Large Models for Machine Monitoring and Fault Diagnostics: Opportunities, Challenges and Future Direction

Xuefeng Chen*[,a], Yaguo Lei*[,a], Yanfu Li*[,b], Simon Parkinson*[,c], Xiang Li[a], Jinxin Liu[a], Fan Lu[a], Huan Wang[b], Zisheng Wang[b], Bin Yang[a], Shilong Ye[b], Zhibin Zhao[a]

[a]School of Mechanical Engineering, Xi'an Jiaotong University, China

[b]Department of Industrial Engineering, Tsinghua University, China

[c]Department of Computer Science, University of Huddersfield, UK

*Joint First Authors

Received May 27, 2025 | Accepted June 20, 2025 | Posted Online Month X, XXXX

Abstract : As a critical technology for industrial system reliability and safety, machine monitoring and fault diagnostics has advanced transformatively with Large Language Models (LLMs). This paper reviews LLM based monitoring and diagnostics methodologies, categorizing them into in-context learning, fine tuning, retrieval augmented generation, multimodal learning, and time series approaches, analyzing advances in diagnostics and decision support. It identifies bottlenecks like limited industrial data and edge deployment issues, proposing a three stage roadmap to highlight LLMs' potential in shaping adaptive, interpretable PHM frameworks.

Keywords: LLMs; context learning; multimodal learning; fault diagnostics

I. INTRODUCTION

Machine monitoring and fault diagnostics is vital for ensuring industrial machinery reliability. Large Models (LMs), particularly Large Language Models (LLMs), are transforming PHM by addressing challenges in condition monitoring and fault diagnostics through capabilities like in-context learning and multimodal reasoning. This paper reviews LM-based approaches (in-context learning, fine-tuning, retrieval-augmented generation, multimodal, time series) and outlines a three-stage roadmap for LM-enabled prognostics and health management (PHM): knowledge-enhanced, task-driven, and self-learning frameworks.

Section II, contributed by Xuefeng Chen, overviews LLM methods for

intelligent maintenance, covering advancements in in-context learning, fine-tuning, and multimodal fusion. Section III, by Yaguo Lei, explores opportunities/challenges in LLM-driven machine fault diagnosis, highlighting progress in foundation models and interpretability. Section IV, authored by Simon Parkinson from the University of Huddersfield, discusses LLMs in engineering monitoring, emphasizing data interpretation and human-in-the-loop applications. Section V, by Yanfu Li, outlines future directions, including industrial knowledge integration, multimodal modeling, and edge optimization for LLM evolution in PHM.

II. Large Models in Prognostics and Health Management

A. OVERVIEW

In recent years, LMs, exemplified by LLMs, have achieved remarkable strides, unveiling colossal potential towards the actualization of artificial general intelligence (AGI). LMs refer to models with massive parameters across various data types, while LLMs are a specific type of the LMs trained exclusively on text to understand and generate natural language, which belong to a subset of LMs. LLMs employ large generative modeling on vast textual corpora. When both dataset scale and model size reach a certain inflection point, they can exhibit astounding generative prowess. The well-known ChatGPT being a paragon, employs techniques like instruction alignment, reinforcement learning, fine-tuning, and thought chain for training and adjustment, equipping LMs with robust generalization, inference, decision-making, and generative capabilities. It is also promising to develop LM-based prognostics and health management (PHM) applications to leverage PHM with the powerful performance, like ChatGPT. Meanwhile, there is a pressing need to address the bottlenecks and practical requirements pertaining to PHM technology. By synergizing applications of LMs, collaborating between ordinary models and LMs, and emphasizing specialized field LMs development, there's an opportunity to refine current PHM operational frameworks, enhance PHM algorithm competencies, and bolster downstream PHM tasks.

This section provides an overview of LM-based methods applied to intelligent maintenance, as illustrated in Fig. 1. It summarizes the main paradigms of LM-based methods from five perspectives: in-context learning (ICL)-based, fine-tuning-based, retrieval-augmented generation (RAG)-based, multimodal-based and time-series-based LMs. Furthermore, details of the paradigms are introduced and related works are reviewed. Additionally, the challenges and future directions related to intelligent maintenance of complex systems are discussed.

Fig. 1. LLMs in PHM

B. ADVANCES OF LARGE MODELS IN PHM

1. In-Context Learning-Based LMs

Most LMs utilized in ICL are LLMs since ICL is a prominent capability of LLMs wherein the model can perform novel tasks by conditioning on a sequence of input-output examples provided in the prompt, without any parameter updates or task-specific fine-tuning. In this setting, the model effectively leverages the contextual information to infer the underlying task and generate appropriate outputs for new inputs. Unlike traditional learning paradigms that require explicit training on labeled datasets, ICL enables zero-shot, one-shot, or few-shot generalization by treating the provided examples as implicit supervision. This approach highlights the model's ability to internalize broad patterns during pretraining and adapt to specific tasks dynamically at the testing time.

In PHM, ICL means enabling an LLM to answer specialized questions, like equipment faults or maintenance protocols, by including relevant technical documents and expert knowledge in the prompt without retraining the model. Specifically, a task-specific query, such as fault diagnosis or maintenance recommendation, is formulated by the user at first. Subsequently, relevant background information, such as technical specifications, operational guidelines, or historical maintenance records, is selected and organized to provide the necessary context for the task. Next, this contextual information is combined with the query into a well-structured prompt that reflects the task's domain-specific requirements. Finally,

LLMs process the prompt holistically and generates a response that reflects both the contextual cues and its pretrained language understanding capabilities, giving fault diagnosis results, maintenance recommendation, etc.

For LLMs like GPT, the generation is autoregressive. Given a prompt $x = (x_1, x_2, ..., x_T)$, whose number is T, the output tokens $y = (y_1, y_2, ..., y_N)$, whose number is N, are generated via:

$$P(y|\ x) = \prod_{t=1}^{N} P(y_t|\ x, y_{<t}) \tag{1}$$

This reflects how the LLM generates output one token at a time, conditioned on the input and previous outputs.

From an end-to-end view, the whole ICL pipeline can be described as:

$$\mathrm{Output} = \mathrm{LLM}(\mathrm{Prompt}(\mathrm{Context}, \mathrm{Query})) \tag{2}$$

where the prompt is the function of the combination of technical context and users' query.

The research progress of ICL-based LMs in the field of PHM. In recent years, numerous scholars have actively explored the application of ICL-based LMs in the field of PHM, investigating their potential for tasks such as intelligent fault diagnosis, predictive maintenance, and knowledge-based decision support across various industrial domains. For instance, Wang et al. [1]proposed a local knowledge base empowered LLM (LKB-E-LLM) that incorporates PHM-specific knowledge through prompt-based retrieval to enhance LLMs. Specifically, it combined text embedding, vector similarity search, and prompt engineering to integrate external technical content into the LLM's input without requiring model retraining. It significantly improved the model's ability to generate accurate, relevant, and professional responses in PHM scenarios. Lukens et al. [2]developed an LLM-based copilot system for maintenance recommendations. A framework using specialized LLM agents, recommender and evaluator, that generates and assesses diagnostic steps in response to sensor alerts, was designed. The system incorporated RAG to enhance contextual relevance and leveraged prompt engineering to structure inputs for reliable, task-specific reasoning. Dave et al. [3]developed a system for explainable fault diagnosis by integrating a physics-based diagnostics tool, named PRO-AID[4], with an LLM. Prompt engineering was utilized to carefully manage and provide specific contextual information to the LLM. This involved feeding the LLM with knowledge about the plant, real-time data from PRO-AID, and the structure of the diagnostic system through a "Symbolic Engine". This contextualization aimed to align the LLM's responses with accurate, physics-based information and diagnostic results, thereby constraining the LLM to prevent hallucinations and enabling it to provide understandable explanations for fault diagnosis. Liao et al. [5]proposed a novel approach for constructing a fine-grained knowledge graph for robotic fault diagnosis. This work leveraged LLMs with a

prompt-engineered industrial nested label classification template to enhance nested named entity recognition. The attention-map aware keyword selection for industrial nested language model and confidence filtering mechanism further improved data augmentation and entity extraction accuracy. Huang et al. [6]introduced a framework for root cause analysis in industrial asset health management with pretrained LLMs. The LLM-enhanced deep root cause analysis (LDRA) method designed prompts to direct LLMs in ranking potential root causes identified by data-driven models, leveraging symptom signals and saliency maps. A multi-LLM debating strategy with self-exclusionary voting reduced biases, improving reliability. Ma et al. [7]proposed a fault diagnostic reasoning pipeline named FDRKG-LLM, integrating LLMs with knowledge graphs (KGs). The method enhanced human-machine collaboration by enabling natural language queries for fault diagnosis in mechanical equipment. Sophisticated prompt engineering facilitated named entity recognition, intent recognition, and subgraph correction, guiding LLMs to leverage KGs effectively, reducing hallucinations, and improving interpretability. Zhou et al. [8]proposed CausalKGPT, an industrial causal knowledge-enhanced LLM for analyzing quality defects in aerospace product manufacturing. By integrating a causal quality-related knowledge graph with a structure causal graph-based sum-product network, the model eliminated pseudo-associations.

2. Fine-Tuning-Based LMs

Fine-tuning in LMs refers to the process of adapting a pretrained model to a specific task or domain by updating its parameters using labeled, task-specific data. While LMs are typically pretrained on diverse textual corpora using self-supervised objectives (e.g., causal or masked language modeling), fine-tuning enables the model to specialize in narrower domains such as failure mode identification, maintenance recommendation generation, or anomaly report summarization.

This adaptation involves continued supervised training, where gradients from a task-specific loss function are used to update the model's parameters. Depending on computational resources and data availability, fine-tuning can be performed at different levels of granularity—from full-model updates to parameter-efficient techniques such as adapter layers, Low-Rank Adaptation (LoRA), and prefix-tuning. By incorporating domain-specific knowledge and aligning the model with operational goals and compliance requirements, fine-tuning enhances the utility of LMs in real-world PHM systems.

