

工业大模型应用技术：综述与展望

秦小林^{1,2}, 王宇^{1,2}, 李弟诚^{1,2}, 徐海文³

(1. 中国科学院成都计算机应用研究所, 四川 成都 610213; 2. 中国科学院大学计算机科学与技术学院, 北京 101408; 3. 中国民用航空飞行学院理学院, 四川 广汉 618307)

摘要: 工业大模型作为新一代人工智能与工业场景深度融合技术, 为重塑制造业智能化体系提供了强有力的技术支撑。随着第四次工业革命的推进, 具备优秀数据处理能力、跨模态融合特性和智能决策的工业大模型已成为推动工业智改数转和工业智能化的核心技术。然而, 其在行业中的应用仍处于初步阶段, 面临数据偏差、行业知识匮乏和算力成本高等挑战, 如何快速落地与广泛应用成为亟待解决的问题。现有综述多聚焦于大模型的宏观理论架构、训练方法及典型行业案例, 缺乏对工业具体场景下关键技术与应用模式的系统性探讨, 难以为典型应用提供全面指导。通过归纳总结工业大模型关键技术和国内外典型应用进展, 进一步阐述了工业大模型技术在实际应用中面临的挑战。同时, 对工业大模型架构体系进行了分析, 为后续研究提供了理论基础、关键技术和典型应用的参考借鉴。

关键词: 工业大模型; 工业智改数转; 工业智能; 工业大模型架构体系; 数据偏差

DOI: 10.11907/rjtk.251555

扫描二维码阅读全文:



中图分类号: TP18; F424

文献标识码: A

文章编号: 1672-7800(2025)010-0001-14

Survey and Perspective on Application Technologies of Industrial Foundation Models

QIN Xiaolin^{1,2}, WANG Yu^{1,2}, LI Dicheng^{1,2}, XU Haiwen³

(1. Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610213, China; 2. School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China; 3. Faculty of Science, Civil Aviation Flight University of China, Guanghan 618307, China)

Abstract: Industrial foundation models (IFMs) represent a new generation of artificial intelligence technologies deeply integrated with industrial scenarios, providing powerful technical support for reshaping the intelligent manufacturing ecosystem. With the advancement of the Fourth Industrial Revolution, IFMs characterized by excellent data-processing capabilities, cross-modal fusion features, and intelligent decision-making—have become the core technologies driving industrial digital-intelligence transformation and industrial intelligence. However, their deployment within industry remains at an early stage and faces challenges such as data bias, lack of domain-specific knowledge, and high computational costs; facilitating rapid adoption and widespread application has thus become an urgent problem to solve. Existing surveys tend to focus on the macro-level theoretical architectures, training methodologies, and representative industry case studies of large models, but they lack a systematic examination of the key technologies and application patterns in concrete industrial scenarios, making it difficult to offer comprehensive guidance for typical use cases. By summarizing the key enabling technologies of IFMs and reviewing major domestic and international application advances, this paper further elucidates the challenges encountered in practical deployments. We have analyzed the architectural framework of industrial foundation models. These models provide a theoretical foundation, essential technologies, and representative applications as valuable references for future research.

Key Words: Industrial Foundation Models (IFMs); industrial digital-intelligence transformation; industrial intelligence; architectural framework of IFMs; data bias

收稿日期: 2025-08-14

基金项目: 国家重点研发计划项目(2023YFB3308601); 四川省委组织部人才专项(2024)

作者简介: 秦小林(1980-), 男, 博士, CCF会员, 中国科学院成都计算机应用研究所研究员、博士生导师, 研究方向为自动推理、空间智能、行业大模型; 王宇(1993-), 男, 中国科学院成都计算机应用研究所博士研究生, 研究方向为自然语言处理、行业大模型; 李弟诚(2003-), 男, 中国科学院成都计算机应用研究所博士研究生, 研究方向为自然语言处理、行业大模型; 徐海文(1978-), 男, 博士, 中国民用航空飞行学院理学院教授、硕士生导师, 研究方向为优化理论与算法、交通运输规划与管理。本文通讯作者: 秦小林。

0 引言

在第四次工业革命浪潮的推动下,信息技术与制造业深度融合,大模型(Large Language Models, LLMs)为工业制造带来了机遇和挑战。制造业作为“立国之本、强国之基”,已被纳入国家重大战略规划。2015年《国务院关于深化制造业与互联网融合的指导意见》(“中国制造2025”)明确提出以新一代信息技术与制造业深度融合为主线,加快推进智能制造^[1]。随着云一边一端算力的普及和工业现场多模态数据(时序、图像、文本等)激增,工业系统从数字化、网络化的模式转入智能化新阶段^[2]。在这一背景下,工业大模型凭借优秀的数据处理和推理能力,成为挖掘工业数据潜在价值的关键技术,能够实现制造全过程精细化管控与优化,显著提升效率、降低成本、提高质量。智能制造已成为“工业5.0”全球制造发展的主流趋势,工业大模型作为核心技术之一,将驱动产业深刻变革,为制造业的跨越式发展注入强劲动力。

近年来,大模型凭借海量训练数据、超大参数和算力,实现了传统AI无法达到的生成与泛化能力^[3]。以自然语言处理为例,GPT^[3-4]、BERT^[5]等模型在文本理解和生成、复杂对话等任务上表现突出^[5];在计算机视觉领域,类似架构也能够生成高保真图像,实现精准识别。大模型已经成为通向人工智能之路的里程碑技术,如2022年底OpenAI推出的ChatGPT,引发了社会的广泛关注^[6-7]。过去AI系统往往针对特定任务设计专用算法,但大模型突破了这一局限,实现了跨任务、跨场景的学习与推理能力,使得同一模型可以服务于不同应用,大幅度提高了算法复用效率,为多领域融合人工智能技术的研究奠定了基础。

大模型技术的核心价值在于其通用性和复用性。2017年,Transformer架构的提出为后续大模型的发展奠定了基础,其注意力机制增强了序列数据处理能力^[8]。基于Transformer的预训练语言模型如BERT^[5]和GPT系列^[3],在海量无监督数据上预训练后,通过少量标注数据微调即可适应各种下游任务,在翻译、问答、摘要等任务上表现优异。这种“基础模型”范式使得同一套参数和结构可以在不同领域复用,并可通过少样本学习或零样本学习在新任务上展现出强大的性能^[7]。现代大模型已经开始支持多模态输入与输出,通过扩展编码器/解码器和交叉注意力等手段,能够同时处理文本、图像、语音等多种数据类型,实现跨模态的关联理解与生成,拓宽了AI在复杂任务中的应用范围。

在时代的浪潮下,我国高度重视制造业智能化发展和人工智能技术创新。制造业数字化智能化被写入《中华人民共和国国民经济和社会发展第十四个五年规划和2035

年远景目标纲要》(简称“十四五”规划)等重要纲要,工业互联网、大数据、云计算等新一代信息通信技术正在深入融入制造业。德国在2013年首次提出“工业4.0”概念,揭开以智能制造为核心的第四次工业革命序幕^[9]。麦肯锡等咨询机构多次指出,数字化和智能化是提升制造业价值创造能力的决定性杠杆^[10]。欧盟委员会于2021年提出欧盟“工业5.0”的概念,强调其核心在于推动以人为本、可持续和韧性的产业系统建设。为迎合工业革命大势,中国工业5.0于2023年提出,其中心思想在于提高国内工业的核心竞争力,为中国新一轮工业革命中占领先机。在这一全球趋势下,传统工业已无法处理工业现场呈爆炸式增长的多模态数据(如时序数据、视觉数据和文本数据等),传统工业信息化势在必行。为此,业界在通用大模型的基础上提出了工业大模型(Industrial Foundation Models, IFMs),该技术通过跨模态自监督预训练、物理机理知识嵌入和专家模型路由等技术,在统一的参数空间内兼顾感知、预测、决策任务,可利用通用或领域知识指导机器完成包括数据分析、预测和决策支持等各类任务。工业大模型技术被视为推动“工业4.0”向“工业5.0”过渡的关键技术支撑。工业大模型强调大规模、多模态和高可靠输出,具有覆盖产品研发、制造、运维等全生命周期应用的潜力。然而,工业场景的独特性也给大模型应用带来了严峻挑战,数据质量与可靠性问题、行业知识与语料库不足、算力成本与技术门槛等挑战严重制约了工业大模型的进一步发展和广泛应用。

