DOI: 10.19650/j.cnki.cjsi.J2514238

故障预测与健康管理领域大语言模型:应用与展望

彭 宇[1],季 拓[1,2],郭楚亮[1]
(1.哈尔滨工业大学电子与信息工程学院 哈尔滨 150001; 2.哈工大郑州研究院 郑州 450000)

摘 要:故障预测与健康管理(PHM)技术通过监测、分析和预测设备健康状态,实现主动维护和风险规避,是保障系统安全稳定运行的关键。尽管基于物理模型和数据驱动的 PHM 方法体系已经较为完整,但是经典方法在面对日渐复杂的工业系统的海量异构数据,特别是非结构化文本和多模态信息时,仍然存在专家知识集成困难,以及泛化能力不足的短板。近年来,Transformer 结构与大语言模型(LLM)方兴未艾,为 PHM 领域内专业知识的有效利用带来新兴的高精度预测范式,其应用潜力和优势包括但不限于知识提取与整合、少样本学习泛化、智能决策支持等。为全面综述大语言模型赋能 PHM 的现状及前景,首先,介绍 PHM 常见任务、Transformer 结构与通用大语言模型基本概念;其次,介绍领域专用大语言模型构建任务中的领域知识构建与注入方法(包括内部参数优化和外部知识增强),给出 PHM 领域专用大语言模型的整体框架;然后,从部件级、子系统级、复杂系统级这 3 个层次,并面向任务深入剖析并综述 PHM 领域大语言模型框架与应用现状,包括故障诊断、健康状态估计、剩余使用寿命预测和异常检测等典型 PHM 领域任务;最后,从模型轻量化、边缘部署、广义复杂系统角度,展望 PHM 领域专用大语言模型未来应用发展的挑战和机遇。

关键词: 大语言模型;故障预测与健康管理;故障诊断;健康状态;剩余使用寿命;异常检测
中图分类号: TP391 TH89 **文献标识码**: A **国家标准学科分类代码**: 510.99

Prognostics and health management domain-specific large language models: Applications and prospects

Peng Yu[1], Ji Tuo[1,2], Guo Chuliang[1]
(*1. School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China; 2. Zhengzhou Research Institute, Harbin Institute of Technology, Zhengzhou 450000, China*)

Abstract: Prognostics and health management (PHM) enables proactive maintenance and risk mitigation by monitoring, analyzing, and forecasting equipment health, undergirding safe and stable system operation. While physics-based and data-driven PHM paradigms are relatively mature, classical approaches struggle to integrate expert knowledge and generalize when confronted with massive heterogeneous data, especially unstructured text and multimodal information generated by increasingly complex industrial systems. Recently, the rapid development of Transformer architecture and large language model (LLM) have opened a high-precision prediction paradigm that efficiently exploits domain expertise, offering advantages such as knowledge extraction and fusion, few-shot generalization, and intelligent decision support. This review comprehensively surveys the current status and future prospects of PHM empowered by LLM. First, canonical PHM tasks, together with Transformer architecture and general LLM, are introduced. Second, domain-specific LLM construction is elaborated with respect to domain-knowledge creation and injection, covering internal parameter optimization and external knowledge augmentation; a unified framework for PHM domain-specific LLM is presented. Third, current PHM domain-specific LLM-based frameworks and applications are dissected in-depth from a task-oriented perspective across components, subsystems, and complex systems levels, focusing on fault diagnosis, state of health estimation, remaining useful life prediction, and anomaly detection. Finally, future challenges and opportunities are outlined regarding model compression, edge deployment, and generalized complex systems.

Keywords: large language model; prognostics and health management; fault diagnosis; state of health; remaining useful life; anomaly detection

收稿日期:2025-07-08 Received Date: 2025-07-08

0　引　　言

故障预测与健康管理(prognostics and health management, PHM)[1-3]是一种利用先进监测、数据分析,对设备或系统进行健康状态评估的综合性技术:故障预测通过分析监测传感器数据评估健康状态,提前发出故障预警;健康管理则根据故障预测结果,制定维护计划,优化健康状态。在航空航天、能源装备等高可靠性需求场景下,PHM 突破了事后维修、定期维护等传统策略的局限性,通过预测性维护有效降低运维成本、规避潜在的灾难性故障,并显著提升系统可靠性与可用性[4-5]。经典 PHM 方法大多基于物理模型和解析方法,通过建立系统退化机理及其数学方程实现预测[6-7]。随着图形处理器(graphics processing unit, GPU)算力与深度残差网络[8]的突破性进展,基于数据驱动的 PHM 方法已成为当前的研究热点,卷积神经网络(convolutional neural network, CNN)[9-11]的特征提取能力,以及长短期记忆网络(long short-term memory, LSTM)[12-14]的序列处理能力在 PHM 领域任务中得到广泛应用。然而,经典方法在面对日渐复杂的工业系统的海量异构数据,特别是非结构化文本和多模态信息时,在专家知识集成和泛化性方面短板凸显[15-16],具体表现为:

1)专家知识集成困难。传统深度学习模型主要擅长处理结构化或规整的数值型传感数据,但对蕴含丰富领域专家经验的非结构化文本数据(例如维修日志、故障报告、操作手册、技术文档)的融合能力十分有限。研究表明,工业 PHM 场景超半数有价值信息存储于非结构化文本中[17-18]。关联文本描述特征与传感器频谱特征,需要复杂的特征工程和先验知识建模,过程繁琐且难以被 CNN、LSTM 等传统模型有效利用[19];

2)模型泛化能力不足。传统模型通常针对特定数据集和工况,面对设备变型、运行条件变化(例如域偏移)、数据稀疏(例如罕见故障模式)等场景时,泛化性能急剧下降[20-21];针对数据稀缺的罕见故障(例如长尾分布),传统模型由于缺乏对领域概念的深层次理解和强大的小样本学习能力,存在严重的漏报风险[22]。

大语言模型(large language model, LLM)[23-24]近年来,迅速发展,为解决上述挑战提供了新途径,其应用潜力和优势包括但不限于:

1)知识提取与整合。LLM 能够直接从海量非结构化文本中学习语义、抽象实体关系、构建领域知识,有效整合维修记录、手册等文本信息的专家经验[25-26];

2)少样本学习泛化。LLM 的自注意力机制能够捕捉长距离依赖和复杂上下文信息,结合大规模预训练中习得的通用知识,对未见过的设备变体、工况变化、乃至跨任务,少样本甚至零样本学习场景下,都具有更强的任务适应性和泛化能力[23,27];

3)智能决策支持。LLM 的文本理解、文本生成和推理能力,能够作为 PHM 智能决策支持系统一部分,通过构建问答系统,可自动化生成维护建议和故障排除步骤,提升维护效率[28]。

近年来,结合 LLM 的 PHM 方法综述大多聚焦利用现有大语言模型的通识能力,包括 LLM 辅助的预测性维护[29],LLM 在 PHM 中的优化技术与应用[30],以及基于 LLM 的表征学习和故障预测的应用潜力[31]。然而,通用 LLM 在 PHM 领域的专业知识相对匮乏,尤其是针对故障诊断、健康状态(state of health, SOH)估计、剩余使用寿命(remaining useful life, RUL)预测、异常检测等具体任务,缺乏对预期效用和解决方案的深入分析,亟需针对领域专用大语言模型的框架探索和应用研究。为归纳概述 PHM 领域大语言模型"是什么? 为何用? 怎么用?"的问题,综述 LLM 赋能故障预测与健康管理的研究现状与应用展望,主要内容包括:

1)简介 PHM 常见任务、通用 LLM 原理、构建方法和局限性,引入领域专用 LLM 必要性;

2)介绍针对领域专用 LLM 的领域知识构建和注入方法,建立 PHM 领域专用 LLM 的整体构建框架;

3)从部件、子系统、复杂系统 3 个层次,针对故障诊断、SOH 估计、RUL 预测、异常检测等具体任务,综述 PHM 领域专用 LLM 的现有应用和潜在价值;

4)从轻量化边缘部署、广义复杂系统等角度,总结并展望 PHM 领域专用 LLM 面临的机遇和挑战。

1　背　　景

1.1　PHM 常见任务

在工业系统的运维管理中,PHM 凭借一系列结构化任务,保障系统的可靠性与安全性[3]。故障诊断、SOH 估计、RUL 预测、异常检测是 4 类代表性 PHM 任务,分别侧重于识别、评估、预测和检测,是实现预防性维护、提升系统韧性和优化资源配置的关键环节。

1)故障诊断

故障是指设备或系统未能按预期执行其规定功能的状态或事件。故障诊断作为 PHM 的基础性任务,其代表性体现在它在设备出现异常时能够迅速定位问题所在,并侧重于 PHM 的"识别"部分。故障诊断通过监测和分析设备或系统运行状态,及时发现并识别故障模式,为设备或系统维护和可靠性管理提供支持。

轴承故障诊断是具有代表性的一类故障诊断任务,在复杂工况下易受磨损、疲劳、润滑不良等因素影响,且故障特征常表现出复杂性和隐蔽性[32-33]。多所高校和科

研机构均制作了轴承故障诊断相关数据集。例如凯斯西储大学数据集(Case Western Reserve University bearing dataset, CWRU)共包含内圈故障、外圈故障和滚动体故障这 3 类故障,每种故障分不同尺寸的人为制造故障点、不同大小电动机马力。数据集文件分为正常数据、12 kHz 采样率驱动端轴承故障、12 kHz 采样率风扇端轴承故障、48 kHz 采样率电机端轴承故障这 4 类,轴承工作状态信息均使用加速度传感器采集。

2)健康状态估计

SOH 估计是 PHM 中关注设备当前"健康度"的关键任务,其代表性体现在为理解设备当前性能衰退程度提供量化依据,并侧重于 PHM 的"评估"部分。SOH 估计基于电池或其他储能设备的运行数据(如电压、电流、温度等),精准评估其当前的健康状态。电池的 SOH 可用容量衰减的形式量化,定义为电池当前可用电量与初始状态总电量的百分比,即:

$$SOH(\%) = \frac{\text{当前电池容量(mAh)}}{\text{初始电池容量(mAh)}} \times 100\% \quad (1)$$

通常当 SOH 降低至 80%时,电池达到第 1 次使用寿命,可在储能电站等领域进行二次利用。电池的 SOH 估计对于优化充电策略、延长设备寿命、提高系统可靠性和安全性,实现预测性维护具有重要意义[34-35]。

电池 SOH 估计任务的数据集可通过自行搭建实验平台方式收集建立。麻省理工学院的 MIT(Massachusetts Institute of Technology)数据集[36]是目前广泛应用的商业锂离子电池数据集之一,包含 124 个商业锂离子磷酸铁锂或石墨电池在快速充电条件下循环的容量数据,循环寿命从 150~2 300 个循环不等,涵盖不同的快速充电策略,记录了电池在循环过程中的电压、电流、温度和内部电阻等信息。

3)剩余使用寿命预测

RUL 预测作为 PHM 中的核心前瞻性任务,其代表性体现在它直接关乎预防性维护的决策时机,侧重于 PHM 的"预测"部分。RUL 预测基于设备的运行状态数据和历史故障信息,精准预测设备剩余的可正常使用时长,对提高工业部件或系统的可靠性和运行安全性、避免致命故障、降低维护成本具有重要意义[37-38]。RUL 定义为设备或系统在达到预设的时效阈值,或其预期功能失效前,所剩余的正常运行时间,即:

$$RUL = T_f - T_0 \quad (2)$$

其中,T_0 为当前时间,T_f 为预计到达使用寿命终点的时间,单位由具体应用场景而定,例如时间(小时、天)、循环次数(航空发动机飞行次数循环、电池充放电循环)、里程数(车辆行驶)等。