Let $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a labeled dataset, where x_i denotes structured or unstructured input data, e.g., sensor logs, maintenance records, or inspection notes, and y_i corresponds to the target output, e.g., fault types, remaining useful life (RUL), or maintenance action.

Given a pretrained model f_θ with parameters θ, fine-tuning aims to optimize

the parameters θ to minimize a task-specific loss function L :

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathrm{L}\left(f_\theta(x_i), y_i\right) \tag{3}$$

For resource-constrained industrial environments or scenarios where labeled data is limited, parameter-efficient fine-tuning approaches can be employed. In these settings, only a small set of task-specific parameters $\phi \subset \theta$ is updated while the core model θ remains frozen:

$$\phi^* = \arg\min_{\phi} \frac{1}{N} \sum_{i=1}^{N} \mathrm{L}\left(f_{\theta,\phi}(x_i), y_i\right) \tag{4}$$

where $f_{\theta,\phi}$ is a model where lightweight modules—such as adapter layers or low-rank projection matrices—are inserted into the frozen base model. This allows for efficient customization of LMs to PHM tasks without the computational overhead of full retraining, while still achieving domain-specific accuracy and compliance.

The research progress of fine-tuning-based LMs in the field of PHM. Tao et al. [9]proposed to textualize vibration data by extracting 24 time and frequency domain features and converting them into natural language descriptions paired with fault modes as question-and-answer inputs for LLMs. The LoRA and quantized LoRA (QLoRA) were applied to fine-tune the LLMs. Zheng et al. [10]conducted an empirical study on fine-tuning LLMs for fault diagnosis of complex systems. Sensor data were converted into prompt datasets and using LoRA for fine-tuning. Zhang et al. [11]proposed a labeled-data-supervised fine-tuning method for LLMs in heating, ventilation, and air conditioning (HVAC) fault diagnosis. It used self-correction to generate datasets and SMOTE for augmentation. The fine-tuned GPT-3.5 with LoRA outperformed GPT-4, with strong generalization and dimension-agnostic ability. Lai et al. [12]proposed a bearing fault diagnosis foundation model, BearingFM. It adopted a cloud-edge-end collaborative framework, used fault-mechanism data augmentation, and designed a contrastive learning model for efficient fine-tuning with small labeled datasets.

3. Retrieval-Augmented Generation-Based LMs

LMs utilized in RAG-based methods are all LLMs since RAG is designed for language-based tasks. RAG is a hybrid framework that enhances LLMs by incorporating external, domain-specific knowledge at the testing time. In PHM, where accurate diagnostics and decision support depend on complex, evolving information sources, such as sensor logs, fault histories, and technical manuals, RAG offers a powerful mechanism for grounding responses in relevant operational data. Unlike standard LLMs, which rely solely on static pretraining, RAG dynamically retrieves relevant documents or records based on the input query, enabling the model to generate outputs that are more context-aware, accurate, and timely.

A typical RAG pipeline for PHM includes a retriever module, using dense vector search or keyword-based techniques, to identify the top-k most relevant maintenance reports, inspection records, or operating manuals. These documents are

concatenated with the original input and passed to a generator model, often a pretrained LLM fine-tuned for industrial terminology. This setup allows the system to deliver high-quality, explainable outputs for tasks such as failure diagnosis, maintenance recommendation, and anomaly interpretation, while mitigating issues like hallucination and outdated knowledge without requiring model retraining.

The research progress of RAG-based LLMs in the field of PHM. Xia et al. [13]proposed FCLLM-DT for bearing fault diagnosis, integrating digital twin, RAG-assisted LLMs, and federated continual learning. For RAG, it constructed a knowledge base from historical data, used a sliding window mechanism in prompts, and updated the queue to generate virtual data, enhancing LLM's accuracy and mitigating hallucination issues. Tao et al. [14]proposed LLM-R, a framework for domain-adaptive maintenance scheme generation, integrating LLMs with hierarchical task-based agents and instruction-level RAG. RAG enhanced the framework by vectorizing maintenance task keywords, retrieving relevant data from a vast knowledge base using a BERT encoder and maximum inner product search, and generating accurate maintenance schemes. It also combined LoRA-KR loss. LLM-R mitigates knowledge conflicts and improves adaptability for complex and small-sample maintenance tasks in diverse domains like aviation and manufacturing. Jurim et al.[15]proposed ChatCNC, a conversational machine monitoring framework for human-centric smart manufacturing. It integrated LLMs and Real-time RAG. RAG dynamically retrieved real-time CNC machine data from IIoT databases to enable context-aware responses, reducing reliance on technical support and enhancing human-data interaction flexibility. Liu et al. [16]proposed an intelligent CNC fault diagnosis system for identifying fault causes, providing repair solutions, and supporting real-time monitoring and maintenance. It integrated LLMs and domain KGs. A multi-source KG is constructed for structured representation. A RAG framework based on KG supports multi-turn interactive diagnosis with real-time data. A dynamic learning mechanism enables knowledge updates.

4. Multimodal-Based LMs

LMs integrate multi-source data, including text, images, audio, and vibration signals, into a unified architecture capable of processing, reasoning, and generating contextually coherent outputs across heterogeneous inputs. Extending beyond traditional text-based LMs, these models employ modality-specific encoders, such as vision encoders for images and specialized modules for industrial signals like vibration or temperature, to transform non-text data into compatible representations. Techniques like cross-attention, joint embeddings, or unified tokenization enable effective fusion of these modalities. In PHM, multimodal LMs can process equipment monitoring data, enhancing fault detection, predictive maintenance, and decision-making. These capabilities advance the development of robust, general-purpose AI systems for complex industrial PHM tasks.

The research progress of multimodal LMs in the field of PHM. Multimodal LMs

for PHM can be broadly categorized into two approaches: alignment and embedding. The alignment approach encodes multi-source information, leveraging methods such as CLIP [17] to align features across modalities at the representation level, ensuring coherent integration of heterogeneous data such as vibration and images. For instance, Li et al. [18]proposed the VSLLaVA pipeline, which integrates vibration analysis expert knowledge into a multimodal LM with signal-question-answer triplets. This pipeline employs LoRA to fine-tune the linear layers of CLIP and the LLM, enhancing performance in signal parameter identification and fault diagnosis for industrial vibration analysis. Alsaif et al. [19]developed a multimodal LM to integrate various modality information, including text, images, audio, vibration signals, and video, employing a CLIP-like modality alignment operation. It maps multimodal features to a unified semantic space through input projectors, linear projectors, and cross-attention mechanisms, ensuring semantic consistency and efficient feature fusion. Chen et al. [20]constructed a large-scale fault diagnosis dataset including vibration time-frequency image-text label pairs and human instruction-ground truth pairs, and employed a multi-scale cross-modal image decoder to extract fine-grained fault semantics, enhancing the accuracy of fault diagnosis reports. Lin et al. [21]developed FD-LLM for fault diagnosis in aero-engines and bearings. The approach integrated multimodal data alignment and fuzzy semantic embedding to process engineering time-series data, addressing challenges like data readability and limited fault samples. Conversely, the embedding approach prioritized one primary source, encoding secondary sources like vibration signals into feature vectors that are integrated into the dominant representation. Jose et al. [22], [23] fine-tuned a LLM on domain-specific texts, embedded inspection notes using the fine-tuned LLM, and used these embeddings to weight other monitoring data, thereby improving the prediction accuracy of machine degradation levels. Peng et al. [24]converted the features of both query and fault-free vibration signals into word embedding and concatenated user instruction text embedding to generate natural language responses for anomaly detection, fault diagnosis, maintenance recommendation, and potential risk analysis tasks. Wang et al.[25] augmented the input signal with semantic information by concatenating embedding signals and prompts and fine-tuned a pretrained LLM for fault diagnosis. Furthermore, some approaches use LLMs as components to leverage linguistic semantics for time-series modeling. Liu et al. [26]employed an LLM as a supervisor for preliminary diagnosis, leveraging brain-inspired chain-of-thought reasoning. Subsequently, small models refined the initial results to achieve precise diagnoses. Zhang et al.[27] utilized an LLM to automatically select the best decomposition level and frequency band by analyzing historical fault data for optimized wavelet packet transform, reducing the subjectivity and uncertainty of manual parameter settings. Du et al. [28]proposed a method to incorporate time-frequency domain information into LLMs, enabling rapid learning of time-frequency data

characteristics for fine-tuning convolutional model. This approach significantly reduces manual operation time.

5. Time Series-Based LMs

Time series-based LMs usually utilize LLMs as encoders to directly process monitoring signals, which do not take pure text as input. By treating time series as sequences of tokens, time series LMs leverage transformer architectures to capture complex temporal patterns and long-range dependencies, which are critical for tasks such as fault diagnosis and RUL prediction. Different types of time series LMs exist based on their input representations and adaptation strategies. Token-based models segment sensor data into fixed-length tokens, while time-frequency models transform signals into spectral representations before tokenization. Prompt-based models introduce task-specific tokens to guide the learning process, and adapter-based models fine-tune lightweight modules while keeping most pretrained parameters fixed for efficient domain adaptation. Additionally, foundation models pretrained on large, diverse datasets using self-supervised objectives provide robust, generalizable representations, enabling few-shot or zero-shot learning across PHM tasks.

The general pipeline begins by normalizing raw time series x_t and segmenting it using a sliding window or patching function $p_i = f(x_{t:t+\Delta})$. Each segment is embedded into a vector:

$$e_i = \phi(p_i) + \varphi(i) \tag{5}$$

where $\phi(\cdot)$ is a learnable embedding function and $\varphi(i)$ adds positional information. The sequence $\{e_i\}$ is then processed by a transformer:

$$H = \text{Transformer}(\{e_i\}) \tag{6}$$

where attention weights model interactions across time. A task-specific head $y = g(H)$ outputs RUL estimates, fault classes, or anomaly scores, supporting maintenance strategies and system health assessment.