现有关于工业大模型的综述缺乏对关键技术和工业场景的概述。如表1所示,文献[11-14]深入分析了通用大模型技术的发展,但缺乏对工业大模型的介绍;文献[15-21]分别从高端新材料、智能制造、电力系统、智能油气等视角解读了工业大模型应用情况;文献[22]提出了工业大模型的五层架构,但未对工业大模型技术和典型应用进行系统性归纳与总结。本文聚焦于最新工业大模型进展,系统归纳了国内外工业大模型的前沿应用,表2展示了国内外12家企业工业大模型应用对比情况。在对当前工业大模型进行研究的基础上,将进一步聚焦工业领域大模型应用现状和技术研究进展,并对未来研究方向进行了展望。

Table 1 Comparative review of IFMs

表1 工业大模型综述比较

文献	大模型通用能力	高端新材料	智能制造	电力系统	智能油气	工业大模型基础架构	工业大模型应用
文献[11-14]	√	×	×	×	×	×	×
文献[15]	√	√	×	×	×	×	×
文献[16-19]	√	×	×	√	×	×	×
文献[20]	√	×	√	×	×	×	×
文献[21]	√	×	×	×	√	×	×
文献[22]	√	×	√	√	×	√	×

Table 2 Comparison of domestic and international IFMs
表 2 国内外工业大模型比较

序号	模型名称	开发机构	代表能力	模型结构/架构特点	典型应用场景	部署方式	国外/国内
1	Siemens Industrial Copilot ^[23-24]	Siemens & Microsoft	工业专家 Copilot,集成 Xcelerator 设计与 OT 数据	Copilot 模式+工业工具链嵌入	设计仿真、设备维护、工艺优化	本地部署+云协同	国外
2	IBM watsonx ^[25]	IBM	企业级数据分析+AI 开发平台,支持多模态任务	多模态+RAG+AutoAI 框架	企业流程分析、文档总结、预测建模	混合云、本地部署	国外
3	NVIDIA ACE ^[26]	NVIDIA	AI 数字人/多模态 Agent,适配边缘与机器人场景	ACE+Avatar+RTX 推理平台	工业数字人、远程控制、机器人交互	GPU 边缘部署+云同步	国外
4	FactoryTalk Design Studio Copilot ^[27]	Rockwell Automation & Microsoft	工业问答助手,支持 MES/HMI/ERP 系统对接	LLM+HMI 接口+知识嵌入式辅助	操作流程辅助、HMI 问答、维护建议	本地部署为主	国外
5	ABB Ability™ Genix Copilot ^[28]	ABB & Microsoft	流程工业知识增强与分析引擎	Genix 知识图谱+大模型融合	流程工业运营分析、设备监控	云边协同	国外
6	Honeywell Forge Intelligent Assistant ^[29]	Honeywell	资产管理、预测性维护与流程优化助手	Assistant 架构+资产上下文理解	智能运维、故障预测、安全巡检	企业本地部署	国外
7	EcoStruxure Copilot ^[30]	Schneider Electric & Microsoft	能源管理与可持续性优化智能助手	EcoStruxure 平台深度集成	碳排分析、能效优化、系统调度	私有云部署	国外
8	盘古大模型 ^[31]	华为	云边协同三层大模型架构,适配矿山等场景	盘古系列+行业专属数据精调,L0/L1/L2 分层	煤矿掘进调度、安全预测、视觉识别	云边一体+矿区边缘推理	国内
9	文心大模型 ^[32]	百度	工业文档解析与图文问答,强化中文理解	文心一言+工业插件化组件	设备文档问答、操作流程分析、质量追溯	百度千帆+私有部署	国内
10	通义大模型 ^[33]	阿里云	通义千问大模型平台,支持 Agent+多模态	Qwen 模型+DashScope 平台+MCP 工具集	工业开发 Agent、研发辅助、文档生成	公有云+私有云+API 调用	国内
11	“DeepSeek”系列 ^[34-35]	深度求索 (DeepSeek)	开源大模型推理平台,工业知识问答与 Agent 构建	DeepSeek-V3/R1+MoE+工具调用支持	智能问答、知识检索、流程文档摘要	模型开源+云平台接入	国内
12	“GLM”系列 ^[36]	清华大学 & 智谱 AI	GLM 系列大模型,支持 Agent、多模态、中文理解	GLM-130B/GLM-4 + AllTools Agent + 多模态	中文工业问答、代码生成、多模态识别	全开源部署+HuggingFace 支持	国内

1 工业大模型架构与定义

1.1 工业大模型定义

工业大模型是指应用于工业领域的大规模预训练人工智能模型,属于行业模型的范畴。该类模型具备跨任务、跨领域的泛化能力,已成为赋能千行百业智能化转型的关键底座。业界逐步形成了“通用模型—行业模型—垂直模型”三级层次的理解体系,如图 1 所示。通用工业模型指面向工业通用知识和通用任务的大模型平台,以理解和管理工业领域普适知识为主要目标。这类模型往往由顶尖 AI 企业基于海量通用数据训练构建,侧重提供最底层、跨行业的基础智能能力模块。行业大模型则是在通用模型基础上,融合特定垂直行业的数据和知识进行二次训练和优化的大模型平台,服务于某一工业垂直领域^[37-38]。行业模型注入了行业专业文档、传感数据等领域信息,具有较强的专业性,能够理解行业特有的术语和规律,在本行业比通用模型更精准。微软亚洲研究院提出通过工业数据后训练 LLMs 以构建工业基础模型,将工业表格、时序等结构化知识嵌入通用模型中,以提升模型对工业数据的理解和推理能力^[39-40]。场景大模型是指面向具体应用场景或任务目标,在通用或行业模型基础上进一步细化微调的专用模型。场景模型以解决工业细分场景中的特定问

题为目的,通常只在企业内部数据上训练,针对特定任务进行了优化,具有很强的针对性和安全性。例如,在某制造企业内部构建的产品质检对话模型,或面向特定产线的故障诊断模型,都属于场景级大模型。场景模型是工业大模型落地应用的“最后一公里”适配,能够将通用/行业模型的能力迁移到企业具体业务中。

三层模型之间具有承继关系,通用模型提供通用智能底座,行业模型在其上进行领域迁移微调,场景模型进一步针对特定任务精调适配。这种层次体系有效平衡了通用能力与专用需求。例如,有调查指出,仅依赖通用大模型往往无法满足高度专有的工业场景需求,约 50% 的企业认为通用模型能力与业务需求存在不匹配,必须通过行业与场景模型加以弥补^[41]。因此,工业大模型的发展趋势是以通用模型为基础,结合行业知识训练行业大模型,再由行业大模型派生出满足特定任务的场景模型^[37]。当前,我国政府也在积极布局这一体系。可以预见,通用、行业、场景三级模型协同,将成为工业智能化的重要技术架构,有力支撑工业领域从数字化向智能化跃升。

