints. These strategies are not yet evaluated in the current study, and are therefore presented as plausible refinement directions rather than validated performance improvements.

An important observation of the present study is that physics-guided prompting can improve the structure of generated signals under the current experimental setting by introducing kinematic and MCSA-related priors into the prompt. However, the corresponding statistical and spectral targets are manually specified guidance parameters, and their sensitivity to alternative settings, operating conditions, and sensing configurations has not been systematically evaluated here. Accordingly, the current results should be interpreted as evidence within the present dataset and prompt design, rather than as a definitive demonstration of cross-condition generalizability. Future work should examine the robustness of these guidance hyperparameters under varying speeds, sensor configurations, sampling conditions, and external datasets before broader generalizability claims are made.

Accordingly, the reported performance gains should not be interpreted as evidence of robustness to unseen operating-speed conditions. A more direct assessment of cross-speed transfer would require training and testing on deliberately separated speed regimes, which is outside the scope of the present revision. The empirical validation in this study is limited to the Paderborn bearing dataset, and therefore the current findings should be interpreted within this sensing and operating setting rather than as a direct demonstration of cross-dataset or cross-domain transferability. Extending the framework to other benchmarks would require nontrivial harmonization of sampling rates, sensor configurations, fault taxonomies, and prompt/guidance settings, which is outside the scope of the present revision.

Several directions for future work remain. The difficulty in synthesizing the load-modulated signatures of inner-race faults suggests that static, text-based prompts may have limited expressiveness for highly non-stationary, multi-component dynamics. Future work could explore hybrid pipelines that combine LLM generation with differential-equation solvers or physics-informed discriminators to enforce time-varying envelope constraints during inference. Compound and simultaneous faults (e.g., bearing spalling with gear pitting or shaft misalignment) would require compositional, multi-label synthesis; extending frameworks such as ChatTS [28] and STELLA [43] to such settings is a natural direction. Addressing these gaps would strengthen LLM-based synthesis for digital twins and reliability engineering [12,41].

7. Conclusions

This paper has presented MechaForge, a physics-guided LLM-based synthesis framework for bearing fault diagnosis under data scarcity. Within the present evaluation setting, the Basic LLM track achieved the strongest overall performance, exceeding both the real-only baseline and the added VAE/GAN conventional generative baselines under the same downstream CNN protocol. At the same time, the Multi-Stage strategy revealed a clear error-propagation issue, and the Physics-Guided track illustrated how manually specified priors can shape the generated sequences under prompt control without necessarily outperforming other augmentation methods in the current configuration. The control experiment further suggested that the generated data contains diagnostically useful, non-random structure rather than behaving as labeled noise. Nevertheless, the present results should still be interpreted with appropriate caution: frequency-domain physical fidelity has not been directly validated, and the experimental evidence remains limited to a single dataset and prompt configuration. Therefore, the current study should be understood as demonstrating the promise of LLM-based diagnostic augmentation in the present setting, rather than as a definitive proof of universal superiority over all generative approaches or of cross-dataset transferability.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app16094566/s1, Supplementary File S1: Prompt templates, output-format rules, parsing logic, and track-specific post-processing criteria for the Basic LLM, Multi-Stage, and Physics-Guided tracks.