The research progress of time series-based LMs in the field of PHM. Wang et al. [29]proposed a time series LM for RUL prediction. By fine-tuning GPT2, it incorporated time series-specific processing: patching to segment sequences into fixed-length patches, positional encoding to inject temporal order, and linear probing for dimension alignment. Self-attention and feed-forward layers in GPT2 were frozen, while residual/normalization layers were updated during fine-tuning. Wang et al. [30]proposed RmGPT, a time series-based LM utilizing a token-based framework with Signal/Prompt/Time-Frequency Task/Fault Tokens. Self-supervised learning via next signal token prediction extracted features, and prompt learning adapts to tasks was integrated. A dual-stage attention transformer was utilized to process multi-channel signals. Pan et al. [31] proposed ParInfoGPT, an LM-based two-stage framework for rotating machine reliability assessment under partial information. It integrated a self-supervised reconstruction network with MI-based masking and a weakly supervised classification network

using a parallel side-adapter. GPT-2 served as the backbone, leveraging pre-trained linguistic capabilities for time-series feature learning. Qin et al. [32]proposed a large fault diagnosis model for rotating machinery based on a dense connection network with depthwise separable convolution (DCNDSC). It designed a dense connection block with depthwise separable convolution (DCDSCB) to capture complex features and proposed a diminutive network fine-tuning strategy to enhance adaptability to new data. Pan et al.[33]explored applying LMs to machinery fault diagnosis for time series analysis. It proposed LLaMA-HFT, a framework using LLaMA2 as the backbone. A hybrid fine-tuning strategy was adopted via freezing part of the bottom blocks while fine-tuning with LoRA on the top blocks. Tao et al. [34]proposed LM4RUL for bearing RUL prediction using pre-trained LMs for time series. It employed local scale perception representation to tokenize vibration data into time-frequency features and uses hybrid embedding learning with selective freezing/fine-tuning. A two-stage fine-tuning strategy adapts pre-trained knowledge to industrial scenarios, enabling long-term RUL prediction without heavy manual feature engineering. Eldele et al.[35] proposed UniFault, a fault diagnosis foundation model for time series. It tackled heterogeneous fault diagnosis data via a preprocessing pipeline with data normalization, sliding window transformation, channel-unification, and cross-domain temporal fusion. A Transformer-based backbone with contrastive self-supervised learning enabled few-shot adaptation, leveraging 9B+ pretraining data points. Chen and Liu[20] proposed a time series LM-based regression framework for RUL prediction. It utilized GPT-2 with self-attention to capture temporal and spatial correlations in multidimensional industrial signals. A unified model structure with the same sliding window and all sensors was adopted.

C. CHALLENGES AND OPPORTUNITIES

1. High-Quality and Large-Scale Industrial Data

The pretraining of LMs rely heavily on vast amounts of high-quality and large-scale data, such as LLMs requiring corpora, while time-series LMs demanding raw data. The scaling law for LMs indicates that model performance is closely tied to dataset size and the number of model parameters[36]. However, in the field of PHM, acquiring sufficient fault data remains a significant challenge. For example, in the fields like wind power and petrochemical intelligent fault diagnosis, the acquisition and organization of knowledge and corpus are relatively straightforward, benefitting from decades of research, accumulated case studies, and industry-standardized documentation. However, for emerging and specialized areas, like C919 aircraft, corpus collection presents challenges, further hampering the development of effective LMs. In addition, while public datasets from fault simulation experiments on mechanical systems are available, their data volume is often limited, and the homogeneity of the data hampers the training of models with strong generalization capabilities. Meanwhile, enterprises are reluctant to share

monitoring data and maintenance records publicly due to concerns over commercial confidentiality. Consequently, constructing a large-scale, high-quality industrial PHM database for pretraining LMs with robust generalizability poses a formidable obstacle.

2. Diversity and Complexity of Industrial Data

Industrial systems comprise diverse equipment and components, each operating under varying conditions and monitored by multiple sensors that collect data such as images, vibration, sound, current, and temperature. For instance, in bearing and gear monitoring, vibration signals are commonly used to assess health conditions, whereas pressure signals are typically collected for fuel control system monitoring. Beyond the inherent heterogeneity of data, the nonstationary nature of signals, such as vibration, complicates feature extraction and modeling. Furthermore, equipment is monitored over extended periods, making it critical to capture long-term dependencies in the data. Consequently, the multimodal, nonstationary, and temporally extended nature of industrial data presents significant challenges for effective modeling in LMs. Specifically, multimodal alignment poses challenges due to the complexity of aligning high-dimensional, heterogeneous data and the absence of a shared semantic space across different modalities, which hinders the unified learning in the representation space.

3. Resource and Time Constraints in System Deployment

LMs, characterized by their vast parameter counts and complex architectures, impose significant computational demands for both training and inference. In industrial settings, deploying these models on edge devices presents substantial challenges. The computational load of LMs typically necessitates high-performance infrastructure, such as powerful GPUs. However, in most industrial scenarios, constraints on power, space, and cost make it impractical to equip edge devices with such advanced hardware. Furthermore, industrial applications, particularly in PHM, require rapid response times to enable early fault detection and timely interventions. The inherently slow inference speed of LMs, driven by their computational complexity, struggles to meet the real-time requirements of industrial systems. Therefore, achieving a balance between model performance and inference speed while deploying efficient LMs on resource-constrained edge devices remains a significant obstacle.

4. Model Generalization in Open Environments

Engineering knowledge exhibits strong domain dependency and contextual coupling. Diverse industrial systems exhibit unique data distributions and physical principles, creating significant gaps between them. Even within a single system, varying operating conditions and failure modes tightly coupled with operational contexts hinder models from adapting to new environments without retraining or fine-tuning. For instance, a model trained under stable conditions may fail under different loads or environmental factors. Furthermore, the absence of a unified framework for encoding engineering knowledge limits the

integration of domain-specific insights into LMs, restricting their ability to generalize learned patterns across domains. Thus, designing LMs with robust generalization capabilities in open, dynamic engineering environments poses a critical challenge.

5. Limited Adaptability of General-Purpose LMs for PHM

General-purpose LMs, such as ChatGPT and DeepSeek, are trained on vast datasets of text and images scraped from the internet, offering strong generalizability. However, their applicability to PHM tasks is limited due to a lack of domain-specific knowledge and monitoring data for industrial equipment, components, and parts. Signals such as images, vibration, temperature, and pressure are critical in PHM but are underrepresented in general datasets. Moreover, LMs are primarily designed for natural language processing, lacking mechanisms for capturing long-term dependencies in time-series data or effectively fusing heterogeneous multisource signals. Consequently, the incomplete adaptability of general-purpose LMs to industrial PHM tasks poses a significant challenge.

D. FUTURE RESEARCH DIRECTIONS

1. Data Generation for Industrial Systems

Training LMs relies on extensive high-quality datasets, which are often scarce in industrial contexts. To overcome this, advanced data generation techniques, such as simulation modeling, digital twin systems, and diffusion models, can synthesize high-fidelity data samples. Simulation modeling can replicate equipment behavior under diverse conditions, while digital twins provide real-time, system-specific data through virtual representations. Diffusion models can generate realistic, diverse signals like vibration or temperature. These approaches reduce dependency on real-world data, enhancing LMs usability and robustness in complex industrial scenarios, particularly for PHM tasks.

2. Fusion and Representation of Multi-Source Data

Industrial systems integrate diverse equipment, sensors, and operating conditions, producing data characterized by strong heterogeneity, non-stationarity, and temporal dependencies. Developing effective feature fusion and representation methods for multi-source heterogeneous data is thus critical. For long-term time-series data such as vibration, temperature, and pressure, models like Transformer can capture long-range dependencies. However, misaligned features in the fusion process may lead to information loss or feature redundancy, ultimately degrading model generalization. To mitigate this, hierarchical alignment strategies can be adopted, aligning data at multiple levels—such as sensor, feature, and semantic space levels—to ensure coherent fusion. In addition, LMs or intelligent agents can be leveraged to manage multi-agent systems, such as swarm robotics, thereby promoting unified and adaptable data representations that effectively handle complex industrial data and enhance PHM performance.

3. Lightweight and Edge-Optimized Industrial LMs

To deploy LMs in resource-constrained industrial edge environments, research into model compression techniques, such as knowledge distillation, pruning, and quantization, is vital. These methods reduce model size and computational demands, enabling efficient operation on edge devices. Additionally, exploring hybrid approaches that distribute computations between edge and cloud systems can balance real-time requirements with model accuracy. By optimizing inference speed while preserving predictive performance, these techniques enhance the applicability of LMs for real-time PHM tasks in industrial settings, addressing the challenges of limited computational resources and stringent latency demands.

4. Knowledge Graphs and Reasoning for LMs in PHM

PHM knowledge for engineering has strong domain dependency and contextual coupling, limiting model generalization in open environments. Building knowledge graphs from unstructured data, such as expert experience, equipment manuals, and maintenance records, provides domain-specific data for LMs. By integrating reinforcement learning-based reasoning, LMs can transfer knowledge across industrial systems, mitigating data distribution gaps. This approach reduces issues from varying operating conditions, enhances generalization in dynamic PHM tasks, and improves interpretability through transparent reasoning, offering PHM engineers reliable decision-making support. In conclusion, aiming at the prevalent transferability challenges of past diagnostic methods, future research endeavors to design a LM that requires no transfer, enabling it to operate efficiently on at least one category of equipment.