1.2 工业大模型架构

工业大模型通常采用分层架构设计,以系统地支撑从底层硬件到顶层应用的完整体系。Lee 等^[42-43]提出了一个四层参考框架,包括硬件基础、计算引擎、AI 算法层和工业应用层。任磊等^[22]提出了一个工业大模型五层架构体系,

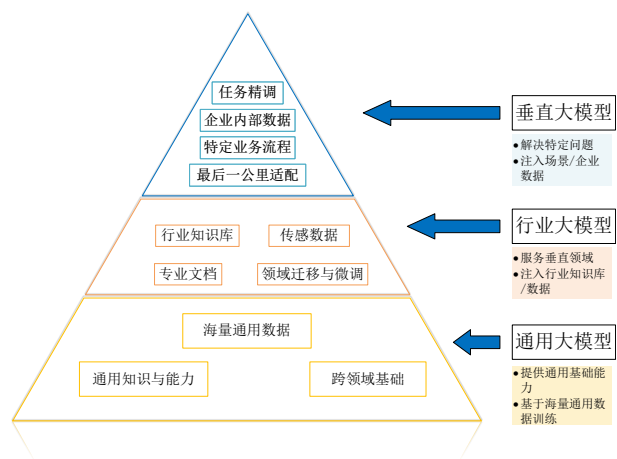


Fig. 1 Definition of IFMs

图1 工业大模型定义

包括基础设施层、基座层、模型层、交互层、应用层。在此基础上对工业大模型的五层架构进行优化,将不同层级的技术要素有机结合,为智能制造等应用场景提供支持。工业大模型技术体系以五层架构为核心框架,涵盖从底层基础设施到顶层应用的完整技术链条。这一架构为工业领域提供了高效、稳定、智能的技术支持,有力推动了工业智能化转型,具体架构如图2所示。

(1)基础设施层。基础设施层是工业大模型的“技术基座”,提供必要的算力、存储和网络支持。算力方面,训练和推理工业大模型需要高性能计算集群,这些集群通过并行计算节点快速处理海量多模态数据,满足智能制造等场景对实时大规模计算的需求。GPU(图形处理单元)和TPU(张量处理单元)等专用加速器在此层发挥关键作用,GPU以其强大的并行浮点运算能力广泛用于模型训练,TPU在深度学习推理中以高效低耗著称,常用于机器视觉控制等实时任务。数据存储方面,工业数据包含结构化与非结构化信息,需要分布式存储和数据库支撑。分布式文件存储系统或对象存储通过跨节点冗余存储,提供扩展性、可靠性和高吞吐量;数据库等新兴存储系统可以对多模态数据进行向量化索引,支持工业场景下基于特征匹配的高速检索。网络连接方面,工业场景下要求极低延迟与高带宽的数据传输。高速光纤网络可保证大规模数据的长距离传输,而新一代5G通信技术通过千兆带宽和毫秒级延迟,为工业大模型提供了边缘实时应用所需的传输保障。基础设施层强调云、边缘和本地计算的协同,云端提供大规模训练与集中推理能力,边缘计算将推理任务下沉至靠近数据源的节点,以降低时延、提高实时性^[44-45]。在对数据安全和隐私要求极高的领域(如国防、关键设备制造),则可能采用本地部署策略,确保数据不离开本地环境。

(2)基座层。基座层是工业大模型的核心支撑,通过构建通用能力模块,为下游应用提供基础模型和优化工具。多模态预训练是其关键技术之一。在工业场景中,数

据来源多样,包括设备日志、CAD文件文本、监控视频和语音告警,以及各种传感器数据(温度、压力、振动等)。传统的单一模态模型难以处理异构信息,而多模态预训练模型通过在大规模工业数据上统一训练,学习不同模态间的关联和共同语义,为生产优化、故障诊断等下游任务提供通用表征能力^[40]。工业机理与知识的融合也是基座层的重要组成,通过将工业领域的专业知识和物理机理(如流体力学、电力系统规律、化工反应机理等)以规则、约束或知识图谱的形式嵌入模型训练与推理过程,可显著提升模型对工业问题的理解和可解释性。基座层提供模型优化工具,针对工业环境中边缘设备算力有限的特点进行模型压缩与加速。其常见技术包括模型剪枝(去除冗余参数)、量化(降低参数精度)、知识蒸馏(将大模型知识迁移到小模型)等,从而大幅减少模型体积和计算量,使得大模型能够在嵌入式设备或边缘节点上实时运行。

(3)模型层。模型层是面向具体需求的核心环节,目标是通过深度适配实现工业场景中任务和领域的高效应用。模型层包括任务导向模型和行业领域模型两大类。任务导向模型通过对基座层预训练模型进行指令微调或任务微调,使其在某一具体工业任务(如设备故障诊断、生产优化、质量检测等)上表现优异。任务导向模型往往结合迁移学习技术,将预训练模型在通用领域的能力迁移到专业任务上,从而缩短开发周期并降低数据标注成本。行业领域模型则进一步聚焦特定制造领域,通过整合行业知识图谱和领域适配层以构建专门化模型。不同工业领域(如航空航天、汽车制造、电力系统、石化冶金等)的业务流程和技术规律差异巨大,通用模型难以直接满足这些特定需求。行业模型通过将该行业的核心知识(如飞行空气动力学、车辆动力学、冶金过程参数等)融入网络结构或输入特征,使模型更好地理解行业内关联并生成精准方案。模型层的设计还包括合理选取输入特征、设计目标函数以及实时学习能力等,使模型能够在实际工业环境中持续适应和优化^[42,46]。

(4)交互层。交互层是工业大模型与用户、环境之间高效沟通和协同的桥梁,其深度融合了多模态交互与工业智能体。多模态交互技术支持文本、语音、图像等多种输入输出方式,使用户能够以自然语言或语音等方式与模型进行交互。文本交互通过自然语言处理技术使模型理解用户意图并生成相应建议(如设备维护中的状态报告或故障处置方案);语音交互则结合语音识别和语音合成,使现场操作人员可通过语音与系统即时沟通,极大提高了工业车间等场景下的便利性和安全性;图像交互利用计算机视觉处理车间监控图像或产品外观,对质量检测和缺陷标注提供直观支持。这些模态间的信息融合让模型可以综合多源输入进行决策,从而更准确地定位问题。