美国国家航天局发布的商用模块化推进系统模拟器(commercial modulur aero propulsion system simulation, C-MAPSS)数据集[39]是 RUL 预测任务训练和测试的经典基准数据集,模拟了一个商用涡扇发动机,使用温度、压力、转速等多种类型传感器[7]。数据集共包含 4 个训练集、测试集和真实寿命集构成的子集,每个子集内均有至少一种工作状态和故障模式。

4)异常检测

异常是指设备或系统运行数据中与预期正常行为模式显著偏离的事件或模式,是 PHM 中"预警"性质的任务,能够及时发现与正常行为模式不符的时间或数据点,侧重于 PHM 的"检测"部分。与故障诊断中关注设备未能按预期执行其规定功能的状态或事件不同,异常检测通过持续监测设备或系统的运行数据,利用机器学习等方法建立正常行为基线,在新的数据偏离该基线时触发警报,提示潜在的故障或风险。因此,异常检测对于故障前干预、避免系统停机,以及减少损失方面至关重要[40-41]。SKAB(skoltech anomaly benchmark)数据集是异常检测任务中具有代表性的数据集之一,通常被视为一个完整的工业系统,包含从测试台传感器收集的多元时间序列数据(例如加速度、电流、电压、温度等),并且提供了异常的标签,可用于评估异常检测方法的性能。

1.2 通用大语言模型

1)Transformer 结构

Transformer 结构[25]通过自注意力机制计算输入序列中元素间相关性,以捕捉序列中各元素的依赖关系,从而处理复杂序列信息,在具有优于循环神经网络的计算性能的同时具有较好的网络子结构稳定性。

如图 1 所示,Transformer 结构[25]具有由多个相同层堆叠的编码器-解码器结构。编码器每个层内包含多头自注意力和前馈神经网络两个子层;解码器额外包含(面向编码器-解码器的)自注意力机制,并在每个子层后采用残差连接和层归一化,以提升收敛稳定性。

Transformer 核心的多头自注意力机制允许模型同时关注序列中的所有其他元素,并计算相关性权重。相比 LSTM 等只能逐步或有限地关注局部信息的传统序列模型,多头机制通过并行使用多个独立自注意力模块,从不同的表示子空间学习信息,更全面地捕捉复杂的长期依赖关系。然而,自注意力机制无法感知序列元素的时间先后关系,无法正确理解序列的物理或逻辑流向。为此,Transformer 引入位置编码并与词嵌入相加,注入关于单词在序列中绝对或相对位置的信息。此外,为提高模型的训练效率和稳定性,每个自注意力和前馈网络子层都集成了残差连接与层归一化:残差连接有助于缓解在深度网络中出现的梯度消失问题,层归一化则能稳定层间输入分布,二者共同作为深度模型优化的关键结构,保障了模型训练的稳定性和效率。

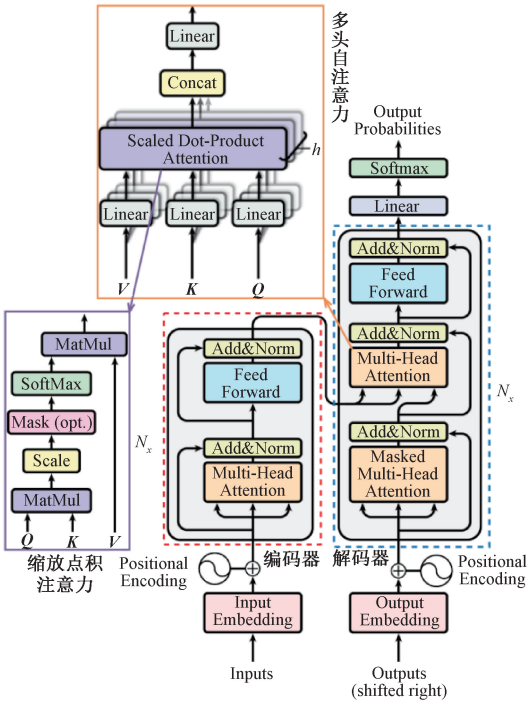

图 1　Transformer 结构

Fig. 1　Transformer architecture

2)大语言模型分类

大语言模型以 Transformer 结构为基本构建模块,依据对文本理解或文本生成能力的不同倾向,选择性地采用编码器和解码器。如图 2 所示,当前主流大语言模型可分为:自编码模型(单编码器)、自回归模型(单解码器)和序列到序列模型(编码器-解码器)。

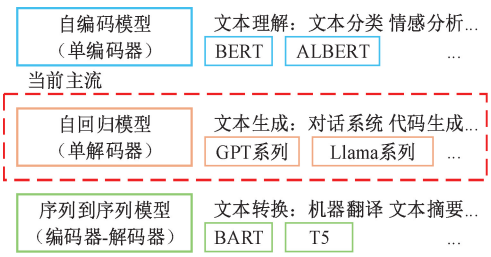

图 2　大语言模型分类

Fig. 2　Large language model classification

LLM 发展早期,BERT(bidirectional encoder representations from transformers)[26]、ALBERT(adversarial learning of BERT)[42]等为代表的自编码模型,采用掩码语言建模和下一句预测方式,能够准确捕捉词语和句子的深层语义关系,从而在文本分类、情感识别等文本理解类任务中表现出色。

随着对话问答任务对文本生成能力的需求日愈强烈,GPT(generative pre-trained transformer)[43]、LLaMA(large language model meta AI)[44]等为代表的自回归模型,通过预测序列中的下一个词进行自回归训练,并逐步生成连贯自然的文本。自回归模型是当前高性能 LLM 的主流架构,具有较低的计算复杂度,被广泛应用于对话系统、代码生成等生成类任务。复杂的序列到序列转换任务,需兼顾文本理解与文本生成能力。BART(bidirectional and auto-regressive transformers)[45]、T5(text-to-text transfer transformer)[46]等序列到序列模型结合 Transformer 结构的编码器和解码器部分,将所有任务统一为"文本到文本"转换问题,实现理解输入序列并生成相应的输出序列,因此被应用于机器翻译、文本摘要生产等文本转换类任务。

3)大语言模型构建方法

大语言模型从最初的通用语言理解能力逐步发展为能够遵循人类指令、安全且有益的通用智能体,需要多阶段的复杂训练过程。如图 3 所示,现代 LLM 的通用训练范式主要包括 4 个核心阶段。

(1)预训练(pre-train),在海量无标注文本数据上自监督学习,通过预测下一个词或被掩盖的词,习得语言的语法、语义和丰富的世界知识[26,43]。这一阶段是 LLM 通用能力的基础,产生具备语言理解与生成能力的基础模型,但输出结果难以与人类偏好完全对齐。

(2)监督微调(supervised fine-tuning),在高质量、人工标注的"指令-响应对"数据集上进行有监督学习,使基础模型更好地理解和遵循人类指令,能够学习到如何针对特定指令生成期望的有效回复[47]。

(3)奖励建模(reward modeling),收集人类对模型不同响应的比较数据,训练独立的奖励模型,以更好地捕捉人类偏好,提升监督微调输出的质量和安全性,并评估任意文本响应与人类偏好的一致性[48]。

(4)对齐与优化(alignment & optimization),将模型行为与人类价值观和偏好进行更深层次的对齐,主要方法包括人类反馈强化学习(reinforcement learning from human feedback, RLHF)和直接偏好优化(direct preference optimization, DPO)。早期广泛采用的 RLHF 引入奖励模型作为强化学习的奖励函数,指导 LLM 通过近端策略优化等算法,生成能够最大化奖励的预测文本[49],从而使输出更细致地趋近于人类偏好,保证结果安全且无害;DPO 通过直接优化模型参数来最大化人类偏好,避免了显式训练奖励模型和复杂的强化学习过程[50]。DPO 通常计算更高效、训练更稳定,在许多场景下能达到与 RLHF 相似甚至更优的对齐效果,是当前 LLM 对齐的重要手段。

通过分阶段迭代上述阶段,LLM 从一个庞大的语言预测器,训练成为一个能够理解复杂指令、生成高质量响应,并与人类价值观高度对齐的人工智能助理。

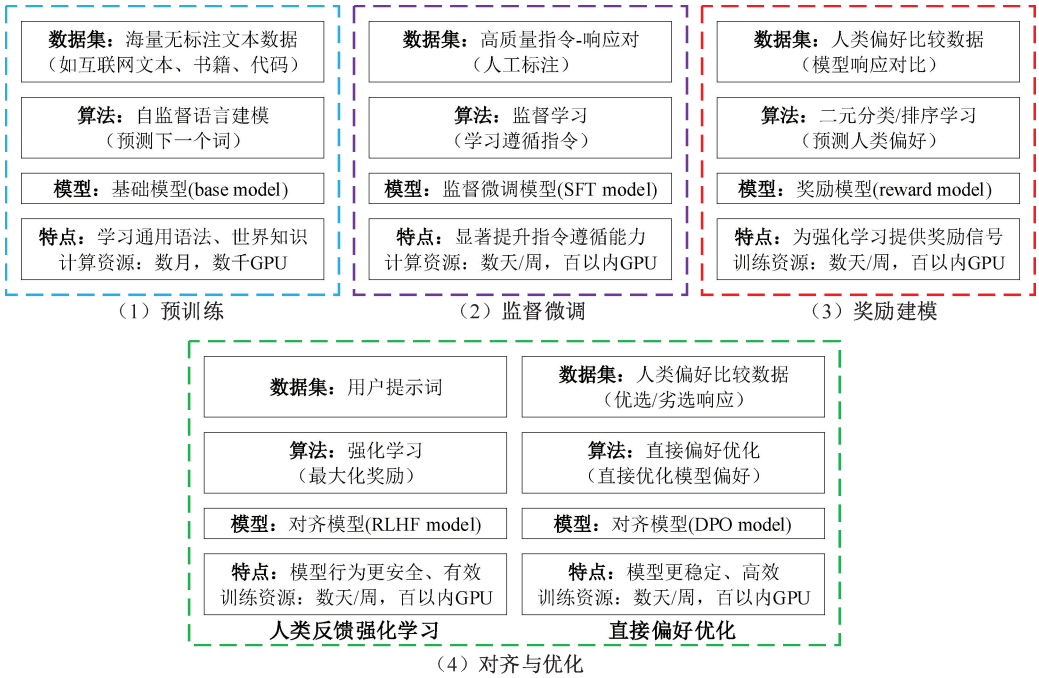

图 3 大语言模型通用训练范式

Fig. 3 General training paradigm for large language models

4)通用大语言模型的局限性

尽管参数量日益增加的基础大语言模型在通用任务中展现出优异的理解和生成能力,但应用于专用领域的特定任务时,仍面临结果可信问题,即:

(1)通用语料库难以覆盖特定领域的深层专业知识体系与术语网络,导致通用模型在面对专业领域的复杂概念时,容易出现知识缺失或理解偏差[51-52];

(2)通用模型的训练语料通常较为宽泛,导致其在专业场景下的推理精度不足,难以捕捉领域特有的细微差别和隐含意义,从而影响决策的准确性[53];

(3)通用大语言模型依赖静态知识,存在难以适应领域数据与知识的动态更新需求、难以反映领域内快速演进的最新信息和行业动态[54],以及难以适配需要高度事实准确和可靠的专业领域[55]等问题。具体而言,通用 LLM 在处理如“某型号航空发动机在巡航阶段出现高频振动”这类具体工程问题时,其局限性尤为突出。由于训练语料中缺乏深度的物理知识和精确的故障模式数据,模型容易出现“幻觉”,即生成看似合理但事实错误的内容,例如将高频振动归因于不相关的润滑油泵压力异常[55]。同时,模型也难以区分不同设备间的细微差异,可能混淆 CFM56 与 GE90 等发动机的特有故障模式[56]。这种高置信度的错误可能导致错误的诊断和维护决策,进而造成严重的安全事故和巨大的经济损失,是不可接受的。