5. Development of PHM-Specific Foundation Models

The limited adaptability of general-purpose LMs to PHM highlights the need for specialized foundation models designed from the ground up. These models should integrate domain-specific knowledge and industrial monitoring data, such as vibration, temperature, pressure, and images. Tailored architectures, like temporal Transformers or graph-based networks, can capture long-term dependencies in time-series data and fuse multi-source signals. Moreover, a key direction of exploration lies in breaking away from the traditional paradigm of merely outputting the diagnosis result. Instead, future efforts should aim to expand these models into comprehensive intelligent maintenance, facilitating end-to-end applications across maintenance planning, decision-making processes, and even robotic detection integration. This whole approach will enable more proactive and autonomous industrial maintenance. Meanwhile, the investigation into the self-learning and self-reasoning capabilities of these PHM foundation models, especially in handling complex and rare industrial cases or anomalies. By empowering models to continuously adapt and reason through challenging scenarios, the industry can achieve higher accuracy in fault detection, enhanced predictive maintenance

capabilities, and ultimately, more robust and reliable industrial systems.

III. Opportunities and Challenges in Large Model-Enabled Machine Fault Diagnosis

A. A BRIEF INTRODUCTION

In recent years, LLMs such as ChatGPT, Qwen, and DeepSeek have emerged as transformative technologies in the field of artificial intelligence[37]. These models, often containing billions or even trillions of parameters, are usually pre-trained on massive data and fine-tuned for a wide range of downstream tasks. With abilities such as in-context learning, instruction following, and chain-of-thought prompting, LLMs have demonstrated exceptional generalization capabilities across diverse natural language processing (NLP) tasks, including but not limited to dialogue systems, text generation, and expert-knowledge-based reasoning.

As the field progresses, LLMs have evolved from language-only models to multimodal large language models (MLLMs), capable of understanding and integrating information across multiple data modalities. Pioneering frameworks such as CLIP[38], BLIP-2[39], and LLaVA[40] have demonstrated the feasibility of aligning visual and textual inputs within a unified semantic space. These advances enable practical applications in cross-modal learning and are starting to reshape vertical industrial fields. In particular, LLMs and MLLMs show immense potential in industrial intelligence, enabling new capabilities for machine fault diagnosis[41-43].

Fault diagnosis of industrial machinery plays a vital role in maintaining the safety, reliability, and operational efficiency of modern industrial systems[44]. It involves detecting, identifying, and classifying failure modes in equipment by analyzing data collected from sensors, such as vibration, temperature, acoustic emission, etc. The general goal is to enable early detection of potential failures to support condition-based maintenance and reduce machine downtime. Over the years, a wide range of fault diagnosis methods have been developed, including signal-processing-based methods[45], model-based methods[46], and data-driven methods[47,48]. In particular, the rapid advancement of deep learning has enabled end-to-end fault diagnosis through automatic feature learning and classification, using models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformer.

While these neural network-based methods have achieved encouraging results in various applications, they also face several limitations. For instance, many models are built with fixed architectures that are specifically tailored to certain operating conditions. As a result, their adaptability may be limited in scenarios such as cross-condition or small-sample learning situations[49]. In addition, a common concern lies in their limited interpretability. These models often function as “black boxes”, offering relatively limited insight into the reasons behind their predictions[50]. Therefore, it is difficult for engineers to

fully assess the reliability of diagnostic results, particularly in cases where interpretability is important for practical implementation.

To address the aforementioned limitations of deep neural network-based methods, researchers have begun exploring how the capabilities and core principles of LLMs can be leveraged to enhance machine fault diagnosis[51]. Although large model-enabled machine fault diagnosis is still in its early stages, current studies have demonstrated its potential to overcome the limitations by taking advantage of the merits of generalization[51,52,9,53], interpretability[10,54,55], and interactivity[56,57]. The current research efforts can be broadly categorized into three directions:

1.Enhancing Generalization via Pretrained Foundation Models

Inspired by the generalization capability of LLMs, some studies have begun to explore the construction of foundation models for fault diagnosis. These approaches adopt a pretraining-then-finetuning paradigm. In such framework, models are first pretrained on large-scale and heterogeneous fault-related datasets and then adapted to specific diagnostic tasks through lightweight finetuning. The goal is to learn generalized feature representations that are resilient to variations in operating conditions, machine types, and data sources. Such models hold promise for enabling cross-machine, cross-condition, and cross-dataset fault diagnosis with minimal additional training. Qin et al. proposed a pretrained large fault diagnosis model based on a densely connected network with depthwise separable convolutions (DCNDSC), enabling accurate and generalized diagnosis across various rotating machines and fault types through effective feature extraction and fine-tuning[52]. Tao et al. proposed an LLM-based bearing fault diagnosis framework that textualizes vibration features and uses fine-tuning to improve cross-condition, small-sample, and cross-dataset generalization[9]. Lei et al. proposed a Transformer-based foundation model for intelligent maintenance that unifies multi-modal data and supports condition monitoring, fault diagnosis, and RUL prediction across components like bearings and gears, with strong generalization and adaptability[53].

2.Improving Interpretability with Knowledge Integration and Reasoning

Another aspect of research focuses on enhancing the interpretability of diagnostic results. By incorporating domain knowledge, signal processing techniques, and structured knowledge representations, these models leverage LLMs' chain-of-thought reasoning abilities. Beyond the diagnostic results, they strive to provide transparent reasoning paths that make the diagnostic logic understandable to human users. Liu et al. proposed a knowledge-enhanced model that embeds aviation assembly knowledge graphs into large language models via prefix-tuning, achieving 98.5% accuracy in industrial fault localization and troubleshooting[54]. Men et al. introduced the To-FD-EKG framework, integrating large language models with fault diagnosis event knowledge graphs to enable traceable

reasoning through structured knowledge and digital twins[55].

3. Enabling Interactive and Multimodal Fault Diagnosis

With the significant progress in MLLMs, researchers are also exploring interactive, and end-to-end diagnostic systems. The corresponding models are designed to process and align multiple data modalities, including vibration signals, textual descriptions, and maintenance logs. Human-in-the-loop diagnostics are supported by these systems, enabling engineers to interact with the model using natural language and to receive contextualized, and multimodal explanations. Lin et al. proposed a multimodal large language model approach with modal alignment, fuzzy semantic embedding, and learnable prompts to improve accuracy in complex equipment fault diagnosis of time-series data[21]. Chen et al. introduced FaultGPT, i.e. a vision-language model that generates fault diagnosis reports from raw vibration signals via instruction tuning and cross-modal decoding[56].

In summary, the aforementioned exploratory directions illustrate how LLM-based technologies could provide a new foundation for more adaptable, explainable, and user-centric machine fault diagnosis systems in complex industrial environments.

B. OPPORTUNITIES AND CHALLENGES

There have been growing interests in applying large models to machine fault diagnosis, offering promising potential to revolutionize traditional diagnostic paradigms. However, this research direction is still in its early stage, with only limited exploratory work reported. Most existing studies are conceptual or conducted under controlled laboratory conditions, and they cannot fully reflect the complexity and variability of real-world industrial environments[57]. As a result, large model-enabled machine fault diagnosis is still far from being practically deployed at scale. To bridge the gap between theoretical exploration and industrial application, several critical challenges should be addressed in future research.

1. Lack of High-Quality Textual Corpora in Fault Diagnosis

A fundamental bottleneck in developing large model-based solutions for machine fault diagnosis is the limited availability of high-quality and domain-specific corpora[11]. Unlike general-purpose natural language processing tasks, which benefit from massive, diverse, and richly annotated datasets, the fault diagnosis domain faces significant data limitations. It lacks large-scale, standardized textual and multimodal resources that capture expert knowledge, diagnostic procedures, signal interpretations, and causal relationships. Most existing datasets are small, task-specific, and often lack detailed annotations or contextual metadata, making them insufficient for training large models that require extensive input. Furthermore, much of the critical knowledge in this field exists in the minds of domain experts or is embedded in unstructured formats such as maintenance reports, logbooks, and research papers. Such knowledge is often difficult to access and

formalize for use in model training and inference.

2. Complex Fault Mechanisms and Challenges in Reasoning

Another key challenge in applying large models to machine fault diagnosis lies in the complexity of fault mechanisms[58]. Unlike classification tasks in other domains, where distinct features can be directly mapped to specific categories, machine fault diagnosis presents greater challenges. Faults such as inner race, outer race, or rolling element defects in bearings often do not exhibit a single and clear indicator. Instead, their signatures are typically subtle, overlapping, and highly context-dependent. These characteristics make reasoning in fault diagnosis particularly challenging, as it often requires inferring latent fault types from ambiguous, indirect, and context-sensitive evidence.

3. Difficulty in Applying Existing Methods from Other Domains to Fault Diagnosis

While LLMs and MLLMs have achieved remarkable success in domains such as NLP and computer vision, directly applying these models to machine fault diagnosis remains a significant challenge. Existing approaches like CLIP rely on aligning well-structured and semantically intuitive modalities. For example, images paired with textual descriptions often exhibit a clear and direct correspondence between visual and linguistic representations. However, this paradigm cannot easily transfer to the fault diagnosis domain.

Vibration or other time-series signals carry complex and implicit information that is not readily interpretable. Extracting meaningful insights from such data typically requires extensive domain knowledge and comprehensive understanding of the signal's temporal and spectral characteristics. Unlike images which offer intuitive patterns, the semantics of signals are abstract and context-dependent. As a result, simple signal-text alignment strategies fall short in capturing the underlying diagnostic knowledge.

Moreover, many current MLLM fine-tuning techniques rely heavily on instruction tuning, which often prioritizes similarity over actual comprehension. In such scenarios, the model cannot well learn how to analyze the features of signals. Instead, it tends to generate responses based on superficial similarities to previously seen examples, leading to inaccurate fault diagnosis. This highlights the demands for domain adapted strategies that can incorporate signal-specific representations and expert knowledge into the learning process.