工业智能体将大模型的智能决策与物理设备执行能力结合起来,实现环境的实时感知与智能控制。智能体可

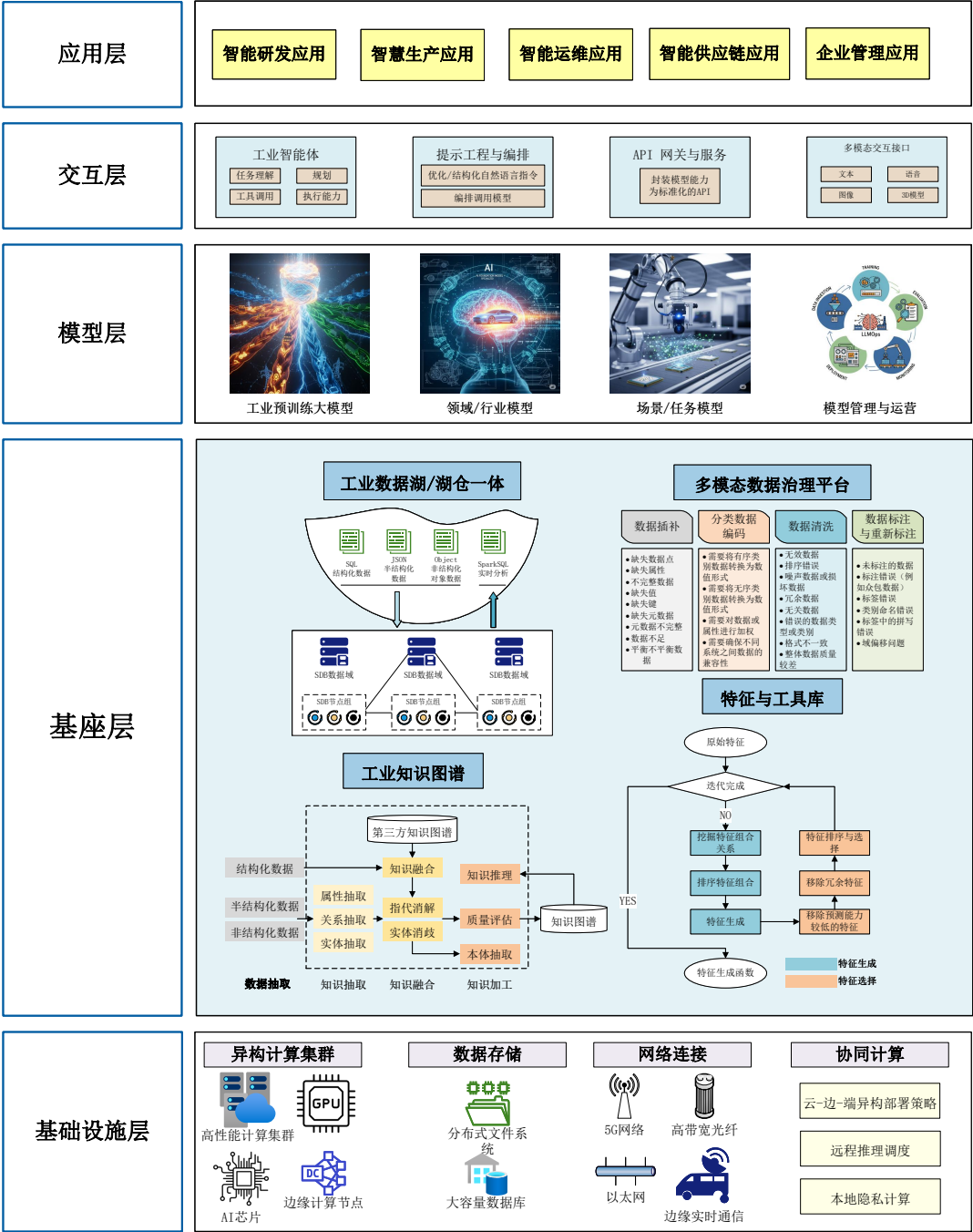


Fig. 2 Industrial Foundation Models (IFMs) architecture

图2 工业大模型架构

通过物联网和传感器网络实时采集设备运行数据,并在本地进行快速处理,减少通信延迟^[44]。结合边缘计算,智能体可以迅速执行模型生成的决策指令,完成复杂的操作任务。智能体常与增强现实(AR)等技术结合,为人机协作提供可视化指导,通过AR叠加生产步骤指引,实现复杂装配任务中的实时辅助。实现交互层性能的关键技术还包括分布式系统设计,使得多个智能体能够协同工作(如多机器人协同),以及交互安全性和鲁棒性设计,确保在严苛工业环境下系统稳定运行^[46]。

(5)应用层。应用层是工业大模型的最终落地环节,直接面向用户需求,体现技术价值。在这一层,工业大模型的功能模块包括智能问答、场景认知、过程决策、终端控制和内容生成等核心能力,它们共同支撑生产、管理和决策等实际业务。智能问答模块集成模型知识库和实时数据处理,可以回答设备状态、工艺流程或故障处理等专业问题,提高信息获取效率并减少对人工经验的依赖。场景认知模块通过图像处理、传感器分析等多模态感知技术对生产环境进行建模,可实时检测车间异常(如设备振动异

常、品质缺陷)并自动提出预警。过程决策模块则针对生产调度、供应链等复杂任务进行智能优化。终端控制功能将模型决策转化为对工业设备的直接控制。西门子和微软演示了利用生成式 AI 通过自然语言生成 PLC 程序的能力——将工程师指令实时转化为自动化控制代码,大幅降低编程复杂度^[47]。内容生成功能主要用于技术文档和报告自动化,通过结合自然语言生成技术和工业知识库,模型可生成故障报告、操作指南、工艺说明等高质量文档。

2 工业大模型关键技术

与通用大模型不同,工业大模型的核心目标是深刻理解并解决实际生产过程中的复杂问题,除需要具备泛化和推理能力外,还需要融合工业领域专业知识,满足工业场景的特殊要求和高标准。如图 3 所示,一条完整的工业大模型构建流程应涵盖从数据处理、模型构建到场景应用的完整链路。

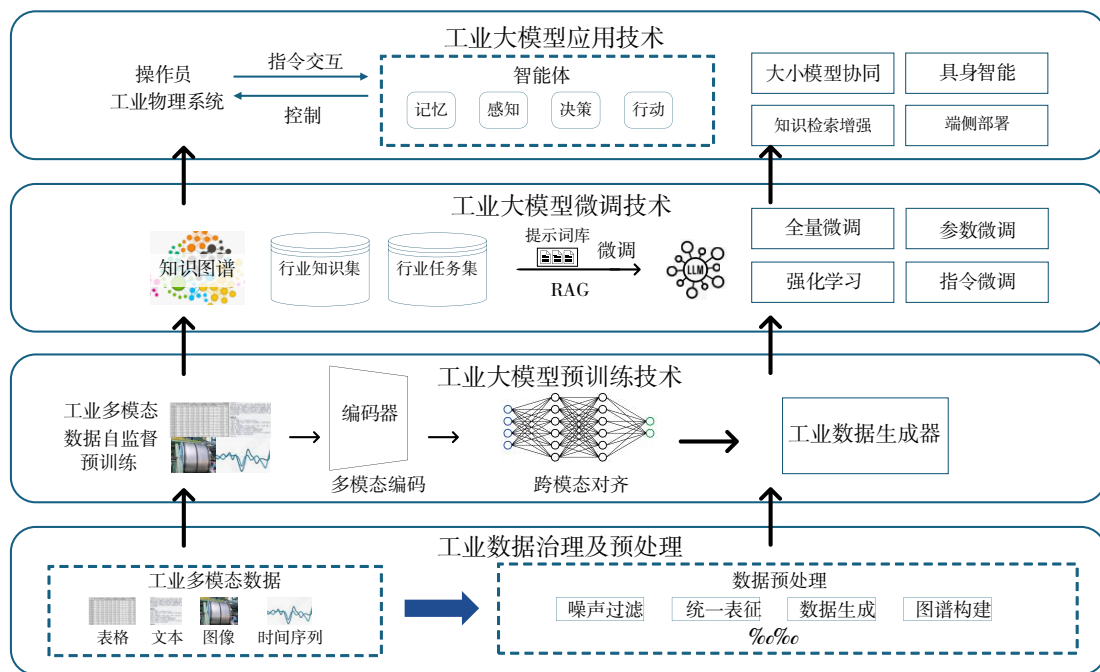


Fig. 3 Building process of Industrial Foundation Models (IFMs)