为克服上述挑战,并充分发挥大语言模型在特定行业和应用场景中的潜力,需要构建具备高度事实准确性的领域专用(domain-specific)大语言模型(也称为垂域模型)。领域专用 LLM 是利用特定领域的语料库和知识,对通用大语言模型进行二次预训练、微调或检索增强等产生的大语言模型,能够深度理解和掌握特定领域的语言模式、专业知识和推理能力。在现阶段涉及专业知识的对话问答、数据分析等任务中,领域专用 LLM 能够适应领域知识的动态性,从而展现出更高的准确性、专业性和可靠性[57-58]。

2 PHM 领域大语言模型框架

尽管大语言模型凭借其卓越的文本理解、生成和推理能力,为 PHM 领域带来了前所未有的机遇。然而,通用 LLM 在面对 PHM 领域特有的复杂多源异构数据和专业知识时,受限于训练语料中领域知识的匮乏,往往难以充分发挥其潜力,构建领域专用大语言模型是现阶段 PHM 智能化发展的必然趋势。本章从领域知识构建和领域知识注入的角度,介绍 PHM 领域专用 LLM,并建立知识构架和知识注入的整体构建框架。

2.1 领域专用大语言模型构建

1)领域知识构建

领域知识构建是通用大语言模型向特定领域专用 LLM 演进的先决条件与应用基础,涉及从多源异构数据中高效地获取、整理、提炼与结构化特定领域内的专业知识。领域知识的质量高低直接决定了模型能够学习和理

解的专业信息的广度和深度,是保证后续知识注入有效性以及领域专用 LLM 性能的关键。领域知识构建通常包括数据采集、知识抽取与表示,以及知识质量管理这 3 个主要环节。

(1)数据采集涵盖了从各种渠道获取海量的非结构化文本、半结构化表格数据及结构化信息。构建高质量的领域语料库是这一阶段的核心任务,为后续的模型训练和知识注入提供了基础性的数据支撑[59]。

(2)知识抽取与表示是核心步骤,通过自然语言处理技术,从非结构化文本中提炼出实体、事件及其相互关系,并将其转化为结构化的知识表示形式。这种结构化表示能够将分散的领域信息组织起来,为大语言模型提供丰富的语义上下文。知识的抽取与表示方法论涵盖了从信息源识别到语义单元抽取的全过程[60]。

(3)知识质量管理贯穿领域知识构建的始终,包括对采集数据和抽取知识的清洗、去重、纠错,以及通过人工审核与标注等方式进行验证,以确保所构建知识的准确性、一致性和完整性。高质量的知识管理是确保知识注入有效性的基础,特别是在结构化知识表示中,数据质量问题对应用效果具有显著影响[61]。

2)领域知识注入

领域知识注入[62]是基于通用 LLM 构建领域专用 LLM 的核心。如图 4 所示,领域专用 LLM 的构建方法根据领域知识注入的实现方式,主要可分为内部参数优化和外部知识增强两大类:内部参数优化直接调整通用模型的权重,使其具备特定领域能力;外部知识增强则通过引入模型外部的信息或机制,使通用大语言模型能够查询检索领域知识。

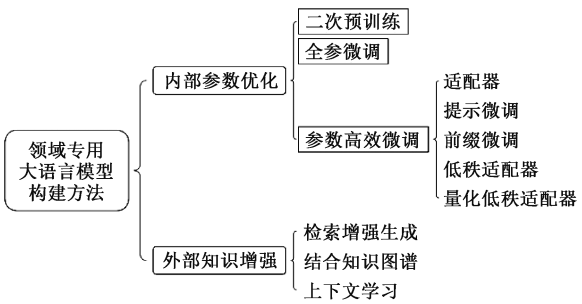

图 4　领域专用大语言模型构建方法概述

Fig. 4　Overview of methods for constructing domain-specific large language models

(1)内部参数优化

二次预训练即领域自适应预训练,在通用大语言模型基础上,利用大量特定领域的无标注数据,进一步预训练[63];全参微调针对领域内特定任务,使用带标注的领域专用数据集,对预训练模型的所有参数进行训练[64]。二次预训练和全参微调是构建领域专用大语言模型的直接方法,但均需要高昂的计算成本,且存在数据集较小时的过拟合问题。

为了降低模型微调的计算量和延迟,参数高效微调[65]仅训练模型的部分参数或添加少量新参数,即可使得通用大语言模型具备领域特定任务所需的语料理解能力。近年来,主流的参数高效微调方法包括:

a. 适配器[66]。在预训练模型的 Transformer 层中插入小型全连接网络,训练新增模块的参数,实现参数共享与模型紧凑化,但也可能增加推理延迟;

b. 提示微调[67]。优化嵌入层的输入提示向量,将离散提示优化问题转为连续提示优化问题,在冻结预训练模型主体参数的情况下,高效引导模型适应领域特定任务,但其性能对初始化和任务复杂度敏感;

c. 前缀微调[68]。在每一层的隐藏状态前添加可训练的前缀向量,并利用前馈网络对其进行参数化,为生成任务提供轻量级适配,同样可能引入推理延迟。

d. 低秩适配器(low-rank adaptation, LoRA)[69]。将权重矩阵分解为两个低秩矩阵的乘积,量化低秩适配器(quantized LoRA, QLoRA)[70]通过将预训练模型量化为 4 位,并保持与 16 位微调相当的性能,大幅减少参数和内存需求,在性能和效率之间取得平衡。

这些方法不仅显著降低了模型微调的计算和存储开销,还有助于缓解全参数微调时可能出现的灾难性遗忘问题,并使得模型的迭代和管理更为灵活。

(2)外部知识增强

性能强大的通用基础大语言模型通常为黑盒模型,使用者难以针对应用领域进行参数微调。为避免调整模型的内部参数,外部知识增强技术旨在通过引入外部信息来提升模型在特定任务上的表现。外部知识库可以包含海量的设备维修手册、历史故障案例报告、技术通讯、甚至是工程师的非结构化工作笔记。

a. 检索增强生成(retrieval augmented generation, RAG)[71]结合了参数化和非参数化记忆,是一种针对知识密集型任务的通用微调方法。在生成文本时,通过联合训练可微分的检索器和生成器,能够从外部非参数记忆中实时获取相关信息,并作为上下文,生成更准确且可溯源的文本。有效解决了模型“幻觉”并提供最新知识,但依赖于外部知识库的质量。

b. 知识图谱[72]以结构化形式,表示实体、概念和相互关联的知识库,有效组织和管理领域知识,可有效弥补通用模型在特定领域知识方面的不足。PHM 领域的知识图谱可以定义部件、传感器、故障模式、物理参数等实体。知识图谱将信息融入训练数据辅助内部参数监督微调[73],可以作为大语言模型的外部工具,在推理时通过查询获取事实性信息,动态地增强知识储备和推理能力[74],从而提供结构化高质量的事实性知识,但通常构

建与维护成本较高。

c. 上下文学习[23]利用通用 LLM 的模式识别和泛化能力,在输入提示中提供任务指令和少量输入-输出示例。通过将领域特定的外部上下文信息直接作为输入,引导模型遵循示例中隐含的任务逻辑,无需进行领域特定微调,即可快速适应新任务[75],但对上下文质量和长度要求较高。

不同外部知识增强方法各有侧重,为黑盒大语言模型提供了不进行参数微调的性能提升途径。检索增强生成擅长提供实时、可溯源的知识;知识图谱则通过提供结构化的事实性知识来弥补模型在特定领域知识上的不足;而上下文学习则以其无需微调的便捷性,在快速适应新任务方面表现突出。

(3)方法对比与选择

内部参数优化与外部知识增强两类方法通过更新模型参数和动态检索外部知识库的方式实现领域知识的注入,并在成本、性能和知识时效性等方面各有侧重。

内部参数优化直接将领域知识固化于模型参数中,推理速度快,一体化程度高。其中,二次预训练和全参微调效果最好,但需要海量领域数据和高昂的计算资源(数百至数千 GPU 月),适用于构建核心基础垂域模型的场景。参数高效微调则是资源受限下的首选,LoRA 及其变体 QLoRA 是当前主流。QLoRA 相较于 LoRA,通过 4 位量化可将微调所需的显存降低约 75%,使得在消费级 GPU 上微调大型模型成为可能,但量化可能带来微小的精度损失。适配器和前缀微调虽也能有效降低训练参数,但可能引入额外的推理延迟。

外部知识增强不改变模型内部参数,灵活性高,尤其适用于无法访问模型权重的闭源 LLM 或需要频繁更新知识的场景。RAG 能够有效缓解模型"幻觉",并提供可溯源的答案,其维护成本主要在于向量数据库的构建和更新,相对较低。知识图谱能提供高度结构化的事实性知识,推理逻辑性强,但其构建与维护成本(包括知识抽取、对齐、更新)远高于 RAG。上下文学习成本最低,无需任何训练,适用于快速验证和原型设计,但受限于示例质量和 LLM 上下文长度。实践中,常结合使用这些方法,例如使用 LoRA 对模型进行初步的领域适配,再结合 RAG 技术让其具备访问最新动态知识库的能力,从而平衡成本与性能[76-77]。

2.2 PHM 领域大语言模型整体框架

经典 PHM 方法难以处理日益复杂的工业系统的多源异构数据和非结构化信息,通用大语言模型虽为 PHM 领域带来机遇,但其在专业知识上的匮乏是显著挑战,尤其是故障诊断、SOH 估计、RUL 预测等具体任务中难以深入分析预期效果和解决方案。领域专用 LLM 凭借其深度融合专业知识和数据的能力,为 PHM 带来了变革,是克服通用 LLM 局限性、充分发挥其在 PHM 智能化发展中潜力的必然趋势[78]。领域专用大语言模型深度融入 PHM 各环节的整体框架如图 5 所示,该框架的核心思想在于,以 LLM 为认知中枢,将原本分散、异构的多模态数据流,通过统一的语义表征,汇聚到 PHM 核心功能层进行分析,最终通过人机交互层,将复杂的分析结果转化为可理解、可执行的智能决策,实现从多模态数据采集,到信息数据分析,再到决策支持的全流程智能化闭环。

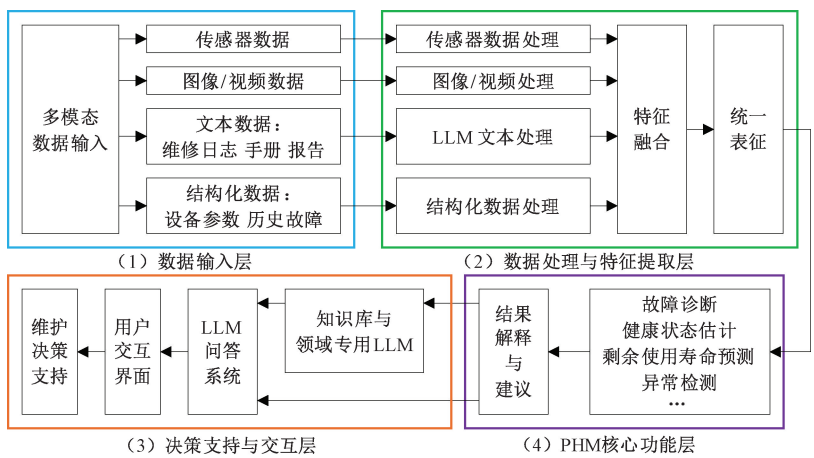

图 5 PHM 领域大语言模型整体框架

Fig. 5 Large language models for PHM framework

1)数据输入层

全面收集设备在设计、制造、测试、运行过程中的多模态信息,涵盖了物理世界和信息世界丰富的数据类型。其中,传感器数据(例如振动、温度、压力、电流等时间序列数据)直接反映了设备的物理状态和动态数据;图像/视频数据(例如摄像头、红外热成像、X 射线等获取)用于