IV. Large Language Models for Monitoring and Diagnostics: Future Opportunities

1 Overview

Monitoring in an engineering context includes a wide range of activities to ensure the correct operation of machinery. This consists of activities ranging from diagnostics, fault detection, and condition monitoring. Diagnostics involve

investigating a known problem to determine the cause, fault detection is concerned with recognising the presence of a problem, and condition monitoring is the continuous assessment of the system to predict potential performance, informing maintenance strategy to mitigate and predict problems [59]. These practices are fundamental in Engineering disciplines where system reliability is critical. For example, machinery in energy generation needs careful monitoring to prevent failure.

Traditionally, monitoring has relied on structured, and mostly numeric data, from sensors and control systems, analysed using data-driven analytical approaches such as machine learning and statistical methods. These approaches have proven to be very useful in identifying and predicting subtle changes in device behaviour. Engineers performing monitoring and diagnostic tasks will be familiar with approaches such as neural networks, fuzzy logic, support vector machines, etc [60]. Each approach has characteristics that make it more appropriate for specific monitoring tasks, depending on data types and analysis aims. While these approaches have proven effective in many scenarios, they often require structured data, domain-specific feature engineering, and extensive manual parameter tuning to yield the best results. Furthermore, these methods often struggle with unstructured data which contain rich contextual originating from maintenance logs, operator notes, and technical manuals, for example.

Recent advances in natural language processing, most notably in LLMs, have presented techniques capable of understanding and reasoning with data sources, exhibiting human-like output. Models such as GPT-3 and LLaMA are trained on vast corpora of text and exhibit strong capabilities in language processing tasks such as comprehension, summarisation, translation, and reasoning. Their ability to process unstructured textual data opens new avenues for enhancing monitoring systems, particularly in interpreting human-generated content and integrating it with sensor-based diagnostics. This potential is being witnessed in many other domains where monitoring is an essential task, such as security [61] and safety compliance[62]. These recent works demonstrate the potential for LLMs in monitoring. In addition to using traditional condition monitoring on structured data, then can introduce processing of unstructured data, enabling more comprehensive analysis that goes beyond identifying something of interest to explaining its significance and even how to mitigate identified challenges[63]. Furthermore, through using LLMs, engineers can gain deeper insight from historical records and support decision-making. This is especially valuable in complex systems where human expertise and machine data must be synthesised for effective diagnostics and maintenance planning.

2 Advance of Large Language Models

LLMs represent a paradigm shift in artificial intelligence. These models are built on transformer architectures and trained on large corpora of text data, enabling them to learn complex language patterns and semantic meaning. They work by having

large amounts of adjustable parameters (weights and biases) that can be changed during training. For example, GPT-3 contains 175 billion parameters. This large size enables them to be able to learn and store a large amount of relationship knowledge.

The key innovation of LLMs lies in their ability to generalise across tasks and domains using different learning strategies. For example, 'few-shot' learning is where the model is trained on a small number of labelled instances to guide its prediction, whereas "zero-shot" learning requires the model to perform tasks without any labelled examples, relying only on pre-trained knowledge. This makes them particularly attractive for engineering applications where labelled data is seldom available or where the monitoring context evolves. Unlike traditional machine learning models that require retraining for each new task, LLMs can adapt to new inputs with minimal additional data.

Emerging research has started to explore the application of LLMs in domains related to engineering monitoring. In predictive maintenance, LLMs have been used to analyse maintenance logs, technician notes, and service records to identify patterns that precede equipment failure[64]. In anomaly detection, LLMs can process textual alerts, error codes, and operator feedback to flag unusual behaviour based on semantic similarity[65]. In industrial automation, LLMs assist in interpreting procedural documents, generating troubleshooting steps, and even translating between technical languages and natural language instructions[82].

Despite these promising developments, the application of LLMs in core engineering monitoring tasks is still underexplored. Most existing studies focus on natural language tasks or general-purpose analytics, with limited attention to the integration of LLMs into real-time monitoring systems or their interaction with sensor-based data streams[19]. Furthermore, the potential of LLMs to bridge structured and unstructured data, such as combining sensor readings with maintenance narratives could provide useful functionality and warrants further investigation. In this work, we examine how LLMs can be used in monitoring and what opportunities and challenges exist in future work.

3 Use cases of LLMs in Monitoring

The use of LLMs beyond traditional AI methods presents new functionality to improve how data is processed, analysed, and interpreted. In this section, five key use cases are presented based on early emerging work.

Data Interpretation and Fusion: Engineering systems generate large volumes of heterogeneous data, including structured sensor outputs and unstructured sources such as maintenance logs, technician notes, and incident reports. Traditional monitoring systems are often limited to focusing only on the former as adopted algorithms are limited to a single data type. LLMs, however, have demonstrated good capabilities for processing unstructured text and can extract actionable insights from these sources [79].

For instance, an LLM can analyse a corpus of maintenance logs to identify recurring fault descriptions, correlate them with specific components or environmental conditions, and flag emerging issues. In addition to interpretation, LLMs can perform data parsing and fusion tasks[81]. This involves integrating information from multiple modalities (e.g., textual logs, sensor metadata, and operational parameters) to construct a holistic view of a system's condition. This integration provides additional context for monitoring tasks.

Anomaly Detection and Fault Diagnosis: LLMs can enhance traditional anomaly detection by identifying patterns in textual data that is related to faults detected in numeric data analysis [65]. For example, subtle shifts in the language used in operator notes, such as negative sentiment regarding the machine's operating condition may indicate deteriorating conditions before analytical tools detect the fault. In fault diagnosis, LLMs can serve as intelligent assistants that retrieve and analyse relevant historical cases. When presented with an error message or fault code, an LLM can search through records to find similar instances, summarise the root causes, and suggest corrective actions. This capability is particularly valuable in complex systems where fault signatures are complex or evolve, with the 'utilisation of unlabelled data mentioned as an open challenge[66].

Human-in-the-Loop Monitoring: Monitoring is not solely a technical process; it often involves human judgment, especially in high-stakes or uncertain scenarios. LLMs can enhance human-in-the-loop monitoring by acting as collaborative partners that summarise diagnostics, highlight anomalies, and propose remedial action. For example, an LLM can generate concise summaries of daily maintenance activities, flag unresolved issues, and recommend follow-up actions. Research undertaken in human-in-the-loop for manufacturing with collaborative robots (widely named cobots) for manufacturing and assembly has demonstrated the potential[67]. Furthering research in human-in-the-loop for monitoring tasks can help prevent unresolved issues from being missed. There is a large body of applicable research in human-in-the-loop with machine learning[68] and early work using LLMs [69]. This has the potential to reduce cognitive load and accelerate decision-making.

Knowledge Extraction and Reasoning: Engineering organisations accumulate vast repositories of knowledge in the form of manuals, service bulletins, and historical maintenance records. Much of this knowledge remains separate from widely implemented monitoring approaches due to its unstructured nature. However, LLMs have proven capabilities in extracting key insights, identifying trends, and surfacing best practices. Beyond extraction, LLMs can perform reasoning tasks, such as inferring causal relationships or generating preventive maintenance strategies. For instance, by analysing a decade of maintenance logs, an LLM might infer that a specific fault tends to occur after a certain sequence of events or under specific environmental conditions[70]. This reasoning capability supports proactive maintenance planning and continuous improvement. Monitoring tasks in engineering can leverage best practices from

security monitoring processes, where appropriate mitigation action will be undertaken once an event has been identified. However, as with the cyber security discipline, the communication of capabilities [71] and their standardisation[72]would emerge as future challenges.

Multimodal Monitoring: Modern monitoring systems increasingly rely on multimodal data sources, including data types coming from visual inspections, acoustic signals, time-series sensor data, and textual reports. While LLMs are inherently text-based, they can be integrated into multimodal frameworks to interpret and contextualise non-textual data[72]. For example, an LLM can be paired with a computer vision model that detects surface cracks in equipment. For example, using LLMs and computer vision to detect issues with material extrusion[73]. The LLM can then correlate these findings with recent maintenance logs to assess severity, suggest causes, and recommend actions. Similarly, audio anomalies detected by signal processing models can be contextualised using LLMs that analyse technician feedback or historical fault narratives. This multimodal synergy enhances the robustness and interpretability of monitoring systems, especially in complex environments where no single data source provides a complete picture. However, there is also the significant challenge that LLMs and other forms of generative AI can produce realistic replica datasets, which means that monitoring systems need to be robust to adversarial attacks [74].

4 Challenges and Future Work

The integration of LLMs into engineering monitoring systems presents beneficial opportunities, but there are a range of challenges and limitations that must be addressed to ensure effective and responsible deployment. This section outlines the key challenges and proposes future research directions to overcome them. Table 1 provides a summary to complement the discussion.

One of the foremost concerns is data privacy and security. Engineering systems often generate sensitive operational data, maintenance records, and failure logs. When such data is used to train or interact with LLMs, there is a risk of unintended exposure or misuse [75]. Moreover, compliance with data protection regulations such as GDPR or industry-specific standards is critical for organisations, both to prevent future attacks and ensure adequate security compliance should an attack occur. Future work must focus on developing privacy-preserving techniques, such as federated learning or differential privacy, and establishing robust governance frameworks for LLM deployment in industrial contexts.

Another significant limitation is the absence of domain-specific training data. Most LLMs are trained on the general-purpose corpus, which lacks the technical depth required for focused monitoring applications. As a result, these models may misinterpret domain-specific terminology or fail to capture subtle fault patterns. Addressing this requires approaches using Retrieval Augmented Generation (RAG) where domain-specific information is

retrieved from documentation and forms part of the LLM query [61]. The curation of high-quality, domain-specific corpora, such as annotated maintenance logs, sensor narratives, and technical enables RAG approaches to retrieve relevant information and improve query response.