图 3 工业大模型构建流程

2.1 工业大模型数据治理与预处理技术

工业数据的治理与预处理是构建工业大模型的基础,其质量直接影响模型的性能与泛化能力。与互联网数据相比,工业数据呈现出多源异构、强时序性、高关联性等显著特征^[21]。多源异构性体现为数据来源广泛,包括传感器、控制系统、MES 系统等,数据类型涵盖结构化时序数据、非结构化文本与图像等;强时序性表现为采集频率高且连续性强,例如设备振动信号可能以毫秒级频率生成^[48];高关联性体现在生产环节中变量间的耦合,如温度、压力与产物纯度之间的动态关系。为了实现高质量的数据治理,需要先对原始数据进行清洗与规范化处理。包括剔除传感器噪声与异常值、插值补全缺失数据、统一物理量纲和编码规则等。同时,为提升数据的结构性和可检索性,往往还需构建工业元数据体系与知识图谱,将设备、工艺、产品等关键实体及其关系进行结构化存储,为后续数据集成、管理和语义查询提供支撑。在这个过程中,工业数据治理的主要任务包括数据质量维护、数据安全与隐私保护、数据标准化以及数据价值挖掘等方面。具体而言,需要在解决数据不完整、不准确、不一致、不及时等问题的

同时,在涉及敏感工业数据和个人信息的情况下,确保数据安全性和隐私性。另一方面,需要考虑如何统一不同数据源的数据格式、协议和标准,以及如何从海量工业数据中提取有价值的信息和知识,为工业决策提供支持。

数据预处理是在上述治理基础上的具体操作,包括对不同来源数据的融合和增强。例如,对图像数据可采用图像增强、合成缺陷图像等方式扩充少量缺陷样本;对时序数据可进行归一化、短时傅里叶变换等;对文本数据可利用领域词典或迁移学习生成补充描述。其中,数据增强尤为关键,如针对工业场景中缺陷样本稀缺的问题,Duan 等^[49]提出利用数据高效的 StyleGAN2 网络并结合“缺陷感知”特征块生成多样化缺陷图像,有效缓解工业质检中缺陷样本匮乏的问题。相比传统的几何变换和噪声添加等基础增强方法,该方法能生成更逼真的缺陷样本并保持视觉特征一致性。此外,针对工业领域大型图像—文本对数据的匮乏,Moenck 等^[50]构建了一个自动化流水线,将网络爬取的工业图文信息整理为大规模训练集,无需人工标注即可生成工业场景特定的语言—图像数据集,用于视觉基础模型预训练。与依赖人工标注的方法相比,该自动化方

案显著降低了数据获取成本并提高了构建效率。同时,可利用知识图谱统一表示设备和工艺信息,使得各种传感器数据和文本说明可以相互关联。或者,可采用自监督预训练任务对未标注的工业数据进行表征学习,结合领域规则过滤不合规数据。例如,Zhou等^[51]指出,集成领域数据并通过严格的数据治理,可以为工业大模型提供高质量数据支撑,使其在掌握大模型通用知识的同时具备工业专业性。相较于传统分散式数据处理方式,这种集成化治理方法能建立统一的质量标准并实现多源数据的有效整合。通过综合的数据治理流程和针对性预处理策略,工业大模型在训练前就能获得结构清晰、高质量且贴合实际生产场景的数据,有效提升工业数据集的质量与多样性,为大模型后续训练奠定坚实基础。

2.2 工业大模型预训练技术

工业大模型的预训练通常是大规模自监督学习过程,从海量工业数据中自动学习通用特征表示,为后续特定工业任务的微调或直接应用提供基础。作为工业大模型构建的核心环节,预训练直接决定了模型泛化能力和迁移性能。基于预训练目标的不同,工业大模型预训练技术可分为基于掩码语言模型(Masked Language Modeling, MLM)的预训练技术、基于对比学习的预训练技术、基于自回归的预训练技术、基于知识增强的预训练技术,以及基于多模态融合的预训练技术。

MLM通过随机掩码输入文本中的部分Token,让模型根据上下文预测被掩码的Token,有效捕捉语义依赖关系。工业大模型中的MLM针对工业文本特点进行了优化,包括对专业实体(设备名称、部件名称、工艺参数)的特定掩码处理,以及考虑工业文本结构特点的结构化掩码策略。例如,贾骏等^[19]提出的PowerBERT采用多层Transformer编码器,设计了字符掩码、实体掩码和片段掩码3种策略,分别学习基础语义、专业术语和段落结构,使模型更好地理解电力文本中的专业知识。

对比学习通过最大化正样本对语义相似度、最小化负样本对相似度以学习有效表示,在工业场景常用于跨模态对齐。CLIP等典型方法通过对比学习同时训练图像和文本编码器,使匹配的图文对在特征空间距离更近。Moenck等^[50]基于工业语言—图像数据集对CLIP进行预训练,适应工业图像和语言对齐任务,能高效利用无标注工业多模态数据,为故障诊断、跨模态检索等任务提供基础。与MLM方法相比,对比学习在多模态融合任务上具有天然优势,但在处理单模态序列建模时不如自回归方法灵活。

自回归预训练基于序列生成,模型根据前序内容预测下一元素,主要应用于工业文本或时序数据生成。该方法可生成设备操作指南、维护报告等,但由于单向性限制,实际应用中常与其他预训练目标结合使用。Ren等^[48]将非线性自回归网络与经验模态分解相结合,用于工业时间序列预测,提升了铁矿石进口量预测等任务准确性。相比对

比学习方法,自回归预训练在序列生成和时序预测任务中更具优势,但在跨模态对齐和特征匹配方面能力有限。

知识增强预训练将工业领域专业知识融入模型,提高对专业概念和关系的理解能力。常见方法是利用工业知识图谱或领域知识库引导预训练过程。Huang等^[52]构建了语言模型与知识图谱结合的预训练框架,通过掩码实体预测和关联预测任务实现图谱节点与文本概念对齐,使模型在学习语言表示同时获取结构化信息。此外,还可通过知识蒸馏或知识注入策略,将专家知识以软标签或损失项形式融入预训练。与自回归方法相比,知识增强预训练更注重结构化知识的显式建模而非序列生成,能够更好地处理概念关系和实体对齐任务,但在动态时序建模方面不如自回归方法。

多模态融合预训练对工业场景中的多种数据进行联合建模,学习统一的多模态表示。工业生产过程中涉及多种异构数据模态(如图像、文本说明、传感器时间序列、结构化表格等),融合这些信息有助于提高故障检测、工艺优化和决策分析等能力。例如,在跨模态缺陷检测场景中,可结合视觉图像与传感器数据共同识别设备异常。在工艺协同优化中,不同模态信息(如设备工况记录和工艺规范)需要对齐以改进生产流程。Wen等^[53]提出的CPIR模型通过引入一个共享潜在中间态,将RGB图像和点云等模态的特征映射到同一隐空间,从而解耦各模态间的特征映射过程,并同时跨模态预测和模态内重建任务,有效提升工业场景下的异常检测性能。Wang等^[54]提出Multi-3D-Memory(M3DM)多模态异常检测方法,设计了Patch-wise Contrastive Learning机制,对点云与RGB图像的局部特征进行无监督对比融合,并在MVTec-3D-AD数据集上获得了优异的异常检测与定位指标结果。对比Wen等^[53]的共享隐空间方法,Wang等^[54]的局部对比融合策略在异常定位精度上更具优势,但CPIR模型在跨模态预测任务中表现更为均衡。