Author Contributions: Conceptualization, X.Z. and X.L.; methodology, X.Z., F.L., and Y.H.; software, Y.H.; validation, Y.H. and D.Y.; formal analysis, F.L.; investigation, X.Z., D.Y., and Y.M.; resources, F.L. and Y.M.; data curation, X.Z. and Y.H.; writing—original draft preparation, X.Z.; writing—review and editing, X.L., F.L., Y.H., D.Y., and Y.M.; visualization, D.Y.; project administration, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Liaoning Provincial Major Science and Technology Program, grant number 2025JH1/11700022.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: Authors Xiyang Zhang, Yi Hu, Dong Yu were employed by the Shenyang CASNC Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wen, L.; Zhang, Y.; Hu, W.; Li, X. The survey of industrial anomaly detection for industry 5.0. *Int. J. Comput. Integr. Manuf.* **2024**, 1–22. [CrossRef]
2. Xia, Y.; Yang, N.; Wang, H.; Wang, X.; Cui, M.; Li, M. Failure propagation prediction of complex electromechanical systems based on interdependence. *Electronics* **2023**, *12*, 2600. [CrossRef]
3. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. [CrossRef]
4. Wen, L.; Li, X.; Gao, L.; Zhang, Y. A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Trans. Ind. Electron.* **2017**, *65*, 5990–5998. [CrossRef]
5. Bono, F.M.; Cinquemani, S.; Chatterton, S.; Pennacchi, P. A Deep Learning Approach for Fault Detection and RUL Estimation in Bearings. In *Proceedings of the NDE 4.0, Predictive Maintenance, and Communication and Energy Systems in a Globally Networked World*; Proceedings of SPIE; SPIE: Bellingham, WA, USA, 2022; Volume 12049, p. 1204908. [CrossRef]
6. Radicioni, L.; Bono, F.M.; Cinquemani, S. Vibration-Based Anomaly Detection in Industrial Machines: A Comparison of Autoencoders and Latent Spaces. *Machines* **2025**, *13*, 139. [CrossRef]
7. Zhang, Z.; Li, S.; Wang, J.; Xin, Y.; An, Z.; Jiang, X. Enhanced sparse filtering with strong noise adaptability and its application on rotating machinery fault diagnosis. *Neurocomputing* **2020**, *398*, 31–44. [CrossRef]
8. Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mech. Syst. Signal Process.* **2018**, *104*, 799–834. [CrossRef]
9. Hakim, M.; Omran, A.A.B.; Ahmed, A.N.; Al-Waily, M.; Abdellatif, A. A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations. *Ain Shams Eng. J.* **2023**, *14*, 101945. [CrossRef]
10. Nectoux, P.; Gouriveau, R.; Medjaher, K.; Ramasso, E.; Chebel-Morello, B.; Zerhouni, N.; Varnier, C. PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In *Proceedings of the IEEE International Conference on Prognostics and Health Management, PHM'12. IEEE Catalog Number: CPF12PHM-CDR*; IEEE: Piscataway, NJ, USA, 2012; pp. 1–8.
11. Smith, W.A.; Randall, R.B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64*, 100–131. [CrossRef]
12. Zhao, J.; Wang, W.; Huang, J.; Ma, X. A comprehensive review of deep learning-based fault diagnosis approaches for rolling bearings: Advancements and challenges. *AIP Adv.* **2025**, *15*, 020702. [CrossRef]
13. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27, pp. 2672–2680.
14. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
15. Hu, T.; Tang, T.; Lin, R.; Chen, M.; Han, S.; Wu, J. A simple data augmentation algorithm and a self-adaptive convolutional architecture for few-shot fault diagnosis under different working conditions. *Measurement* **2020**, *156*, 107539. [CrossRef]
16. Ding, Y.; Ma, L.; Ma, J.; Wang, C.; Lu, C. A generative adversarial network-based intelligent fault diagnosis method for rotating machinery under small sample size conditions. *IEEE Access* **2019**, *7*, 149736–149749. [CrossRef]
17. Guo, Q.; Li, Y.; Liu, Y.; Gao, S.; Song, Y. Data Augmentation for Intelligent Mechanical Fault Diagnosis Based on Local Shared Multiple-Generator GAN. *IEEE Sens. J.* **2022**, *22*, 9598–9609. [CrossRef]
18. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29, pp. 2234–2242.
19. Lee, S.; Jeong, H.; Kwon, J. Transformer-based GAN with Multi-STFT for rotating machinery vibration data analysis. *Electronics* **2024**, *13*, 4253. [CrossRef]
20. Dong, G.S.; Wan, H.P.; Luo, Y.; Todd, M.D. A fast sparsity-free compressive sensing approach for vibration data reconstruction using deep convolutional GAN. *Mech. Syst. Signal Process.* **2023**, *188*, 109937. [CrossRef]
21. Zhu, C.; Lin, W.; Zhang, H.; Cao, Y.; Fan, Q.; Zhang, H. Research on a bearing fault diagnosis method based on an improved wasserstein generative adversarial network. *Machines* **2024**, *12*, 587. [CrossRef]
22. Jiang, Y.; Pan, Z.; Zhang, X.; Garg, S.; Schneider, A.; Nevmyvaka, Y.; Song, D. Empowering time series analysis with large language models: A survey. *arXiv* **2024**, arXiv:2402.03182. [CrossRef]
23. Zhang, X.; Chowdhury, R.R.; Gupta, R.K.; Shang, J. Large language models for time series: A survey. *arXiv* **2024**, arXiv:2402.01801. [CrossRef]
24. Abdullahi, S.; Usman Danyaro, K.; Zakari, A.; Abdul Aziz, I.; Amila Wan Abdullah Zawawi, N.; Adamu, S. Time-Series Large Language Models: A Systematic Review of State-of-the-Art. *IEEE Access* **2025**, *13*, 30235–30261. [CrossRef]
25. Kottapalli, S.R.K.; Hubli, K.; Chandrashekhara, S.; Jain, G.; Hubli, S.; Botla, G.; Doddaiah, R. Foundation models for time series: A survey. *arXiv* **2025**, arXiv:2504.04011. [CrossRef]

26. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
27. Lin, H.; Yu, M. A novel distributed PV power forecasting approach based on Time-LLM. *arXiv* **2025**, arXiv:2503.06216. [CrossRef]
28. Xie, Z.; Li, Z.; He, X.; Xu, L.; Wen, X.; Zhang, T.; Chen, J.; Shi, R.; Pei, D. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv* **2024**, arXiv:2412.03104. [CrossRef]
29. Lee, J.; Jeon, H.; Kim, U.; Kim, M. BearGen: LLM-guided signal generation framework for bearing fault diagnosis. *Adv. Eng. Inform.* **2026**, *71*, 104400. [CrossRef]
30. Rousseau, C.; Boschi, T.; Cornacchia, G.; Salwala, D.; Pascale, A.; Moreno, J.B. Forging Time Series with Language: A Large Language Model Approach to Synthetic Data Generation. *arXiv* **2025**, arXiv:2505.17103. [CrossRef]
31. Lin, J.; Wang, Y.; Wu, J.; Luo, H.; Pei, Z.; Wang, J. Thoth: Mid-Training Bridges LLMs to Time Series Understanding. *arXiv* **2026**, arXiv:2603.01042. [CrossRef]
32. Rezazadeh, N.; Caputo, F.; Aversano, A.; Lamanna, G.; De Luca, A.; Perfetto, D. Prototype-attention domain adaptation for explainable bearing fault diagnosis. *Procedia Struct. Integr.* **2026**, *80*, 411–417. [CrossRef]
33. Randall, R.B.; Antoni, J. Rolling element bearing diagnostics—A tutorial. *Mech. Syst. Signal Process.* **2011**, *25*, 485–520. [CrossRef]
34. Cao, H.; Su, S.; Jing, X.; Li, D. Vibration mechanism analysis for cylindrical roller bearings with single/multi defects and compound faults. *Mech. Syst. Signal Process.* **2020**, *144*, 106903. [CrossRef]
35. Lessmeier, C.; Kimotho, J.K.; Zimmer, D.; Sextro, W. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *Proceedings of the PHM Society European Conference*; PHM Society: Bilbao, Spain, 2016; Volume 3, Issue 1. [CrossRef]
36. Leite, V.; da Silva, J.G.B.; Torres, G.L.; Veloso, G.F.C.; da Silva, L.E.B.; Bonaldi, E.L.; de Oliveira, L. Bearing fault detection in induction machine using squared envelope analysis of stator current. In *Bearing Technology*; InTech: London, UK, 2017.
37. Antoni, J.; Randall, R.B. The spectral kurtosis: Application to the vibratory surveillance and diagnostics of rotating machines. *Mech. Syst. Signal Process.* **2006**, *20*, 308–331. [CrossRef]
38. Xie, S.M.; Raghunathan, A.; Liang, P.; Ma, T. An explanation of in-context learning as implicit bayesian inference. *arXiv* **2021**, arXiv:2111.02080.
39. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388. [CrossRef]
40. Zhang, W.; Li, C.; Peng, G.; Chen, Y.; Zhang, Z. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mech. Syst. Signal Process.* **2018**, *100*, 439–453. [CrossRef]
41. Zhang, S.; Zhang, S.; Wang, B.; Habetler, T.G. Deep Learning Algorithms for Bearing Fault Diagnostics—A Comprehensive Review. *IEEE Access* **2020**, *8*, 29857–29881. [CrossRef]
42. Zhao, Z.; Ye, S.; Qi, L.; Ni, H.; Fei, S.; Tong, Z. A PI-Dual-STGCN Fault Diagnosis Model Based on the SHAP-LLM Joint Explanation Framework. *Sensors* **2026**, *26*, 723. [CrossRef]
43. Fan, J.; Zhao, H.; Wei, L.; Rao, J.; Li, G.; Yuan, J.; Xu, W.; Qi, Y. STELLA: Guiding Large Language Models for Time Series Forecasting with Semantic Abstractions. *arXiv* **2025**, arXiv:2512.04871. [CrossRef]

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.