发现表面缺陷、结构损伤等视觉异常[79];非结构化文本数据(例如维修日志、故障报告、操作手册、专家经验等)提供丰富领域知识的信息源[19];结构化数据(例如设备参数、故障记录、维护计划等)则为分析提供基础信息。

2)数据处理与特征提取层

对原始数据进行清洗、转换等操作,并从中提取出高价值特征。包含不同模态数据的专业处理模块,如传感器、图像、结构化数据处理模块负责各自模态的降噪、特征工程和模式识别。其中,文本处理模块是 LLM 发挥关键作用之处,通过自然语言理解、命名体识别、关系抽取等技术,从海量文本数据中自动抽取、组织和构建知识图谱,将原本的非结构化数据转化为机器可理解和利用的形式[80]。随后会通过多模态融合技术将其整合[81],如在输入层直接拼接特征的早期融合、在决策层整合结果的晚期融合、通过交叉注意力机制在中间层面对特征进行动态加权以实现深度交互的混合融合[82]。

最终,融合后的特征会形成一个统一的表征向量,该向量再通过一个线性投影层[83]进行维度映射和空间对齐,从而转换为与词元表兼容的伪词元序列[82-83],以适配 Transformer 模型的统一输入结构,为后续核心功能层的分析预测提供高质量、一体化的数据基础[84]。

3)PHM 核心功能层

基于统一的表征进行分析和预测,包含故障诊断、SOH 估计、RUL 预测、异常检测等 PHM 领域内核心任务的功能,领域专用 LLM 融合历史运行数据、状态监测指标、环境因素等多模态数据,全面评估设备的整体健康状况,生成详细的健康报告,并进行趋势分析和风险评估。解释与建议模块确保 LLM 不仅给出分析结果,还能用自然语言解释原因,并提供初步的维护和建议,以增强系统的可解释性[85]

4)决策支持与交互层

将复杂分析结果转化为可操作的智能决策。此层引入了知识库与 PHM 领域专用 LLM,存储 PHM 领域的专业知识、历史故障案例、维修规程等,并由 LLM 持续学习和更新,提供强大的推理能力。PHM 系统的领域专用 LLM 问答系统是与用户交互的核心:LLM 通过理解用户自然语言提问的查询意图,获取 RUL 预测、故障诊断、维护方案、故障分析等信息,提供多层次、多角度的精准回答[86]。用户交互界面直观展示 PHM 分析结果、健康趋势、维护计划等。最终,维护决策支持模块基于 LLM 分析结果,智能推荐最佳维护时机,生成详细的维护工单,并持预测性维护、预防性维护等维护策略的制定与实施[87]。

综上,基于领域专用 LLM 的 PHM 框架在数据处理和问答决策方面,能够展现出多角度的特点和能力:

1)通过强大的多模态理解与融合,深度理解并处理 PHM 全流程中产生的多源异构数据;

2)具备智能知识抽取与推理能力,能够从海量非结构化文本中构建 PHM 领域知识图谱;

3)具备上下文感知和理解,能够提供高度定制化的诊断和预测结果,且通过提供诊断或预测结果的依据,增强决策建议的解释性与透明度;

4)通过人机协同与知识增益,辅助 PHM 工程师提高工作效率,通过交互反馈持续学习优化模型能力。

3　层次化 PHM 任务领域专用大语言模型

PHM 领域任务涵盖从不同层次对设备及系统进行监测和分析。领域专用 LLM 能够赋能 PHM 系统,有效应对不同层次化任务的挑战,例如,部件级任务中,侧重于传感器数据处理与特征识别,建模复杂度相对较低,主要通过学习多模态数据来替代复杂的物理建模;子系统级任务中,侧重于部件间耦合效应的理解和多模态数据融合分析,建模复杂度提升,需处理更复杂的逻辑关联;复杂系统级任务中,聚焦于宏观决策支持与系统风险评估,通过深度整合各层次信息并进行高级推理,从局部洞察转向全局优化。

本章将从部件、子系统、复杂系统 3 个层次介绍 PHM 领域专用大语言模型的应用,并对已有的研究进行分析讨论,各层次任务及领域专用大语言模型举例如图 6 所示。

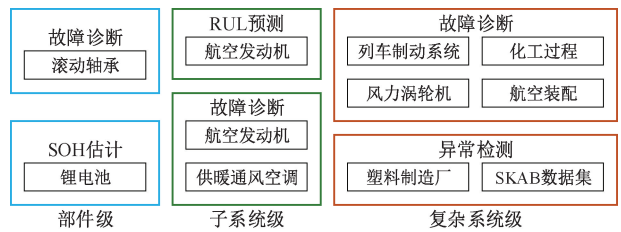

图 6　PHM 领域大语言模型概述

Fig. 6　Overview of PHM domain-specific LLM

3.1　部件级 PHM

现代工业体系中,各类设备或系统均由不同部件协同工作,任何单一部件的潜在故障都有可能导致整个设备或系统的性能下降甚至停机,从而造成巨大的经济损失和安全风险。部件级 PHM 是设备健康管理的基础层次,用于对单一、独立的机械部件(如轴承、齿轮、阀门等)进行状态监测、故障诊断等任务。在这一层级,LLM 的引入显著降低了对复杂物理模型的依赖,展现出低建模复杂度的优势。

LLM 能够通过学习海量的多模态数据,包括传感器时间序列、图像、以及非结构化的维修日志和操作手册,自动捕捉部件的健康特征和故障模式,无需显式构建复杂的物理方程[19]。例如,LLM 可以从维修人员对异响、异味、轻

微振动等现象的语言描述中,结合历史数据,直接识别并预测部件的早期故障,极大地简化了从原始数据到诊断结果的路径[80]。例如,Mamba[88]等线性模型的应用,能进一步提升处理速度并降低计算复杂度,使其在处理长序列数据时更具优势,使得部件级 PHM 能够更快部署和迭代。

1)部件故障诊断

面对跨条件、小样本和跨数据集等适应性挑战,Tao 等[89]提出了一种基于 LLM 的轴承故障诊断框架,如图 7 所示,将振动数据的时域和频域特征编码为文本特征向量,使用 QLoRA 方法对 ChatGLM2-6B[90]模型进行参数高效微调,以增强其对轴承振动信号特征的理解和分析能力。在 CWRU 等 4 个公开数据集上,分别进行的单数据集、单数据集跨条件、全数据跨数据集,以及有限数据跨数据集等各类适应性实验表明,该框架能有效提升模型的泛化能力,跨数据集学习后的诊断准确率有明显提升,具有复杂工程场景的适用性。

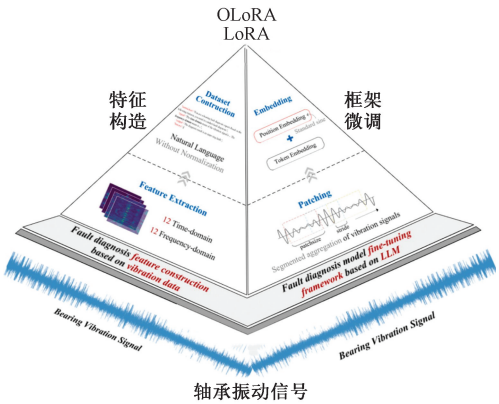

图 7 基于 LLM 的轴承故障诊断框架

Fig. 7 LLM-based framework for bearing fault diagnosis

此外,Pang 等[91]提出了 LLaMA-HFT 故障诊断框架,基于 LLaMA2[92]使用混合微调策略,即仅冻结底层模块的部分参数,对其他顶层模块使用 LoRA 参数高效微调。在帕德博恩大学(Paderborn University,PU)数据集[93]上,LLaMA-HFT 准确率和 F1 分数均优于小规模 Transformer 模型。

Peng 等[94]提出了一种名为 BearLLM 的轴承健康管理框架。首先,基于 CWRU 等公开数据集构建了一个多模态轴承健康管理的数据集,结合了机械振动信号和健康管理语料库;然后,使用该数据集训练故障分类网络,以统一振动信号表示为输入,通过监督学习提取振动信号特征并转换为词嵌入,作为 LLM 的输入。最后,采用 LoRA 参数高效微调策略对 Qwen2-1.5B[95]模型进行微调,以适应轴承健康管理任务。实验表明,BearLLM 在 9 个公共故障诊断数据集上,超越了专门针对单数据集设计的方法,展现出了优秀的泛化能力。

面对机械装备故障诊断中对高精度和智能化方法的需求,Lin 等[96]提出了一种基于多模态大语言模型的故障诊断框架 FD-LLM,通对 Vicuna-7B[97]模型进行模型对齐训练、模糊语义嵌入、可学习的提示嵌入,以及 LoRA 参数高效微调,使模型能够理解和处理时间序列数据。在 CWRU 数据集的轴承故障诊断实验中,FD-LLM 在准确率、精确率、和 F1 分数指标上均高于 98%,均高于使用 CNN[98]、小规模 Transformer 等方法。

2)部件 SOH 估计

Yunusoglu 等[99]提出了一种基于 LLM 框架的电池 SOH 估计方法,对预训练 BERT 模型输入层进行了定制,将电池循环数据和瞬间放电数据转化为 LLM 可理解的嵌入向量表示。同时,利用多头自注意力机制捕捉长期依赖关系,以及前馈神经网络提取复杂模式、位置编码保留序列信息;并且在架构顶部添加回归头,以输出 SOH 估计值。在包含 8 个电池单元不同循环条件老化数据的实验数据集中,该方法相比 SVM、CNN 等经典方法具有更低的平均绝对误差(mean absolute error,MAE)。

Feng 等[100]提出电池 SOH 估计框架 GPT4Battery,结合了 GPT-2 模型和测试时间训练[101]技术。在预训练阶段,利用从实验数据集中提取的电池电压-容量曲线特征对模型进行训练,通过物理引导的自监督学习和掩码重构损失函数,使模型能够利用电池的基本物理原理进行学习,优化了模型对这些特征的提取和理解能力。测试时间训练阶段,采用前缀提示适应策略[67-68],在输入数据前附加少量可学习的提示信息,微调模型参数,从而在降低计算成本的同时,提高了模型的适应性和泛化能力。在马里兰大学 CALCE 等多个锂离子电池数据集实验中,表现出了超过 LSTM、小规模 Transformer 等方法的均方误差(root mean square error,RMSE)、MAE 评价指标,并展现出优秀的跨数据集泛化能力。

Sun 等[102]提出了基于预训练时间序列基础模型的 TimeGPT[103]电池 SOH 估计框架,利用 143 个具有 6 种不同正极材料的锂离子电池的循环数据,使其能够捕获电池特有的健康退化模式的非线性和复杂性,从而实现高精度的 SOH 估计。在多个锂离子电池数据集上的实验结果表明,该方法在 RMSE、MAE 等评估标准上优于 LSTM、小规模 Transformer 等方法,为锂离子电池 SOH 估计提供了一种高效且高精度的新范式。

3.2 子系统级 PHM

子系统级 PHM 关注由多个相互作用的部件构成的功能单元的健康管理,例如液压系统、传动系统或冷却系统等。与部件级 PHM 的独立性不同,子系统级 PHM 需处理部件间的复杂耦合效应和故障传播路径。领域专用 LLM 在此层级的主要优势在于能够实现更准确的多模态融合数据分析。

子系统的数据来源除传感器外,还有系统级的性能参数、控制指令以及反映部件间相互影响的文本记录。LLM 能够深度整合这些异构数据,通过对文本描述中因果关系、功能依赖的理解,构建子系统级的逻辑关联[84]。例如,在识别子系统异常时,LLM 能够综合分析来自不同部件的多传感器数据[104]、系统运行参数以及相关操作日志与历史维护记录中的非结构化文本信息,从而更精确地诊断问题来源并识别潜在的故障传播链。这种融合分析能力提升了子系统故障诊断的准确性,并为优化维护策略提供了更全面的视角。