The interpretability and trustworthiness of LLM outputs also pose a challenge. Although LLMs can generate sensible and realistic responses, the underlying reasoning is often opaque making it difficult for the recipient to understand any reasoning or find evidence. In safety-critical environments, engineers must be able to understand and justify the basis of any recommendation or insight. Future research should prioritise the development of explainable AI techniques tailored to LLMs, enabling users to trace outputs back to source data or model logic.

Real-time performance and deployment constraints further complicate the use of LLMs in monitoring. Many industrial applications require low-latency responses and operate in environments with limited computational resources [76]. Current LLMs, especially those with billions of parameters, are computationally intensive and may not be suitable for edge deployment. Research into model compression, quantization, and efficient inference architectures is essential to enable real-time, on-device monitoring. In some work, authors focus on improving the efficiency of function calls from edge devices [77]. In other recent works, authors focus on embedding LLMs in FPGAs for signal processing tasks; however, in some sensitive applications, there is a need to host the LLM within the organisation's control to ensure that there is no unintended exposure.

Beyond these core challenges, several opportunities exist. One promising direction is the development of hybrid models that combine LLMs with physics-based or signal-processing models. While LLMs excel at interpreting unstructured data and language, physics-based models offer precise, mechanistic insights into system behaviour. Integrating these approaches can yield more robust and interpretable monitoring systems.

Another area of focus should be the creation of benchmarking and evaluation frameworks specific to LLMs in engineering monitoring. Current benchmarks are often geared toward general NLP tasks and do not reflect the unique demands of industrial diagnostics. Establishing standardised datasets, metrics, and evaluation protocols will facilitate meaningful comparisons and accelerate progress in the field [78]. However, there is a need to ensure that LLMs are trained and benchmarked on datasets where the correct permissions have been granted, otherwise, there is a potential for copyright infringement [80].

Finally, the integration of LLMs with edge computing and IoT infrastructure represents a significant challenge. By deploying optimised LLMs on edge devices, organisations can enable localised, real-time analysis without relying on cloud connectivity and minimise security and privacy concerns. This is particularly valuable in remote or bandwidth-constrained

environments, where latency and data ownership are key concerns.

Table 1: Summary of Challenges, Limitations and areas of Future work

Challenge/Area	**Limitation**	**Future Work**
Data privacy and security	Risk of exposing sensitive data; regulatory compliance requirements	Explore developing privacy-preserving techniques and compliance frameworks
Lack of domain-specific training data	General-purpose models lack technical depth	Collection of domain-specific training data for anonymisation and sharing within the research community
Interpretability and trust	Opaque reasoning behind LLM outputs limits user confidence	Develop and leverage explainable AI techniques for LLMs
Real-time performance and deployment constraints	High computational demands restrict edge deployment and low-latency use	Optimise LLMs for real-time and resource-constrained environments
Hybrid models between LLMs and traditional ML models	LLMs alone may lack a mechanistic understanding of system behaviour	Create test datasets and use case studies to serve as benchmarks with hybrid model implementation and evaluation
Benchmarking and evaluation frameworks	Lack of domain-specific benchmarks for LLM performance in monitoring	Establish standardised practices, datasets and evaluation protocols

Integration with edge computing and IoT	Limited support for on-device inference in industrial settings	Explore real-time, localised monitoring through edge deployment of LLMs either through efficient API calls or small local hosting

V. LLM Future Direction

Recent advances in large vision-language models (LVLMs) and LLMs have brought transformative opportunities to industrial visual monitoring and prognostics and health management (PHM). Wang et al. introduced DefectGLM, the first LVLM tailored for wafer defect detection, which significantly improved semantic understanding and domain-specific text generation through large-scale multimodal data and contrastive domain adaptation[83]. Building on this, they developed IVMMF, an intelligent monitoring and maintenance framework integrating local knowledge bases with vision-language models to enable end-to-end automation from image recognition to maintenance recommendation[84]. Li et al. systematically reviewed the application of foundational models like ChatGPT in PHM, proposing a roadmap for the AI 2.0 era that emphasizes transitions from single-task to multimodal systems and from offline modeling to real-time intelligence[85]. In digital twin systems, Sun et al. employed an LLM-driven multi-agent architecture to enhance perception of global temporal features, improving decision-making intelligence and traceability[86]. Liu et al. developed a fault diagnosis system combining an aerospace assembly knowledge graph with LLMs, using prefix tuning for efficient knowledge integration and inference, demonstrating strong performance in complex industrial scenarios[54]. Together, these studies highlight how large models are overcoming traditional bottlenecks in industrial AI, driving PHM systems toward greater efficiency, intelligence, and autonomy.

LLMs have demonstrated significant potential in the field of PHM, offering a novel approach to overcoming the limitations of traditional methods in generalization, interpretability, and verification. Leveraging their strengths in generalization, logical reasoning, and natural language generation, researchers have proposed a three-stage progressive paradigm for LLM-driven PHM. The first stage, *knowledge-enhanced PHM*, integrates LLMs with enterprise knowledge bases, expert rules, and historical maintenance data, substantially improving equipment state understanding and fault diagnosis accuracy. The second stage, *task-driven PHM*, highlights LLMs' capabilities in task planning, resource allocation, and decision support, enabling the generation of executable maintenance strategies from natural language inputs. The third stage, *self-learning PHM*, envisions systems capable of continual adaptation and evolution across new equipment, conditions, and tasks through LLM-based continuous learning and optimization. Collectively, LLMs are emerging as the core engine driving PHM systems from knowledge integration and intelligent decision-making toward autonomous evolution, laying the

foundation for the next generation of cognitively capable and self-governing intelligent assurance systems.

Despite the promising potential of LLMs in industrial prognostics and health management (PHM), their widespread adoption faces several critical challenges and technical bottlenecks. Most existing LLMs are trained on general-purpose corpora, lacking deep modeling capabilities for industrial scenarios, specialized terminology, and heterogeneous data sources. This limits their ability to accurately interpret high-dimensional, weakly supervised signals embedded in equipment states, particularly when dealing with unstructured or temporal data such as vibration signals, thermal images, and system logs. Moreover, industrial PHM tasks are often highly customized and context-specific, yet current LLMs struggle to tightly integrate with domain-specific knowledge graphs and expert rule systems, leading to limited interpretability and verifiability of their inferences in engineering practice—particularly problematic for safety-critical systems. Additionally, industrial data are typically private and highly distributed, posing challenges in balancing data isolation, privacy protection, and model generalization during training and deployment. Issues of computational cost and latency remain pressing, especially in edge or resource-constrained environments, where LLMs often fall short in inference efficiency and lightweight deployment. More fundamentally, current LLMs lack true "industrial memory" and "evolutionary cognition"—they are unable to continually update their knowledge base or construct causal understanding of novel fault patterns as human experts do. Addressing these limitations will require breakthroughs in domain-specific knowledge injection, enhanced causal reasoning, multimodal temporal modeling, and adaptive model compression, paving the way toward industrial-grade LLMs with integrated cognitive, decision-making, and self-evolution capabilities.

The future development of LLMs in industrial prognostics and health management (PHM) will evolve toward greater specialization, intelligence, and deployability. This evolution is expected to follow several key trajectories:

1. Integration of Industrial Knowledge and Expert Capabilities

Future LLMs will shift from general corpus-based training toward deep integration with industrial ontologies, manuals, fault databases, maintenance logs, and expert knowledge. By incorporating structured knowledge graphs, symbolic rules, and few-shot learning mechanisms, these knowledge-enhanced LLMs will move beyond surface-level language understanding to model causal fault chains and maintenance logic in a traceable and interpretable way. This will empower LLMs with expert-level cognitive capabilities—enabling analogical reasoning, root cause analysis, fault localization, and maintenance recommendation—thus serving as reliable knowledge partners for engineers.

2. Deep Modeling of Multimodal and Heterogeneous Information

Industrial PHM inherently involves diverse modalities such as images, audio, vibration signals, text, and logs. Next-generation LLMs must evolve into multimodal perception-fusion-reasoning systems capable of processing time series, thermographs, and structural diagrams within a unified semantic space. Leveraging temporal and spatial encoding mechanisms in transformer architectures, these models can perform long-term modeling of dynamic system behaviors under noisy, nonlinear, and variable-load conditions—enhancing robustness and stability in real-world scenarios.

3. Explainability, Safety, and Trustworthiness Assurance

As LLMs assume increasingly critical roles in industrial decision-making, their outputs must be explainable and trustworthy to ensure system safety. A comprehensive assurance framework must be established, incorporating causal graph-based decision explanations, confidence scoring and anomaly detection modules, and human-in-the-loop mechanisms for high-risk interventions. Additionally, to address privacy concerns, privacy-preserving architectures based on federated learning, encrypted computation, and localized fine-tuning should be developed to enable secure cross-enterprise and cross-system model collaboration.

4. Industrial-Grade Deployment and Resource-Constrained Optimization

Real-world industrial settings demand lightweight, real-time, and scalable models. Advancements in efficient inference engines, model compression and distillation, and heterogeneous computing support will be essential for deploying LLMs on edge nodes, control terminals, and embedded platforms. Modular design and microservice packaging will further enable flexible integration and rapid iteration, promoting the transition from lab-scale feasibility to production-grade control.

5. Autonomous Learning and Self-Evolution

As industrial environments and task demands evolve, static LLMs struggle to keep pace with emerging knowledge needs. The future lies in models that learn independently, adapt proactively, and improve continuously—shifting from tool-like intelligence to long-term companion-like systems. Key advancements would focus on feedback-driven learning (e.g., ReAct, Reflexion) for self-reflection and iterative optimization. Lifelong learning frameworks will enable knowledge accumulation and adaptation across dynamic domains in PHM. Integrating cognitive architectures and world models will further enhance reasoning and environmental understanding. Ultimately, the shift from models that “answer questions” to those that can “think, learn, and evolve” will mark a fundamental leap forward in the path toward truly intelligent PHM systems.