2.3 工业大模型微调与检索增强生成

工业大模型在实际部署中需要针对特定领域和任务进行优化,以满足工业场景的精准性和可靠性要求。如图4所示,微调和RAG是两种主要的优化策略,它们通过不同技术路径提升模型在工业应用中的表现。微调通过调整模型参数以学习领域知识,而RAG则通过外部知识库增强模型的信息获取能力^[55]。这两种方法各有优势,可以根据具体应用场景的需求和资源约束进行选择或组合使用。

工业大模型微调技术主要包括监督学习微调、强化学习微调和参数高效微调3种策略。监督学习微调利用工业领域的标注数据对预训练模型作进一步训练,能够显著提升模型在特定任务上的表现^[5,22]。参数高效微调技术如LoRA通过在模型中引入可训练的低秩矩阵,仅微调少量参数即可实现有效的领域适应。这种方法可显著降低

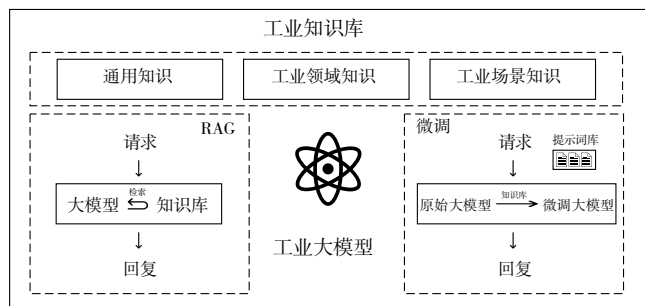


Fig. 4 RAG and fine-tuning application paradigms

图4 检索增强生成与微调应用范式

计算资源需求,同时避免灾难性遗忘问题。QLoRA 进一步结合量化技术,在保持性能的同时大幅减少内存占用^[56]。强化学习微调则通过构建奖励函数,使模型输出更符合工业场景的安全性和合规性要求。相较于参数高效微调专注于计算效率,强化学习微调更注重输出质量的控制和安全性约束,但训练过程更为复杂。

检索增强生成技术通过将预训练的参数化记忆与非参数化外部知识库相结合,为工业大模型提供动态、实时的知识支持。RAG 系统包含检索器和生成器两个核心组件,检索器从外部知识库中检索相关信息,生成器基于检索结果生成最终输出^[55]。在工业应用中,RAG 特别适用于需要频繁更新知识或处理大量技术文档的场景,如设备维护手册查询、工艺规程检索等^[57]。与强化学习微调需要精心设计奖励函数不同,RAG 通过外部知识检索提供信息增强,实现难度相对较低但对检索系统质量依赖较高。高级 RAG 技术从简单的 naive RAG 发展到 advanced RAG 和 modular RAG,包括多跳推理、层次化检索和自适应检索策略,进一步提升了系统的智能化水平^[58]。此外,混合检索增强生成方法在智能制造领域的专用问答系统中展现出显著优势,通过结合向量检索和知识图谱检索,有效提升了系统的响应准确性^[59]。

微调和 RAG 在工业应用中各有适用场景和优势。微调更适合任务相对固定、有充足标注数据的场景,能够深度定制模型行为,但需要较高的计算资源且更新知识需要重新训练。RAG 则更适合知识变化频繁、需要实时信息的场景,具有知识更新灵活、计算开销相对较小的优势,但检索质量直接影响生成效果^[60]。对比 RAG 和常见微调技术在长尾数据集上的性能差异,结果显示,在所有的实体中,微调均能产生一定的性能提升,且在极受欢迎和极不受欢迎类别中的提升最为显著。而 RAG 通过与微调技术结合,能够获得更优结果,但其提升会随着模型规模的增大而减小。

2.4 多智能体协同优化技术

在工业场景中,多智能体系统能够将复杂的生产流程或能源系统划分为若干协同子系统,每个智能体负责局部决策并通过通信或集中训练实现全局优化。典型应用包括生产调度、能耗管理、设备协同控制等。近年来,研究者

多采用多智能体深度强化学习或分布式优化方法,将车间或设备看作自治智能体,通过去中心化执行、集中式训练等框架提升系统可扩展性和鲁棒性。例如,Zhu 等^[61]针对多能源工业园区提出了基于多智能体深度强化学习的能量管理框架,通过去中心化执行、集中训练的方式最小化园区长期能耗成本,同时满足负载需求。Jang 等^[62]针对半导体工厂的动态调度问题提出了“领袖—跟随”结构的多智能体 RL,将调度任务拆分给不同子代理并用规则约束避免决策失误,实验结果表明该方法在多变负荷情况下能获得更高的调度效率和鲁棒性。相比 Zhu 等^[61]的去中心化执行模式,Jang 等^[62]的“领袖—跟随”结构通过层次化决策提供了更强的协调控制能力,但在系统扩展性方面不如完全去中心化的方法灵活。

3 工业大模型应用

工业大模型通过与工业场景的深度融合,推动了生产、运营和管理等环节的智能化转型,逐步实现“自适应、自决策、自执行”的工业智能化阶段。研究表明,大模型在工业研发设计、生产制造、运营管理和产品服务优化等方面展现出巨大潜力,但其应用仍面临数据质量、领域知识对齐、模型可靠性与实时性等挑战^[63]。本文探讨了大模型在工业全链条中的应用,如图 5 所示,涵盖研发设计、智能制造、供应链与物流优化、智能运维、高性能材料等 5 个主要领域。

3.1 智能研发与设计生成

工业大模型正在融入计算机辅助设计和工程仿真工具,实现设计流程的智能化升级。经过微调的大规模语言模型可以理解工程几何信息并生成高质量的 CAD 草图和参数化设计。例如,Wu 等^[64]提出 CAD-LLM 框架,利用预训练语言模型处理工程草图数据,微调后能够生成复杂机械零件的参数化设计。

在芯片与硬件设计方面,工业大模型通过构建面向芯片设计的域适配模型,在生成 EDA 脚本等任务上可以获得超过通用大模型的表现,其基本生产流程如图 6 所示。Li 等^[65]利用对话式 LLMs 辅助从架构规范到硬件描述语言的自动生成。在一个案例中,LLMs 帮助定义指令集、实现组件设计并修复错误,最终实现了一款 8 位累加器微处理器设计,并成功生成了首个完全由 AI 输出的硬件描述代码完成流片。Liu 等^[66]提出 ChipNeMo,通过采用定制化分词器、领域自适应持续预训练等领域适配技术以提升大模型的芯片设计效率,并在工程助手聊天机器人、EDA 脚本生成和漏洞总结分析等 3 个核心芯片设计应用中实现了显著的性能提升。

此外,工业大模型在模拟电路设计领域也开始发挥作用。Yin 等^[67]提出了 ADO-LLMs 框架,将大模型与贝叶斯优化相结合,用于模拟电路尺寸优化。该方法利用模型的