1)子系统 RUL 预测

Wang 等[105]在 RUL 预测任务中,较早地探索了 GPT 模型应用,提出了基于微调 GPT-2[106]预训练模型的方法,如图 8 所示。首先,该方法首先对输入时间序列进行分块处理,将连续数据转化为固定长度的子序列,以适配 GPT-2 的输入格式并降低数据复杂度;然后,数据嵌入层利用位置编码和值嵌入,将时间序列转换为 GPT-2 可理解的向量表示,输出层则将模型的输出维度转换为预测剩余寿命所需的单维度数值;最后,在微调阶段,GPT-2 模型中的大部分参数(包括自注意力和前馈神经网络层,以及残差和归一化层)被冻结,仅更新重构的数据嵌入层和输出层的参数。

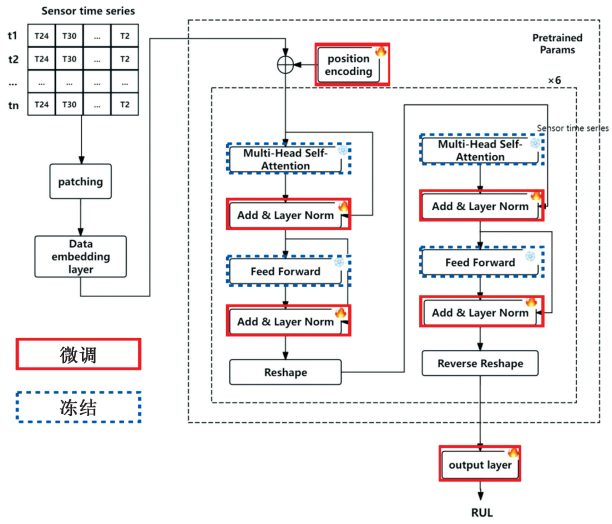

图 8　GPT 微调模型用于 RUL 预测的框架

Fig. 8　Fine-tuned GPT for RUL prediction framework

在来源于真实飞行数据且更加复杂的新数据集 N-CMAPSS[107]上的实验表明,该方法在多个子数据集上的 RMSE 结果优于领域适配 Transformer[108]和深度高斯过程[109]两种对照方法。

此外,Chen 等[110]提出了一种利用 GPT-2 模型 RUL 预测的框架:通过滑动窗口和线性嵌入层,将多维时间序列信号转换为 GPT-2 模型可理解的输入序列。在微调阶段,该框架使用 LoRA 高效微调方法,冻结了 GPT-2 模型的前 20 层,只对最后 4 层进行微调,从而在保留模型通用性能的同时实现了领域适配,并显著降低了计算资源消耗。微调后的 GPT-2 模型输出的隐藏状态通过一个线性回归头预测最终的 RUL。在 CMAPSS 数据集 FD002 和 FD004 两个相对复杂的子集实验中,该框架在 RMSE 和评分函数上的结果均优于多种基于 CNN、LSTM 等传统深度神经网络的方法。

为解决 RUL 预测中高维、噪声传感器数据带来的挑战,Chen 等[111]进一步提出了因果推理数据剪枝框架,通过因果推理识别与 RUL 具有稳健因果关系的传感器信号进行数据剪枝,并结合参数高效微调策略。在 N-CMAPSS 数据集上仅使用 10% 的数据进行微调,即可实现比现有方法低 26% 的 RMSE,并显著减少 90% 的训练时间,展示了高效且高精度的工业 RUL 预测潜力。

2)子系统故障诊断

工业智能化进程中,对系统整体进行状态监测是 PHM 领域的重要需求,以确保复杂生产或运行过程的连续性和高效性。与零件故障诊断不同,基于深度学习的系统故障诊断方法,通常需要针对特定模型进行训练,且不同故障特征存在重叠,易发生误诊,亟需鲁棒性更强的诊断策略[96]。

在 3.1 节第 1 部分中提及的 FD-LLM,凭借其多模态理解与逻辑推理能力,不仅适用于轴承等部件故障诊断任务,还可用于子系统级 PHM 任务。在来自于波音公司 CFM56-7B26、GE90-115B 等不同型号航空发动机的真实数据集[112]的实验中,准确率、精确率和 F1 分数上均超过 90%,显著高于使用 CNN、基础 Transformer 等方法。同时,该方法具备问答功能,通过提示学习和交互式推理,在用户输入工程数据后,系统能够生成诊断结果并提供维护建议,实现了人机交互式的故障诊断。

在供暖、通风和空调(heating, ventilation, and air conditioning, HVAC)系统故障诊断任务中,Zhang 等[113]提出了一种通过标签数据监督微调 LLM 的方法,通过设计 LLM 自纠正策略,自动基于标签数据生成微调数据集,并采用数据增强方法自适应更新微调数据集。实验表明,微调的 GPT-3.5 模型故障诊断准确率显著提高,与 GPT-4 模型相比平均诊断准确率高出 31.1%,且具备较强的泛化能力。

3.3　复杂系统级 PHM

复杂系统级 PHM 是最高层次的健康管理,涵盖了由多个子系统组成的大型、异构工业系统,如智能工厂、航空航天器或发电站。这一层面决策涉及多目标优化、资源调配和业务影响评估。LLM 在此层级的核心贡献在于提供完备、可靠的问答决策建议,从而实现对整个系统性能、安全和运营效率的宏观管理[114]。

面对海量的跨子系统数据和复杂的运营目标,LLM 能够将所有层次的健康信息与系统级的运行计划、环境因素、甚至供应链数据进行融合[115]。其强大的推理能力使得 LLM 能够识别系统层面的瓶颈、预测系统级故障对生产任务的影响、评估整体运营风险,并基于这些洞察,以自然语言的形式提供完备的决策支持,例如,推荐最优的生产调度、备件库存策略或跨部门协同维护方案。这些建议不仅考虑了技术可行性,还能权衡经济效益和安全风险,从而实现全局最优。

1)复杂系统故障诊断

Zheng 等[116]提出的用于复杂系统故障诊断的微调预训练 LLM 框架,通过数据集到自然语言提示的转换和监督式微调,对开源模型 LLaMa-2 和闭源模型 GPT-3.5-Turbo 进行 LoRA 参数高效微调。在高速列车制动系统真实数据集和田纳西-伊斯曼化工过程仿真数据集的实验表明,该方法在 F1 分数等标准下接近甚至超越 SVM、CNN 等基准方法,具有一定的应用前景。

在风力涡轮机故障诊断任务中,Men 等[117]提出了一种名为故障诊断事件思考知识图谱的框架,包含知识建模和知识推理。知识建模包含故障诊断事件知识图谱、抽象事件本体和物理世界的数字孪生,共同构成虚拟故障诊断数字孪生环境;知识推理以 GPT-3.5 为核心,在 VFD-DTEnv 中执行感知-思考-行动循环,实现对故障原因的推理和维护决策支持,从而有效提升了风力涡轮机故障诊断的准确性和可解释性。

在航空装配故障诊断任务中,Liu 等[118]提出了一种将知识图谱嵌入到 LLM 的联合知识增强框架,通过知识图谱中的图结构大数据对 ChatGLM-6B[90]模型进行前缀微调,并在推理阶段集成了子图生成-检索模块,增强模型专业知识的同时降低了计算量。在真实场景的数据集实验中,达到了 98.5% 的故障诊断准确率,显著高于未微调 LLM 的 41.5%。

2)复杂系统异常检测

Russell-Gilbert 等[119]基于 Llama 3-8B[120],提出了一种名为 AAD-LLM 的异常检测框架。通过统计过程控制技术对时间序列数据进行预处理,以确定正常行为的基线,并将时间序列数据分割成较小的窗口。再通过提示工程和文本模板将统计信息和领域知识整合到输入提示中,从而引导 LLM 进行推理。在塑料制造厂案例中,AAD-LLM 取得了 70.7% 的准确率和 77.0% 的 F1 分数。在 SKAB 异常检测基准数据集的实验中,AAD-LLM 的 F1 分数接近甚至超过了部分基于 LSTM 等方法,展现出了动态工业环境中异常检测的潜力。

此外,Russell-Gilbert[121]基于 Llama 3.1-8B[120],在 AAD-LLM 的基础上提出了 RAAD-LLM 异常检测框架,通过动态检索外部知识库中的相关信息,增强了模型对时间序列数据中异常的识别能力,以及在数据稀疏的工业环境中的异常检测能力。相比于 AAD-LLM,RAAD-LLM 在塑料制造案例中的准确率提升至 88.6%,F1 分数达到 91.9%。在 SKAB 数据集上,RAAD-LLM 的 F1 分数高于基于 LSTM 等对比方法,在动态工业环境中可作为更加准确且实用的异常检测方法。综上,PHM 领域大语言模型代表性研究对比如表 1 所示。

当前研究呈现出的特点为:

(1)方法路径清晰:研究从文本化传感器数据[89]向真正的多模态融合[96]演进,模型能力不断增强;

(2)偏重诊断与评估:大部分工作集中在故障诊断和 SOH/RUL 评估,而更前端的异常检测和更后端的维护决策优化与 LLM 的结合尚属蓝海;

(3)首选参数高效微调:LoRA 及其变体因低成本和高效性,是目前领域专用 LLM 构建的主流方法;

(4)初现端到端问答能力:部分研究[96]已展现出从数据输入到自然语言问答的端到端能力,预示着未来 PHM 系统将更具交互性和易用性。

综观 PHM 领域大语言模型的现有应用,其代表性研究目前仍受限于高质量、多模态标注数据的稀缺性,直接影响了模型的学习效果;模型对齐与泛化能力提升的复杂性,使得 LLM 难以在多样化的工业场景中实现普适性应用;以及对领域先验知识和专家经验的依赖性,限制了其自主学习和适应能力。同时,在复杂异构数据的深度融合与交互式推理方面,LLM 仍存在解释性与鲁棒性瓶颈,亟需提升跨任务、跨场景的可信赖、可泛化的解释能力。此外,硬件部署的处理实时性与计算效率也是亟待解决的关键问题,特别是在资源受限的边缘设备上,如何实现低延迟、高吞吐量的推理,是在工业 PHM 应用中面临的重大挑战。

4 趋势与挑战

随着大语言模型技术的不断发展,基于模型的 PHM 方法正在人工智能时代迎来前所未有的变革。大语言模型以及领域专用大语言模型,凭借其强大的特征学习能力、知识融合和推理能力,为 PHM 领域的各种任务带来新的解决方案。但将 PHM 领域专用大语言模型在实际应用中有效地落地,仍然面临诸多挑战。本章将从大模型轻量化、边缘端侧部署、广义复杂系统的角度展望 PHM 领域专用大语言模型的发展机遇,并分析其在实时性需求与计算资源限制、数据质量与成本、多物理场景仿真与可扩展性方面的挑战。本章将从大模型轻量化、边缘端侧部署、广义复杂系统的角度展望 PHM 领域专用大语言模型的发展机遇,并分析其在实时性需求与计算资源限制、数据质量与成本、多物理场景仿真与可扩展性方面的挑战。