Through these multi-faceted advancements, LLMs will continue to enhance their intelligence, domain

specificity, and practicality in industrial PHM, ultimately evolving from cognitive assistants to autonomous operation partners, and powering the next generation of intelligent manufacturing and maintenance ecosystems.

References

[1]. H. Wang and Y.-F. Li, "Empowering ChatGPT-Like Large-Scale Language Models with Local Knowledge Base for Industrial Prognostics and Health Management".

[2]. S. Lukens, L. H. McCabe, J. Gen, and A. Ali, "Large Language Model Agents as Prognostics and Health Management Copilots," PHM_CONF, vol. 16, no. 1, Nov. 2024.

[3]. A.J. Dave, T. N. Nguyen, and R. B. Vilim, "Integrating LLMs for Explainable Fault Diagnosis in Complex Systems," Feb. 08, 2024, arXiv: arXiv:2402.06695.

[4]. T. N. Nguyen, T. Downar, and R. Vilim, "A probabilistic model-based diagnostic framework for nuclear engineering systems," Annals of Nuclear Energy, vol. 149, p. 107767, 2020.

[5]. X. Liao, C. Chen, Z. Wang, Y. Liu, T. Wang, and L. Cheng, "Large language model assisted fine-grained knowledge graph construction for robotic fault diagnosis," Advanced Engineering Informatics, vol. 65, p. 103134, May 2025.

[6]. H. Huang, T. Shah, J. Karigiannis, and S. Evans, "Physics and Data Collaborative Root Cause Analysis: Integrating Pretrained Large Language Models and Data-Driven AI for Trustworthy Asset Health Management," 2024.

[7]. Y. Ma, S. Zheng, Z. Yang, H. Pan, and J. H. and, "A knowledge-graph enhanced large language model-based fault diagnostic reasoning and maintenance decision support pipeline towards industry 5.0," International Journal of Production Research, vol. 0, no. 0, pp. 1–22, 2025.

[8]. Zhou, "CausalKGPT: Industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing," Advanced Engineering Informatics, 2024.

[9]. L. Tao, H. Liu, G. Ning, W. Cao, B. Huang, and C. Lu, "LLM-based framework for bearing fault diagnosis," Mechanical Systems and Signal Processing, vol. 224, p. 112127, Feb. 2025.

[10]. S. Zheng, K. Pan, J. Liu, and Y. Chen, "Empirical study on fine-tuning pre-trained large language models for fault diagnosis of complex systems," Reliability Engineering & System Safety, vol. 252, p. 110382, Dec. 2024.

[11]. J. Zhang, C. Zhang, J. Lu, and Y. Zhao, "Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning," Applied Energy, vol. 377, p. 124378, Jan. 2025.

[12]. Z. Lai, C. Yang, S. Lan, L. Wang, W. Shen, and L. Zhu, "BearingFM: Towards a foundation model for bearing fault diagnosis by domain

knowledge and contrastive learning," International Journal of Production Economics, vol. 275, p. 109319, Sep. 2024.

[13]. Y. Xia et al., "FCLLM-DT: Enpowering Federated Continual Learning With Large Language Models for Digital-Twin-Based Industrial IoT," IEEE Internet Things J., vol. 12, no. 6, pp. 6070–6081, Mar. 2025.

[14]. L. Tao et al., "LLM-R: A Framework for Domain-Adaptive Maintenance Scheme Generation Combining Hierarchical Agents and RAG".

[15]. J. Jeon et al., "ChatCNC: Conversational machine monitoring via large language model and real-time data retrieval augmented generation," Journal of Manufacturing Systems, vol. 79, pp. 504–514, Apr. 2025.

[16]. Y. Liu, Y. Zhou, Y. Liu, Z. Xu, and Y. He, "Intelligent Fault Diagnosis for CNC Through the Integration of Large Language Models and Domain Knowledge Graphs," Engineering, 2025.

[17]. A.Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," Feb. 26, 2021, arXiv: arXiv:2103.00020.

[18]. Q. Li et al., "VSLLaVA: a pipeline of large multimodal foundation model for industrial vibration signal analysis," Sep. 03, 2024, arXiv: arXiv:2409.07482. Accessed: Oct. 30, 2024.

[19]. K. M. Alsaif, A. A. Albeshri, M. A. Khemakhem, and F. E. Eassa, "Multimodal Large Language Model-Based Fault Detection and Diagnosis in Context of Industry 4.0," Electronics, vol. 13, no. 24, p. 4912, Dec. 2024.

[20]. Y. Chen and C. Liu, "Remaining Useful Life Prediction: A Study on Multidimensional Industrial Signal Processing and Efficient Transfer Learning Based on Large Language Models".

[21]. L. Lin, S. Zhang, S. Fu, and Y. Liu, "FD-LLM: Large language model for fault diagnosis of complex equipment," Advanced Engineering Informatics, vol. 65, p. 103208, May 2025.

[22]. S. Jose, K. T. P. Nguyen, K. Medjaher, R. Zemouri, M. Lévesque, and A. Tahan, "Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models," Expert Systems with Applications, vol. 255, p. 124603, Dec. 2024.

[23]. S. Jose, K. T. P. Nguyen, K. Medjaher, R. Zemouri, M. Lévesque, and A. Tahan, "Bridging expert knowledge and sensor measurements for machine fault quantification with large language models," in 2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), Boston, MA, USA: IEEE, Jul. 2024, pp. 530–535.

[24]. H. Peng, J. Liu, J. Du, J. Gao, and W. Wang, "BearLLM: A Prior Knowledge-Enhanced Bearing Health Management Framework with Unified Vibration Signal Representation," Dec. 16, 2024, arXiv: arXiv:2408.11281.

[25]. W. Wang and D. Wang, "An Innovative Foundation Model for Bearing Prognostics and Health Management

Through Pre-Trained Large Language Models," 2025.

[26]. Y. Liu et al., "Brain-like Cognition-Driven Model Factory for IIoT Fault Diagnosis by Combining LLMs With Small Models," IEEE Internet Things J., pp. 1–1, 2024.

[27]. Zhang, S. Li, T. Hong, C. Zhang, and W. Zhao, "Enhanced Fault Prediction for Synchronous Condensers Using LLM-Optimized Wavelet Packet Transformation," Electronics, vol. 14, no. 2, p. 308, Jan. 2025.

[28]. W. Du et al., "Channel attention residual transfer learning with LLM fine-tuning for few-shot fault diagnosis in autonomous underwater vehicle propellers," Ocean Engineering, vol. 330, p. 121237, Jun. 2025.

[29]. P. Wang, S. Niu, H. Cui, and W. Zhang, "GPT-based equipment remaining useful life prediction," in ACM Turing Award Celebration Conference 2024, Changsha China: ACM, Jul. 2024.

[30]. Y. Wang et al., "RmGPT: Rotating Machinery Generative Pretrained Model," Sep. 26, 2024, arXiv: arXiv:2409.17604. Accessed: Nov. 18, 2024.

[31]. Z. Pang, Y. Luan, J. Chen, and T. Li, "ParInfoGPT: An LLM-based two-stage framework for reliability assessment of rotating machine under partial information," Reliability Engineering & System Safety, vol. 250, p. 110312, Oct. 2024.

[32]. Y. Qin, T. Zhang, Q. Qian, and Y. Mao, "Large Model for Rotating Machine Fault Diagnosis Based on a Dense Connection Network With Depthwise Separable Convolution," IEEE Trans. Instrum. Meas., vol. 73, pp. 1–12, 2024.

[33]. Z. Pang, H. Zhang, and T. Li, "Hybrid Fine-Tuning in Large Language Model Learning for Machinery Fault Diagnosis," in 2024 IEEE 22nd International Conference on Industrial Informatics (INDIN), Beijing, China: IEEE, Aug. 2024, pp. 1–6.

[34]. L. Tao et al., "Pre-Trained Large Language Model Based Remaining Useful Life Transfer Prediction of Bearing".

[35]. Eldele et al., "UniFault: A Fault Diagnosis Foundation Model from Bearing Data," Apr. 02, 2025, arXiv: arXiv:2504.01373.

[36]. J. Kaplan et al., "Scaling Laws for Neural Language Models," Jan. 23, 2020.

[37]. A.Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit and L. Jones, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[38]. A.Radford, J. W. Kim, C. Hallacy, A. Ramesh and Gabriel Goh "Learning transferable visual models from natural language supervision," in International conference on machine learning, 2021: PmLR, pp. 8748-8763.

[39]. J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in International conference on machine learning, 2023: PMLR, pp. 19730-19742.

[40]. H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," Advances

in neural information processing systems, vol. 36, pp. 34892-34916, 2023.

[41]. Q. Zhang, C. Xu, J. Li, Y. Sun, J. Bao, and D. Zhang, 'LLM-TSFD: An industrial time series human-in-the-loop fault diagnosis method based on a large language model', Expert Systems with Applications, vol. 264, p. 125861, 2025.

[42]. K. Du., B. Yang, K. Xie, N. Dong and Z. Zhang, 'LLM-MANUF: An integrated framework of Fine-Tuning large language models for intelligent Decision-Making in manufacturing', Advanced Engineering Informatics, vol. 65, p. 103263, 2025.

[43]. X. Pan, W. Zhuang, S. Wen, W. Yu, J. Bao, and X. Li, 'A context-aware KG-LLM collaborated conceptual design approach for personalized products: A case in lower limbs rehabilitation assistive devices', Advanced Engineering Informatics, vol. 66, p. 103422, 2025.

[44]. Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, 'Applications of machine learning to machine fault diagnosis: A review and roadmap', Mechanical Systems and Signal Processing, vol. 138, p. 106587, 2020.