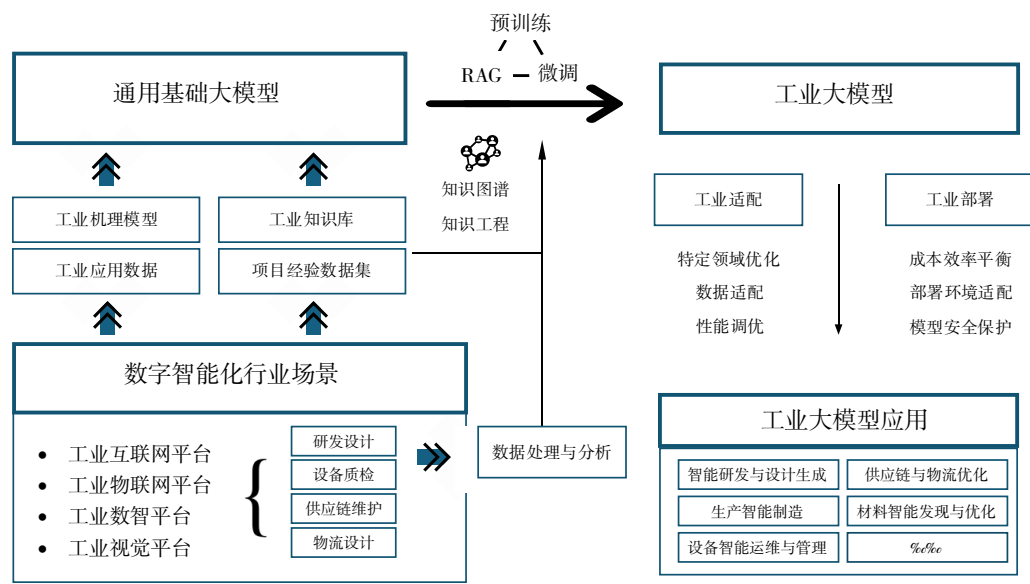


Fig. 5 Applications of IFMs
图5 工业大模型应用

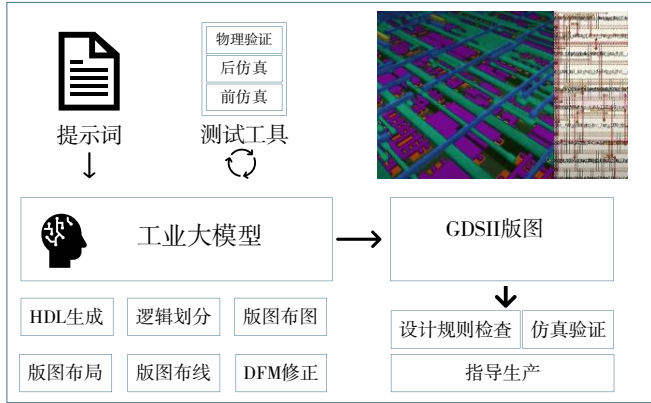


Fig. 6 Digital IC design process of IFMs
图6 工业大模型的数字芯片设计流程

先验知识快速生成可行设计点,并通过迭代优化验证,提高了模拟电路设计的效率和质量。德州仪器也提出了一种将电路设计知识编码为结构化文本的大模型通用框架LLM-USO,通过构建知识摘要并在相似子电路间复用,显著提升了优化效率^[68]。

3.2 大模型驱动的智能制造

在生产制造环节,大模型通过代码生成和智能决策促进生产模式向柔性化转型。一方面,工业大模型可将自然语言指令转换为机器可执行的程序,简化了传统编程流程并减少人工错误。例如,Fakih等^[69]提出的LLM4PLC框架利用用户反馈与外部验证引导大模型生成可编程逻辑控制器程序代码,该方法将PLC代码生成的成功率从47%提升到72%,专家评估的代码质量得分从2.25/10提升至7.75/10。另一方面,在生产排程与控制方面,基于大模型的调度系统可以应对复杂多任务场景。比如,Jeong等^[70]在多车间、多任务的复杂生产环境中,将大模型与历史调度数据相结合,实现对新任务的最优排程建议。通过预测

生产线上各工序的资源需求和瓶颈,大模型辅助的调度方案能够提高生产线利用率并缩短交付周期。

3.3 设备智能运维与管理

在设备管理和维护领域,大模型聚焦于状态监测和预测性维护。研究表明,LLMs可以分析设备的传感器数据、维护记录和运行日志,自动发现潜在故障模式并生成维保建议^[63]。与传统机器学习方法相比,大模型能够持续学习设备的运行规律,不断更新故障预测模型,从而提高预测准确度和维护效率。例如,Palma等^[71]在风电场压缩机故障检测中采用了基于Qwen-2.5B大模型的框架,通过融合振动信号、温度曲线和运行日志,实现了92.3%的故障检测召回率,并将维护相关成本降低约18%。这些成果表明,工业大模型通过多源数据的融合分析,能够为设备的健康管理提供精准的预测和决策支持。

3.4 智能供应链与物流协同优化

在供应链规划与物流运输中,工业大模型通过对海量数据的分析,可以实现需求预测和路径优化的智能化。通过综合历史销售数据、市场趋势和社交媒体反馈等信息,工业大模型对未来需求进行更精准的预测。例如,采用大模型进行情感分析能够量化消费者评价和舆情趋势,为生产计划和库存管理提供参考依据。与此同时,大模型可通过调用天气、交通等实时数据接口,对物流路线和配送计划进行优化。例如,Hadi等^[72]发现通过引入实时交通和天气信息后,大模型驱动的调度系统可以显著缩短运输距离、降低物流成本并提高响应速度。大模型通过智能分析和外部信息融合,极大增强了供应链的预测能力和供应链响应的灵活性。

3.5 高性能材料智能发现与仿真优化

工业大模型依托大数据分析和仿真优化技术,能够加速新材料的设计与研发。通过挖掘海量材料数据库和高

通量仿真数据,大模型可以快速优化材料的化学成分和微观结构,从而提高材料性能^[73-74]。此外,与传统基于图神经网络的方法不同,工业大模型能够从材料的文本描述中直接预测材料性能,为材料基因工程提供了全新范式。

Niyongabo 等^[75]开发的 LLM-Prop 框架首次证明了大语言模型在晶体材料性能预测中的卓越能力。该方法基于 T5 编码器架构,通过微调预训练模型直接从晶体文本描述预测材料性能,在带隙预测任务上比当前最先进的图神经网络方法 ALIGNN 提升约 8%。该模型能够有效捕获空间群对称性和 Wyckoff 位点等关键信息,这些信息在传统图神经网络中难以准确建模^[76]。

Ding 等^[77]提出的 MatExpert 框架通过模仿人类专家的材料发现过程,实现了自主的晶体结构生成和性能预测。该系统结合大语言模型与对比学习技术,在材料生成任务中超越了现有最先进方法。Jia 等^[78]开发的 LLMatDesign 系统利用 LLMs 智能体将人类指令转化为材料修改操作,通过自我反思机制快速适应新任务,实现了零样本材料设计。

这些工业大模型在材料科学中的成功应用正在重新定义材料发现范式,从传统的基于物理模拟和实验试错转向基于文本理解和语言推理的智能预测。这一转变不仅提高了材料设计效率,还为跨学科知识整合和创新材料概念生成开辟了新道路。

4 工业大模型应用挑战与展望

4.1 工业大模型应用挑战

目前,工业大模型从实验室和云端真正部署到复杂、严苛且高度动态的工业环境中,仍然面临着一系列独特的挑战。其中的核心挑战可分为以下 4 类:

(1) 工业大模型部署困难。将工业大模型成功融入到实际生产流程之中是一项涉及基础设施、数据安全、成本控制和法规遵循的系统性工程。首先,在设施投入与运维成本方面,大模型的训练和推理需要巨大的算力支持,需要企业进行高昂的硬件投入,例如搭建高性能 GPU 服务器。图 7 是一些典型大模型的预估训练成本^[79],企业若选择自托管开源模型,仅租赁一台搭载 A100GPU 的服务器就可能造成每月数千美元的花费,而这通常只是处理中等查询量的起步费用。此外,模型部署后的持续监控、维护、扩展以及应对“冷启动”延迟等运维问题,都需要专业的大语言模型运维团队支持,进一步增加了人力和运营开支。与此同时,监管合规的复杂性也为大模型的工业应用带来了新的挑战。各国政府和监管机构正在加速制定相关法规,以确保 AI 系统的使用是公平、透明和负责任的,企业需要为证明其 AI 系统的合规性做好准备,包括详细记录决策过程和确保模型透明度^[80]。