表 1　PHM 领域大语言模型代表性研究对比

Table 1　Comparison of PHM domain-specific LLMs representative research

层次	任务	代表研究	LLM 骨架	关键创新	数据集	主要方法与效果
部件级	故障诊断	[89]	ChatGLM2-6B	振动信号文本化、QLoRA 微调	CWRU 等数据集跨数据集和条件应用	提出振动信号文本化和 QLoRA 微调,跨数据集学习后诊断准确率明显提升
		[91]	LLaMA2	基于 LoRA 的混合高效微调	PU 数据集	采用基于 LoRA 的混合微调策略,准确率和 *F*1 分数均优于小规模 Transformer 模型
		[94]	Qwen2-1. 5B	多模态数据集构建、LoRA 微调	CWRU 等数据集构成的多模态数据集	通过构建多模态数据集和 LoRA 微调,在 9 个公共数据集上展现优秀泛化能力
		[96]	Vicuna-7B	多模态对齐、时序数据处理	CWRU 数据集	采用多模态对齐和 LoRA 微调,准确率、精确率、*F*1 分数均高于 98%
	SOH 估计	[99]	BERT	定制输入层、回归预测	8 个电池单元循环老化数据集	通过定制输入层和回归预测头,平均绝对误差低于 SVM、CNN 等经典方法
		[100]	GPT-2	物理引导自监督学习、测试时间训练	CALCE 等数据集	提出物理引导自监督学习和测试时间训练,采用前缀提示适应策略微调,*RMSE*、*MAE* 优于对比方法,展现出优秀的跨数据集泛化能力
		[102]	TimeGPT	预训练时序模型捕获退化模式	143 个锂离子电池循环数据集	利用预训练时序模型捕获健康退化模式,*MAE*、*RMSE* 优于 *LSTM*、小规模 Transformer 模型
子系统级	RUL 预测	[105]	GPT-2	冻结主干网络、层微调	N-CMAPSS 数据集	通过时间序列分块并冻结主干网络微调输入输出层,*RMSE* 结果优于对照方法
		[110]	GPT-2	滑动窗口、LoRA 高效微调	C-MAPSS 数据集	采用滑动窗口和 LoRA 高效微调,*RMSE* 和评分函数优于多种传统深度神经网络
		[111]	GPT-2	因果推理数据剪枝	C-MAPSS 数据集	提出因果推理数据剪枝框架,仅用 10% 数据微调,实现 26% 更低 *RMSE*,训练时间减少 90%
	故障诊断	[96]	Vicuna-7B	多模态理解、问答推理	波音飞机发动机真实数据集	基于多模态理解和问答推理能力,实现人机交互式诊断,指标均超过 90%
		[113]	GPT-3. 5	监督微调、自纠正策略	HVAC 数据集	提出通过标签数据监督微调 LLM,诊断准确率比 GPT-4 平均高 31. 1%
复杂系统级	故障诊断	[116]	LLaMA-2	数据集到自然语言提示的转换	列车制动系统、化工过程数据集	采用数据集到自然语言提示转换和 LoRA 微调,*F*1 分数接近甚至超越 SVM、CNN 等基准方法
		[117]	GPT-3. 5	知识建模、循环推理	风力涡轮机故障数据集	提出基于知识图谱的故障诊断框架,通过 GPT-3. 5 循环推理,提升诊断准确性和可解释性
		[118]	ChatGLM-6B	知识图谱嵌入、前缀微调	航空装配数据集	通过知识图谱嵌入和前缀微调,诊断准确率达 98. 5%,显著高于未微调 LLM
	异常检测	[119]	Llama 3-8B	时间序列数据分割、提示工程	塑料制造厂数据集、SKAB 数据集	基于统计控制和提示工程,在塑料制造厂案例中准确率 70. 7%,*F*1 分数 77. 0%,SKAB 数据集 *F*1 分数接近或超过部分 LSTM 方法
		[121]	Llama 3. 1-8B	检索增强生成	塑料制造厂数据集、SKAB 数据集	采用检索增强生成技术,在塑料制造厂案例中准确率提升至 88. 6%,*F*1 分数达 91. 9%,SKAB 数据集 *F*1 分数高于对比方法

4.1 发展趋势

1)大模型轻量化

PHM 领域专用大语言模型在训练时依赖海量计算资源,云端推理也需高性能 GPU。然而,面对飞机跨洋飞行时发动机传感器异常等场景,通过不稳定的卫星通信链路进行云端处理会导致显著延迟,怠误故障响应最佳时机[122]。同时,军工、关键基础设施等领域对数据隐私要求极高,频繁与云端数据交互难以满足数据隐私和安全需求[123]。因此,模型剪枝、量化、知识蒸馏等轻量化技术正日益普及,以使大语言模型能在保持输出结果质量时,大幅削减参数规模和任务计算负载,从而为资源受限边缘端设备的高效本地数据采集、本地数据处理奠定基础。

2)边缘端侧部署

随着网络轻量化技术的进步,以及模型呈指数增长的能力密度[124],使得将 PHM 领域专用 LLM 部署到工业设备的边缘端正成为可能。边缘部署侧重于硬件层面的考量,包括优化成本、能效、实际算力以及硬件利用率。通过将经过轻量化处理的模型直接部署在靠近数据源的边缘设备上,可以有效降低对昂贵云端算力的依赖,减少数据传输延迟,实现更快的故障响应。

此外,本地化处理更好地满足了数据安全需求。这将进一步推动 PHM 技术在更多行业和场景中的应用,提升设备的可靠性和运行效率,显著降低运维成本。

3)广义复杂系统

当前 PHM 系统通常针对特定设备类型或行业定制开发,存在知识孤岛和复用性差的问题。大语言模型凭借其强大的通用语言理解、跨模态知识融合和泛化能力,为构建广义复杂系统提供了潜在途径[125]。故可引入世界基础模型(world foundation model)[126],通过人工智能构建交互系统的内部模拟环境,使其能够预测未来状态并理解复杂系统内部的因果关系。实现广义复杂系统建模的关键挑战,在于克服高精度多物理场仿真时的数据质量参差不齐与计算资源受限问题。

世界基础模型增强的领域专用大语言模型可从海量异构数据中自主学习并构建普适的 PHM 领域知识,使得跨设备知识迁移和少样本学习成为可能。无需针对每一种新型设备或特定故障模式从零开始构建模型,通过少量数据和简单的指令,即可实现模型的快速适配和性能优化。这将极大降低 PHM 系统的开发和部署门槛,加速 PHM 技术在不同工业场景中的普及和应用,最终实现知识的普惠化和经验的通用化[127]。

4.2 技术挑战

1)实时性需求与计算资源限制

工业 PHM 应用中,实时故障预警和紧急维护决策对系统的响应速度有着极高的要求。这意味着系统必须能够迅速完成数据的处理、分析和决策。然而,当前主流的预训练大语言模型由于其庞大的参数量和复杂的推理计算,在计算能力受限的边缘设备上部署和维护成本高昂。例如,GPT-3 模型拥有 1 750 亿参数[23],即使是经过优化的小型 LLM,其推理延迟在边缘设备上也难以满足工业 PHM 极高的实时性需求[128]。这种高推理延迟直接阻碍了 LLM 在工业 PHM 应用场景中进行实时故障预警和紧急维护决策。

同时,随着模型剪枝、量化等轻量化技术快速发展,在保证 PHM 高精度和鲁棒性的前提下,如何利用有限的计算资源实现 LLM 的低延迟、高吞吐量推理,仍然是严峻的技术挑战。此外,即使经过量化等优化,LLM 的持续推理仍然可能迅速耗尽电池,难以适应功耗敏感的工业场景[129]。考虑到绿色制造和可持续发展的社会背景,LLM 的训练和部署所伴随的巨大能源消耗也值得关注,对其能效比的优化将显得尤为重要[130]。

2)数据质量与成本

尽管大语言模型展现出从非结构化数据中抽取知识的强大能力,但工业 PHM 领域的数据依然面临着严峻的质量与成本挑战。首先,高质量的设备故障数据,特别是导致设备失效的早期退化数据,因设备制造商和运营商致力于避免故障发生,往往稀缺且获取难度大[131]。其次,特定故障模式(尤其是罕见故障或长尾事件)的数据样本极不均衡,LLM 在学习这些非典型现象时可能面临困难,影响其诊断和预测的准确性[22]。

其次,工业数据普遍存在多源异构、格式不统一、噪声多、缺失值普遍等问题,需要大量人力进行清洗、标注和预处理,带来了高昂的时间和经济成本[132]。如何有效利用少量带标签数据和大量无标签数据进行半监督或自监督学习,并降低数据获取和标注成本,是 LLM 在 PHM 领域发展的重要制约。

3)多物理场景仿真与可扩展性

工业设备通常涉及力学、热学、电学等多个物理场的复杂耦合,其仿真对数据质量和计算资源都有极高要求。然而,实际工业数据往往质量参差不齐,且获取受限。如何在有限的计算资源下,利用世界基础模型有效整合不同物理场的知识,并克服数据质量问题,实现对复杂系统行为的准确预测和因果关系的理解,是当前 PHM 领域亟待攻克的关键方法层面挑战[133]。

此外,确保仿真方法在面对海量异构工业系统时的可扩展性也至关重要。这要求模型不仅能够适应来自不同类型、不同规模设备的复杂数据流,还需要有效地进行知识迁移,从而在多样化的工业环境中实现真正意义上的普适性应用。现有 AI 模型在跨领域、跨设备部署时常面临数据分布差异和领域适应的难题,如何在不牺牲性

能的前提下,实现模型在广泛工业场景中的可靠推广,是当前 PHM 领域 LLM 的重要挑战[134]。

5 结　论

在大语言模型时代背景下,传统 PHM 方法在处理多模态数据、融合领域知识以及提升泛化能力等方面仍有挑战,而通用大语言模型又难以直接适配领域专业知识深度和预测精度的特定需求。为全面综述 PHM 领域专用大语言模型的研究现状,从 PHM 常见任务和指标、Transformer 结构和通用大语言模型出发,深入阐述领域专用大语言模型构建的领域知识注入方法、PHM 领域专用大语言模型的整体框架,并从部件级、子系统级、复杂系统级这 3 个层次针对故障诊断、健康状态估计、剩余使用寿命预测、异常检测等任务,分析了各类专用大语言模型的设计架构、构建策略及其在实际应用中展现的效果与潜力。

PHM 领域专用大语言模型的发展潜力巨大,在大模型轻量化、边缘端侧部署、广义复杂系统等多个方向具备发展机遇,但同时也面临实时性需求与计算资源限制、数据质量与成本、多物理场景仿真与可扩展性的挑战。随着硬件 AI 计算能力的不断增强和大语言模型性能的持续优化,PHM 领域专用大语言模型将朝着功能更全面、响应更迅速、应用更广泛的先进方向稳步进化。展望未来,PHM 领域大语言模型的发展,不仅是预测精度的提升,也是人机协作范式的革命:能够与工程师深度对话、共同诊断、协同决策的数字专家伙伴,从而深刻变革维护和管理复杂工业系统的方式。

参考文献

[1] VICHARE N M, PECHT M G. Prognostics and health management of electronics[J]. IEEE Transactions on Components and Packaging Technologies, 2006, 29(1): 222-229.

[2] 彭宇, 刘大同, 彭喜元. 故障预测与健康管理技术综述[J]. 电子测量与仪器学报, 2010, 24(1):1-9.
PENG Y, LIU D, PENG X Y. A review: Prognostics and health management [J]. Journal of Electronic Measurement and Instrumentation, 2010, 24(1):1-9.

[3] SHEPPARD J W, KAUFMAN M A, WILMER T J. IEEE Standards for prognostics and health management[J]. IEEE Aerospace and Electronic Systems Magazine, 2009, 24(9): 34-41.

[4] JARDINE A K S, LIN D M, BANJEVIC D. A review on machinery diagnostics and prognostics implementing condition-based maintenance [J]. Mechanical Systems and Signal Processing, 2006, 20(7): 1483-1510.

[5] CUESTA J, LETURIONDO U, VIDAL Y, et al. A review of prognostics and health management techniques in wind energy [J]. Reliability Engineering & System Safety, 2025, 260: 111004.

[6] GOEBEL K, SAHA B, SAXENA A, et al. Prognostics in battery health management[J]. IEEE Instrumentation & Measurement Magazine, 2008, 11(4): 33-40.

[7] SAXENA A, GOEBEL K, SIMON D, et al. Damage propagation modeling for aircraft engine run-to-failure simulation[C]. International Conference on Prognostics and Health Management, 2008: 1-9.

[8] HE K M, ZHANG X Y, REN SH Q, et al. Deep residual learning for image recognition[C]. 29th IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[9] SATEESH BABU G, ZHAO P L, LI X L. Deep convolutional neural network based regression approach for estimation of remaining useful life [C]. Database Systems for Advanced Applications, 2016: 214-228.