[45]. S. Yang, L. Ling, X. Li, J. Han, and S. Tong, "Industrial battery state-of-health estimation with incomplete limited data toward second-life applications," Journal of Dynamics, Monitoring and Diagnostics, vol. 3, no. 4, pp. 246-257, 2024.

[46]. Y. Zhu, T. Guo, X. Li, Y. Zhang, and W. Zhang, "Domain generalization prognosis method for lithium-ion battery state of health with transformer and multi-kernel MMD," Journal of Dynamics, Monitoring and Diagnostics, vol. 3, no. 4, pp. 311-323, 2024.

[47]. X. Li, S. Yu, Y. Lei, N. Li, and B. Yang, 'Dynamic vision-based machinery fault diagnosis with cross-modality feature alignment', IEEE/CAA J. Autom. Sinica, vol. 11, no. 10, pp. 2068–2081, 2024.

[48]. C. Liu, X. Li, X. Chen, and S. Khan, 'Neuromorphic computing-enabled generalized machine fault diagnosis with dynamic vision', Advanced Engineering Informatics, vol. 65, p. 103300, 2025.

[49]. B. Yang, Y. Lei, X. Li, N. Li, X. Si, and C. Chen, 'A dynamic barycenter bridging network for federated transfer fault diagnosis in machine groups', Mechanical Systems and Signal Processing, vol. 230, p. 112605, 2025.

[50]. X. Li, W. Zhang, X. Li, and H. Hao, 'Partial domain adaptation in remaining useful life prediction with incomplete target data', IEEE/ASME Trans. Mechatron., vol. 29, no. 3, pp. 1903–1913, 2024.

[51]. K. Jeon and G. Lee, 'Hybrid large language model approach for prompt and sensitive defect management: A comparative analysis of hybrid, non-hybrid, and GraphRAG approaches', Advanced Engineering Informatics, vol. 64, p. 103076, 2025.

[52]. Y. Qin, T. Zhang, Q. Qian, and Y. Mao, 'Large model for rotating machine fault diagnosis based on a dense connection network with depthwise separable

convolution', IEEE Trans. Instrum. Meas., vol. 73, pp. 1–12, 2024.

[53]. Y. Lei, X. Li, X. Li, N. Li, and B. Yang, "Research on Large Model for General Prognostics and Health Management of Machinery," Journal of Mechanical Engineering, vol. 61, no. 6, pp. 1–13, 2025.

[54]. P. Liu, L. Qian, X. Zhao, and B. Tao, 'Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly', IEEE Trans. Ind. Inf., vol. 20, no. 6, pp. 8160–8169, 2024.

[55]. C. Men, Y. Han, P. Wang, J. Tao, and C.-G. Huang, 'The interpretable reasoning and intelligent decision-making based on event knowledge graph with LLMs in fault diagnosis scenarios', IEEE Trans. Instrum. Meas., vol. 74, pp. 1–16, 2025.

[56]. J. Chen, R. Huang, Z. Lv, J. Tang, and W. Li, " FaultGPT: Industrial fault diagnosis question answering system by vision language models," arXiv, preprint arXiv:2502.15481, 2025.

[57]. J. Zhang, C. Zhang, J. Lu, and Y. Zhao, 'Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning', Applied Energy, vol. 377, p. 124378, 2025.

[58]. Li, P. K.-Y. Wong, X. Tao, J. Ma, and J. C. P. Cheng, 'An interactive system for 3D spatial relationship query by integrating tree-based element indexing and LLM-based agent', Advanced Engineering Informatics, vol. 66, p. 103375, 2025.

[59]. Shagluf, A., Parkinson, S., Longstaff, A. P., & Fletcher, S. (2018). Adaptive decision support for suggesting a machine tool maintenance strategy: from reactive to preventative. Journal of Quality in Maintenance Engineering, 24(3), 376-399.

[60]. Pimenov, D. Y., Bustillo, A., Wojciechowski, S., Sharma, V. S., Gupta, M. K., & Kuntoğlu, M. (2023). Artificial intelligence systems for tool condition monitoring in machining: Analysis and critical review. Journal of Intelligent Manufacturing, 34(5), 2079-2121.

[61]. Bolton, R., Sheikhfathollahi, M., Parkinson, S., Basher, D., & Parkinson, H. (2025). Multi-Stage Retrieval for Operational Technology Cybersecurity Compliance Using Large Language Models: A Railway Casestudy. arXiv preprint arXiv:2504.14044.

[62]. Bolton, R., Sheikhfathollahi, M., Parkinson, S., Vulovic, V., Bamford, G., Basher, D., & Parkinson, H. (2025). Document Retrieval Augmented Fine-Tuning (DRAFT) for safety-critical software assessments. arXiv preprint arXiv:2505.01307.

[63]. Akhtar, S., Khan, S., & Parkinson, S. (2025). LLM-based event log analysis techniques: A survey. arXiv preprint arXiv:2502.00677.

[64]. Angelopoulos, J., Manettas, C., & Alexopoulos, K. (2024, October). Industrial Maintenance Optimization Based on the Integration of Large Language Models (LLM) and Augmented Reality (AR). In European Symposium on Artificial Intelligence in

Manufacturing (pp. 197-205). Cham: Springer Nature Switzerland.

[65]. Elhafsi, A., Sinha, R., Agia, C., Schmerling, E., Nesnas, I. A., & Pavone, M. (2023). Semantic anomaly detection with large language models. Autonomous Robots, 47(8), 1035-1055.

[66]. Niu, G., Dong, X., & Chen, Y. (2023). Motor fault diagnostics based on current signatures: A review. IEEE Transactions on Instrumentation and Measurement, 72, 1-19.

[67]. Keshvarparast, A., Battini, D., Battaia, O., & Pirayesh, A. (2024). Collaborative robots in manufacturing and assembly systems: literature review and future research agenda. Journal of Intelligent Manufacturing, 35(5), 2065-2118.

[68]. Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. Future Generation Computer Systems, 135, 364-381.

[69]. Amirizaniani, Maryam, et al. "Developing a framework for auditing large language models using human-in-the-loop." arXiv preprint arXiv:2402.09346 (2024).

[70]. Wadhwa, S., Hassanzadeh, O., Bhattacharjya, D., Barker, K., & Ni, J. (2024, November). Distilling Event Sequence Knowledge From Large Language Models. In International Semantic Web Conference (pp. 237-255). Cham: Springer Nature Switzerland.

[71]. Alshaikh, O., Parkinson, S., & Khan, S. (2025). A Contextual Framework to Standardise the Communication of Machine Learning Cyber Security Characteristics. Computer Standards & Interfaces, 104015.

[72]. Wu, J., Gan, W., Chen, Z., Wan, S., & Yu, P. S. (2023, December). Multimodal large language models: A survey. In 2023 IEEE International Conference on Big Data (BigData) (pp. 2247-2256). IEEE.

[73]. Wu, G., Cheng, C. T., & Pang, T. Y. (2024, November). Defect Classification and Localization in Material Extrusion with Multi-Modal Large Language Models. In 2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS) (pp. 539-544). IEEE.

[74]. Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., & Parkinson, S. (2023). A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. Ieee Access, 11, 144497-144529.

[75]. Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. ACM Computing Surveys, 57(6), 1-39.

[76]. Chow, K., Tang, Y., Lyu, Z., Rajput, A., & Ban, K. (2024, May). Performance optimization in the llm world 2024. In Companion of the 15th ACM/SPEC International Conference on Performance Engineering (pp. 156-157).

[77]. Paramanayakam, V., Karatzas, A., Anagnostopoulos, I., & Stamoulis, D. (2025, March). Less is more: Optimizing function calling for llm

execution on edge devices. In 2025 Design, Automation & Test in Europe Conference (DATE) (pp. 1-7). IEEE.

[78]. Maini, P., Jia, H., Papernot, N., & Dziedzic, A. (2024). LLM Dataset Inference: Did you train on my dataset?. Advances in Neural Information Processing Systems, 37, 124069-124092.

[79]. Huang, W., Abbeel, P., Pathak, D., & Mordatch, I. (2022, June). Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In International conference on machine learning (pp. 9118-9147). PMLR.

[80]. Baack, S., Biderman, S., Odrozek, K., Skowron, A., Bdeir, A., Bommarito, J., ... & Wolf, T. (2025). Towards Best Practices for Open Datasets for LLM Training. arXiv preprint arXiv:2501.08365.

[81]. Ma, Z., Chen, A. R., Kim, D. J., Chen, T. H., & Wang, S. (2024, April). Llmparser: An exploratory study on using large language models for log parsing. In Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (pp. 1-13).

[82]. Kleidermacher, H. C., & Zou, J. (2025). Science Across Languages: Assessing LLM Multilingual Translation of Scientific Papers. arXiv preprint arXiv:2502.17882.

[83]. H. Wang, C. Li, and Y.-F. Li, "Large-Scale Visual Language Model Boosted by Contrast Domain Adaptation for Intelligent Industrial Visual Monitoring," IEEE Trans. Ind. Inform., vol. 20, no. 12, pp. 14114 – 14123, Dec. 2024.

[84]. H. Wang, C. Li, Y.-F. Li, and F. Tsung, "An Intelligent Industrial Visual Monitoring and Maintenance Framework Empowered by Large-Scale Visual and Language Models," IEEE Trans. Ind. Cyber-Phys. Syst., vol. 2, pp. 166 – 175, 2024.

[85]. Y.-F. Li, H. Wang, and M. Sun, "ChatGPT-like large-scale foundation models for prognostics and health management: A survey and roadmaps," Reliab. Eng. Syst. Saf., vol. 243, p. 109850, 2024.

[86]. Y. Sun, Q. Zhang, J. Bao, Y. Lu, and S. Liu, "Empowering digital twins with large language models for global temporal feature learning," J. Manuf. Syst., vol. 74, pp. 83 – 99, 2024.