[10] 伍济钢, 文港, 杨康. 改进一维卷积神经网络的航空发动机故障诊断方法[J]. 电子测量与仪器学报, 2023, 37(3): 179-186.
WU J G, WEN G, YANG K. Improved one-dimensional convolutional neural network for aero-engine fault diagnosis[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(3): 179-186.

[11] WANG H T, WANG R H, YANG J. Unsupervised domain adaptive migration learning-based approach to bearing remaining useful life prediction [J]. Instrumentation, 2025, 12(1): 37-47.

[12] ZHENG SH, RISTOVSKI K, FARAHAT A, et al. Long short-term memory network for remaining useful life estimation[C]. 2017 IEEE International Conference on Prognostics and Health Management, 2017: 88-95.

[13] ZHANG Y ZH, XIONG R, HE H W, et al. Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries[J]. IEEE Transactions on Vehicular Technology, 2018, 67(7): 5695-5705.

[14] 靳乐怡, 王珏, 叶红军, 等. 基于 LSTM 的卫星导航系统服务性能监测方法研究[J]. 电子测量技术, 2024, 45(22): 149-156.
JIN L Y, WANG J, YE H J, et al. Research on LSTM-

based performance monitoring techniques for satellite navigation systems services[J]. Electronic Measurement Technology, 2024, 45(22): 149-156.

[15] YAN R Q, ZHOU ZH, SHANG Z G, et al. Knowledge driven machine learning towards interpretable intelligent prognostics and health management: Review and case study[J]. Chinese Journal of Mechanical Engineering, 2025, 38(1): 31-61.

[16] TAO L F, LI SH Y, LIU H F, et al. An outline of prognostics and health management large model: Concepts, paradigms, and challenges[J]. Mechanical Systems and Signal Processing, 2025, 232: 112683.

[17] REINSEL D, GANTZ J, RYDNING J. Data age 2025: The datasphere and data-readiness from edge to core[R/OL]. International Data Corporation, 2018.

[18] 中国大数据产业发展指数报告(2024 版)[R]. 北京大数据研究院, 2024.
China big data industry development index report(2024 Edition)[R]. Beijing Institute of Big Data Research, 2024.

[19] ZHONG K, JACKSON T, WEST A, et al. Natural language processing approaches in industrial maintenance: A systematic literature review[J]. Procedia Computer Science, 2024, 232: 2082-2097.

[20] LI X, DING Q, SUN J Q. Remaining useful life estimation in prognostics using deep convolution neural networks[J]. Reliability Engineering & System Safety, 2018, 172: 1-11.

[21] ZHANG CH, LIM P, QIN A K, et al. Multiobjective deep belief networks ensemble for remaining useful life estimation in prognostics[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2306-2318.

[22] REN ZH J, LIN T T, FENG K, et al. A systematic review on imbalanced learning methods in intelligent fault diagnosis[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 3246470.

[23] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[24] 夏润泽, 李丕绩. ChatGPT 大模型技术发展与应用[J]. 数据采集与处理, 2023, 38(5): 1017-1034.
XIA R Z, LI P J. Large language model ChatGPT: Evolution and application[J]. Journal of Data Acquisition and Processing, 2023, 38(5): 1017-1034.

[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010.

[26] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.

[27] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[J]. ArXiv preprint arXiv:2108.07258, 2021.

[28] LUKENS S, MCCABE L H, GEN J, et al. Large language model agents as prognostics and health management copilots[C]. Annual Conference of the PHM Society, 2024, 16(1):3906.

[29] DENG H, NAMOANO B, ZHENG B, et al. From prediction to prescription: Large language model agent for context-aware maintenance decision support[C]. PHM Society European Conference, 2024, 8(1): 478-487.

[30] YU T Y, TANG J Y, YU Q Y, et al. Large language models for PHM: A review of optimization techniques and applications[J]. Autonomous Intelligent Systems, 2025, 5(1): 1-14.

[31] LI Y F, WANG H, SUN M X. ChatGPT-like large-scale foundation models for prognostics and health management: A survey and roadmaps[J]. Reliability Engineering & System Safety, 2024, 243: 109850.

[32] WEN L, LI X Y, GAO L, et al. A new convolutional neural network-based data-driven fault diagnosis method[J]. IEEE Transactions on Industrial Electronics, 2018, 65(7): 5990-5998.

[33] 李军宁, 罗文广, 陈武阁. 面向振动信号的滚动轴承故障诊断算法综述[J]. 西安工业大学学报, 2022, 42(2): 105-122.
LI J N, LUO W G, CHEN W G. Overview of algorithms for rolling bearing fault diagnosis based on vibration signal[J]. Journal of Xi′an Technological University, 2022, 42(2): 105-122.

[34] YANG S J, ZHANG C P, JIANG J CH, et al. Review on state-of-health of lithium-ion batteries: Characterizations, estimations and applications[J]. Journal of Cleaner Production, 2021, 314(10): 128015.

[35] 胡晓亚, 郭永芳, 张若可. 锂离子电池健康状态估计

方法研究综述[J]. 电源学报, 2022, 20(1): 126-133.
HU X Y, GUO Y F, ZHANG R K. Review of state-of-health estimation methods for lithium-ion battery[J]. Journal of Power Supply, 2022, 20(1): 126-133.

[36] SEVERSON K A, ATTIA P M, JIN N, et al. Data-driven prediction of battery cycle life before capacity degradation[J]. Nature Energy, 2019, 4(5): 383-391.

[37] WANG Y D, ZHAO Y F, ADDEPALLI S. Remaining useful life prediction using deep learning approaches: A review[J]. Procedia Manufacturing, 2020, 49: 81-88.

[38] 孙见忠, 王卓健, 闫洪胜, 等. 航空预测性维修研究进展[J]. 航空学报, 2025, 46(7): 6-29.
SUN J ZH, WANG ZH J, YAN H SH, et al. Research advances in aircraft predictive maintenance[J]. Acta Aeronautica et Astronautica Sinica, 2025, 46(7): 6-29.

[39] LIU Y, FREDERICK D K, DECASTRO J A, et al. User's guide for the commercial modular aero-propulsion system simulation (C-MAPSS): Version 2[R/OL]. NASA, 2012.

[40] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: A survey[J]. ACM Computing Surveys, 2009, 41(3):1541882.

[41] 彭喜元, 庞景月, 彭宇, 等. 航天器遥测数据异常检测综述[J]. 仪器仪表学报, 2016, 37(9): 1929-1945.
PENG X Y, PANG J Y, PENG Y, et al. Review on anomaly detection of spacecraft telemetry data[J]. Chinese Journal of Scientific Instrument, 2016, 37(9): 1929-1945.

[42] LAN ZH ZH, CHEN M D, GOODMAN S, et al. ALBERT: A lite bert for self-supervised learning of language representations[C]. International Conference on Learning Representations, 2020.

[43] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[R/OL]. OpenAI, 2018.

[44] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models[J]. ArXiv preprint arXiv:2302.13971, 2023.

[45] LEWIS M, LIU Y H, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]. 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7871-7880.

[46] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(140): 1-67.

[47] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners[C]. International Conference on Learning Representations, 2022.

[48] CHRISTIANO P, LEIKE J, BROWN T B, et al. Deep reinforcement learning from human preferences[J]. Advances in Neural Information Processing Systems, 2017, 30: 4302-4310.

[49] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[50] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model[J]. Advances in Neural Information Processing Systems, 2023, 36: 53728-53741.

[51] LEE J, YOON W, KIM S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[52] CHALKIDIS I, FERGADIOTIS M, MALAKASIOTIS P, et al. LEGAL-BERT: The Muppets straight out of law school[C]. Findings of the Association for Computational Linguistics: EMNLP 2020, 2020: 2898-2904.

[53] LU W, LUU R K, BUEHLER M J. Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities[J]. NPJ Computational Materials, 2025, 11(1).

[54] GURURANGAN S, MARASOVIĆ A, SWAYAMDIPTA S, et al. Don't stop pretraining: Adapt language models to domains and tasks[C]. 58th Annual Meeting of the Association for Computational Linguistics, 2020: 8342-8360.

[55] HUANG L, YU W J, MA W T, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 3703155.

[56] HEESCH R, EILERMANN S, WINDMANN A, et al. Evaluating large language models for real-world engineering tasks [J]. ArXiv preprint arXiv: 2505.13484, 2025.

[57] LING CH, ZHAO X J, LU J Y, et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey [J]. ACM Computing Surveys, 2025, 58(3): 3703155.

[58] ZHENG J W, HONG H H, LIU F Y, et al. DragFT: Adapting large language models with dictionary and retrieval augmented fine-tuning for domain-specific machine translation[J]. ArXiv preprint arXiv:2402.15061, 2024.

[59] HU L M, LIU Z Y, ZHAO Z W, et al. A survey of knowledge enhanced pre-trained language models [J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(4): 1413-1430.

[60] JI SH X, PAN SH R, CAMBRIA E, et al. A survey on knowledge graphs: Representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494-514.

[61] XUE B C, ZOU L. Knowledge graph quality management: A comprehensive survey [J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(5): 4969-4988.

[62] SONG Z R, YAN B, LIU Y H, et al. Injecting domain-specific knowledge into large language models: A comprehensive survey [J]. ArXiv preprint arXiv: 2502.10708, 2025.

[63] WU CH Y, LIN W X, ZHANG X M, et al. PMC-LLaMA: Toward building open-source language models for medicine [J]. Journal of the American Medical Informatics Association, 2024, 31(9): 1833-1843.

[64] SINGHAL K, TU T, GOTTWEIS J, et al. Toward expert-level medical question answering with large language models[J]. Nature Medicine, 2025, 31(3): 943-950.

[65] LIALIN V, DESHPANDE V, YAO X, et al. Scaling down to scale up: A guide to parameter-efficient fine-tuning[J]. ArXiv preprint arXiv:2303.15647, 2023.

[66] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP [C]. 36th International Conference on Machine Learning, 2019, 97: 4944-4956.

[67] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning [C]. Empirical Methods in Natural Language Processing. 2021: 3045-3059.

[68] LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[C]. 59th Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, 2021: 4582-4597.

[69] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[J]. ArXiv preprint arXiv: 2106.09685, 2021.

[70] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. QLoRA: Efficient finetuning of quantized LLMs [J]. Advances in Neural Information Processing Systems, 2023, 36: 10088-10115.

[71] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[72] WANG Q, MAO ZH D, WANG B, et al. Knowledge graph embedding: A survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.

[73] ZHAO Y, ZHANG J J, ZHOU Y, et al. Knowledge graphs enhanced neural machine translation [C]. 29th International Joint Conference on Artificial Intelligence, 2021: 4019-4025.

[74] VEDULA N, PARTHASARATHY S. FACE-KEG: Fact checking explained using knowledge graphs [C]. ACM International Conference on Web Search and Data Mining, 2021: 526-534.

[75] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620(7972): 172-180.

[76] QIN Y F, FU X, ZHANG ZH X, et al. Key technologies for application of coal mine domain large model based on LoRA fine-tuning and RAG fusion[J]. Journal of Mine Automation, 2025, 51(8): 34-42, 50.

[77] CALONGE D S, SMAIL L. Optimizing retrieval-augmented generation (RAG) for colloquial cantonese: A lora-based systematic review[J]. ArXiv preprint arXiv: 2508.08610, 2025.

[78] CHKIRBENE Z, HAMILA R, GOUISSEM A, et al. Large language models (LLM) in industry: A survey of

applications, challenges, and trends[C]. 2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT, 2024: 229-234.

[79] REN ZH G, FANG F ZH, YAN N, et al. State of the art in defect detection based on machine vision[J]. International Journal of Precision Engineering and Manufacturing-Green Technology, 2022, 9(2): 661-691.

[80] CLANCY R, ILYAS I F, LIN J. Scalable knowledge graph construction from text collections[C]. Workshop on Fact Extraction and VERification, 2019: 39-46.

[81] LI S T, TANG H. Multimodal alignment and fusion: A survey[J]. ArXiv preprint arXiv:2411.17040, 2025.

[82] NAGRANI A, YANG SH, ARNAB A, et al. Attention bottlenecks for multimodal fusion[J]. Advances in Neural Information Processing Systems, 2021, 34: 14200-14213.

[83] AN J, LEE J, LEE J, et al. Towards LLM-centric multimodal fusion: A survey on integration strategies and techniques[J]. ArXiv preprint arXiv: 2506.04788, 2025.

[84] LAHAT D, ADALI T, JUTTEN C. Multimodal data fusion: An overview of methods, challenges, and prospects[J]. Proceedings of the IEEE, 2015, 103(9): 1449-1477.

[85] RIBEIRO M T, SINGH S, GUESTRIN C. 'Why should I trust you?' Explaining the predictions of any classifier[C]. Conference on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2016, 2016: 97-101.

[86] ZHANG C, CHEN J X, LI J T, et al. Large language models for human-robot interaction: A review[J]. Biomimetic Intelligence and Robotics, 2023, 3(4): 100131.

[87] OGUNFOWORA O, NAJJARAN H. Reinforcement and deep reinforcement learning-based solutions for machine maintenance planning, scheduling policies, and optimization[J]. Journal of Manufacturing Systems, 2023, 70: 244-263.

[88] GU W Q, DAO T. Mamba: Linear-time sequence modeling with selective state spaces[J]. ArXiv preprint arXiv: 2312.00752,2023.

[89] TAO L F, LIU H F, NING G AO, et al. LLM-based framework for bearing fault diagnosis[J]. Mechanical Systems and Signal Processing, 2025, 224: 112127.

[90] ZENG AO H, XU B, WANG B W, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools[J]. ArXiv preprint arXiv: 2406.12793, 2024.

[91] PANG ZH D, ZHANG H, LI T. Hybrid fine-tuning in large language model learning for machinery fault diagnosis[C]. 2024 IEEE 22nd International Conference on Industrial Informatics. 2024: 1-6.

[92] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. ArXiv preprint arXiv:2307.09288, 2023.

[93] LESSMEIER C, KIMOTHO J K, ZIMMER D, et al. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification[C]. PHM Society European Conference. 2016, 3(1):1577.

[94] PENG H T, LIU J W, DU J S, et al. BearLLM: A prior knowledge-enhanced bearing health management framework with unified vibration signal representation[C]. AAAI Conference on Artificial Intelligence, 2025, 39(19): 19866-19874.

[95] BAI J Z, BAI SH, CHU Y F, et al. Qwen technical report[J]. ArXiv preprint arXiv:2309.16609, 2023.

[96] LIN L, ZHANG S H, FU S, et al. FD-LLM: Large language model for fault diagnosis of complex equipment[J]. Advanced Engineering Informatics, 2025, 65: 103208.

[97] JI B. VicunaNER: Zero/Few-shot named entity recognition using vicuna[J]. ArXiv preprint arXiv: 2305.03253, 2023.

[98] RUAN D W, WANG J, YAN J Q, et al. CNN parameter design based on fault signal analysis and its application in bearing fault diagnosis[J]. Advanced Engineering Informatics, 2023, 55: 101877.

[99] YUNUSOGLU A, LE D, ISIK M, et al. Battery state of health estimation using LLM framework[C]. 2025 26th International Symposium on Quality Electronic Design, 2025: 1-8.

[100] FENG Y Y, HU G SH, LI X D, et al. Adapting amidst degradation: Cross domain li-ion battery health estimation via physics-guided test-time training[J]. ArXiv preprint

arXiv:2402. 00068, 2024.

[101] SUN Y, WANG X L, LIU ZH, et al. Test-time training with self-supervision for generalization under distribution shifts[C]. 37th International Conference on Machine Learning, 2020, 119: 9229-9248.

[102] SUN W J, WU C, XIE CH D, et al. Fine-tuning enables state of health estimation for lithium-ion batteries via a time series foundation model[J]. Energy, 2025, 318: 134177.

[103] GARZA A, CHALLU C, MERGENTHALER-CANSECO M. TimeGPT-1[J]. ArXiv preprint arXiv:2310. 03589, 2023.

[104] KIBRETE F, WOLDEMICHAEL D E, GEBREMEDHEN H S. Multi-sensor data fusion in intelligent fault diagnosis of rotating machines: A comprehensive review[J]. Measurement, 2024, 232: 114658.

[105] WANG P W, NIU SH ZH, CUI H L, et al. GPT-based equipment remaining useful life prediction[J]. ACM Turing Award Celebration Conference, 2024: 159-164.

[106] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[R/OL]. OpenAI, 2019.

[107] ARIAS CHAO M, KULKARNI C, GOEBEL K, et al. Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics[J]. Data, 2021, 6(1):6010005.

[108] LI X Y, LI J J, ZUO L, et al. Domain adaptive remaining useful life prediction with transformer[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 3200667.

[109] ZENG J Q, LIANG ZH L. A deep gaussian process approach for predictive maintenance[J]. IEEE Transactions on Reliability, 2023, 72(3): 916-933.

[110] CHEN Y, LIU CH. Remaining useful life prediction: A study on multidimensional industrial signal processing and efficient transfer learning based on large language models[J]. ArXiv preprint arXiv:2410. 03134, 2024.

[111] CHEN Y, LIU CH. Causal inference based transfer learning with LLMs: An efficient framework for industrial RUL prediction[J]. ArXiv preprint arXiv:2503. 17686, 2025.

[112] LIN L, HE W H, FU S, et al. Novel aeroengine fault diagnosis method based on feature amplification[J]. Engineering Applications of Artificial Intelligence, 2023, 122: 106093.

[113] ZHANG J, ZHANG CH B, LIU J, et al. Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning[J]. Applied Energy, 2025, 377: 124378.

[114] LI SH, PUIG X, PAXTON C, et al. Pre-trained language models for interactive decision-making[J]. Advances in Neural Information Processing Systems, 2022, 35: 31199-31212.

[115] SHAHIN K I, LAZAROVA-MOLNAR S. Digital twins in prognostics and health management: A review of diagnostic, remaining useful life, and predictive maintenance applications[C]. 2024 8th International Conference on System Reliability and Safety, 2024: 720-729.

[116] ZHENG SH W, PAN K, LIU J, et al. Empirical study on fine-tuning pre-trained large language models for fault diagnosis of complex systems[J]. Reliability Engineering & System Safety, 2024, 252: 110382.

[117] MEN CH H, HAN Y, WANG P, et al. The interpretable reasoning and intelligent decision-making based on event knowledge graph with LLMs in fault diagnosis scenarios[J]. IEEE Transactions on Instrumentation and Measurement, 2025, 74: 3550999.

[118] LIU P F, QIAN L, ZHAO X W, et al. Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly[J]. IEEE Transactions on Industrial Informatics, 2024, 20(6): 8160-8169.

[119] RUSSELL-GILBERT A, SOMMERS A, THOMPSON A, et al. AAD-LLM: Adaptive anomaly detection using large language models[C]. 2024 IEEE International Conference on Big Data, 2024: 4194-4203.

[120] GRATTAFIORI A, DUBEY A, JAUHRI A, et al. The llama 3 herd of models[J]. ArXiv preprint arXiv: 2407. 21783, 2024.

[121] RUSSELL-GILBERT A. RAAD-LLM: Adaptive anomaly detection using LLMs and RAG integration[D]. Mississippi State: Mississippi State University, 2025.

[122] XUE H H, ZHOU H, HUANG B, et al. Edge computing for internet of things: A survey[C]. 2020 International Conferences on Internet of Things and IEEE Green Computing and Communications, 2020: 755-760.

[123] ALWARAFY A, AL-THELAYA K A, ABDALLAH M,

et al. A survey on security and privacy issues in edge-computing-assisted internet of things[J]. IEEE Internet of Things Journal, 2021, 8(6): 4004-4022.

[124] XIAO CH J, CAI J, ZHAO W L, et al. Densing law of LLMs[J]. Nature Machine Intelligence, 2025.

[125] ZHANG H T, SEMUJJU S D, WANG ZH CH, et al. Large scale foundation models for intelligent manufacturing applications: A survey[J]. Journal of Intelligent Manufacturing, 2025.

[126] HA D, SCHMIDHUBER J. Recurrent world models facilitate policy evolution [J]. Advances in Neural Information Processing Systems, 2018, 31: 2455-2467.

[127] REN L, WANG H T, DONG J B, et al. Industrial foundation model[J]. IEEE Transactions on Cybernetics, 2025, 55(5): 2286-2301.

[128] PANG G S, SHEN CH H, CAO L B, et al. Deep learning for anomaly detection: A review [J]. ACM Computing Survey, 2021, 54(2):3439950.

[129] FRIHA O, FERRAG M A, KANTARCI B, et al. LLM-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness[J]. IEEE Open Journal of the Communications Society, 2024, 5: 5799-5856.

[130] STRUBELL E, GANESH A, MCCALLUM A. Energy and policy considerations for modern deep learning research [C]. 34th AAAI Conference on Artificial Intelligence. 2020, 34(9): 13693-13696.

[131] LI CH J, LI SH B, FENG Y X, et al. Small data challenges for intelligent prognostics and health management: A review [J]. Artificial Intelligence Review, 2024, 57(8): 1-52.

[132] ESCOBAR C A, MCGOVERN M E, MORALES-MENENDEZ R. Quality 4.0: A review of big data challenges in manufacturing[J]. Journal of Intelligent Manufacturing, 2021, 32(8): 2319-2334.

[133] ATTARAN M, CELIK B G. Digital twin: Benefits, use cases, challenges, and opportunities [J]. Decision Analytics Journal, 2023, 6: 100165.

[134] PHUYAL S, BISTA D, BISTA R. Challenges, opportunities and future directions of smart manufacturing: A state of art review[J]. Sustainable Futures, 2020, 2: 100023.

作者简介

彭宇,2004 年于哈尔滨工业大学获得博士学位,现为哈尔滨工业大学教授、博士生导师,主要研究方向为虚拟仪器和自动测试、故障预测与健康管理、可重构计算等。
E-mail: pengyu@hit.edu.cn

Peng Yu received his Ph.D. degree from Harbin Institute of Technology in 2004. He is currently a professor and a Ph.D. advisor at Harbin Institute of Technology. His main research interests include virtual instruments and automatic test technologies, prognostics and health management, and reconfigurable computing, etc.

季拓,2023 年于哈尔滨工业大学获得学士学位,现为哈尔滨工业大学电子与信息工程学院测控工程系在读硕士研究生,主要研究方向为神经网络与深度学习、故障预测与健康管理、大语言模型。
E-mail: 24S105138@stu.hit.edu.cn

Ji Tuo received his B.Sc. degree from Harbin Institute of Technology in 2023. He is currently a master student in the Department of Measurement and Control Engineering, School of Electronics and Information Engineering, Harbin Institute of Technology. His main research interests include neural networks and deep learning, prognostics and health management, and large language models.

郭楚亮(通信作者),2019 年于哈尔滨工业大学获得学士学位,2024 年于浙江大学获得博士学位,现为哈尔滨工业大学副研究员,主要研究方向为 FPGA 加速、边缘 AI 软硬协同设计。
E-mail: chuliang007@hit.edu.cn

Guo Chuliang (Corresponding author) received his B.Sc. degree from Harbin Institute of Technology in 2019, and Ph.D degree from Zhejiang University in 2024. He is currently a research associate professor at Harbin Institute of Technology. His main research interests include FPGA-based acceleration, and co-design for edge AI applications.