(2) 超长上下文与实时多模态交互。工业场景往往涉

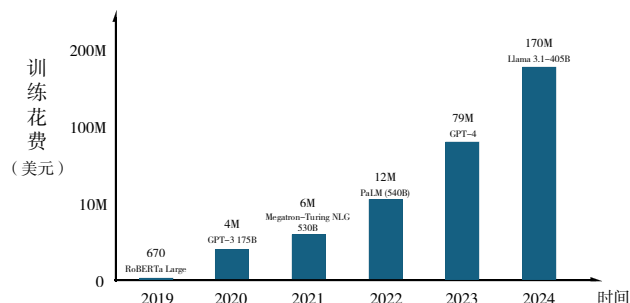


Fig. 7 Estimated training cost of typical large language models

图 7 典型大模型预估训练成本

及其庞杂的上下文信息,例如生产手册、维修日志、设备运行日志和长时视频流等,而传统大型语言模型的上下文窗口有限,对超长序列的处理能力不足,成为大模型实际应用中的一大瓶颈。工业应用常常需要处理时间跨度极长、内容极为丰富的文档和数据流。例如,对一台大型设备进行故障诊断,模型可能需要一次性分析其长达数十年的设计文档、维修手册、历史运行日志和传感器时间序列数据。当前,尽管主流大模型不断扩展其上下文窗口,但高效处理数百万甚至更长 Token 序列仍面临巨大挑战,包括计算量激增、注意力机制效率下降和“上下文丢失”等技术瓶颈。

实时多模态交互能力是实现高级人机协作和环境感知的核心。工厂环境是一个充斥着多种数据模态的场所,包括来自摄像头的视觉图像、来自麦克风的语音、来自传感器的振动和温度读数,以及存储在系统中的 CAD 设计图和工艺流程文件。工业大模型需要能够实时地融合与理解这些异构数据,才能完成复杂任务。例如,质量检测员可以通过语音描述缺陷,同时用摄像头指向缺陷部位,模型需要结合这两种输入,实时判断缺陷类型并给出处理建议。实现这种流畅的实时交互,需要解决不同模态数据在语义层面的精准对齐、模型端到端推理延迟等难题,这也是当前多模态大模型研究的前沿方向^[81]。

(3) 工业大小模型协同。为平衡工业大模型强大的分析能力与边缘端对低延迟、高隐私性的刚性需求,“大小模型协同”或“端云协同”架构已成为行业重要趋势。如图 8 所示,该架构的核心在于让不同规模的模型在云端和边缘端各司其职,形成优势互补,而非相互替代。Reddi 等^[82]指出,以往通过 RLHF 优化后的小型模型有时能胜过更大的未对齐模型,这意味着对特定任务可用紧凑、定制化的 SLM 替代通用 LLMs。针对协同推理,一些研究发展了模型切分和分层机制。Zhang 等^[83]提出的 EdgeShard 框架将大模型划分为多个“碎片”并部署到分布式边缘设备上,通过动态规划优化设备选择和模型分片,成功降低了推理延迟 50%,吞吐量提升 2 倍。这说明在工业物联网场景下,可以将一个大模型横向分片,利用多台边缘节点协同计算。

另有方法通过知识蒸馏将 LLMs 的能力转移给轻量级模型。Li 等^[84]提出混合蒸馏框架,将大模型的多路径推理

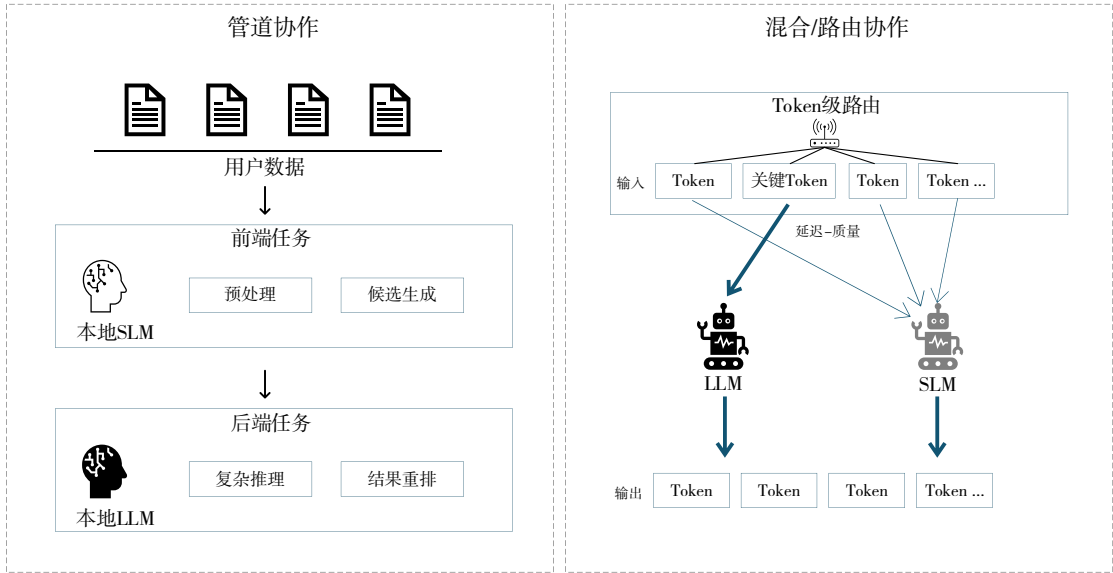


Fig. 8 Basic framework example of large-small model collaboration

图8 大小模型协同基本框架示例

能力融合进小模型,使得一个7B参数模型在数学推理上超过了GPT-3.5Turbo。工业上可利用此类技术,将大模型的“能力”压缩到更小的模型中,在边缘端设备上实现快速响应。更宏观地,研究人员提出了层级化的多层模型体系结构。Sarkar等^[85]构想了一种分层分布式LLMs架构,通过层次化的知识组织(上层通用大模型,下层领域小模型)为不同设备和应用需求提供灵活部署。该架构允许根据硬件能力和任务复杂度自动选择合适层级的模型,例如资源受限设备只加载任务相关的微型模型,从而节省计算资源。Chen等^[86]进一步对这类架构进行了系统化分析,提出了集中式、混合式、分散式等多种边缘智能设计策略。在集中式方案中,边缘设备作为终端,主要计算在后端服务器进行;而在混合或分散式方案中,模型和推理任务在边缘和云之间灵活拆分或在多设备之间共享,实现对时延、隐私与资源的平衡。综合来看,通过将大模型与小模型在架构、推理流程中组合使用,可以在保证工业系统智能化能力的同时,满足边缘计算的实时性和效率要求。

(4)可信工业大模型。考虑到工业大模型承担的专业生产任务对准确性的高度要求,模型可信性在工业智能系统中至关重要,主要包括可靠性、安全性、公平性、鲁棒性和可解释性等维度。可靠性方面,工控环境中模型必须在各种工况下稳定运行。然而,一项制造业实证研究发现,一个在实验环境表现良好的模型,在面对真实生产环境中的数据分布变化时经常会发生性能波动甚至失效^[87]。因此,工业场景中需持续监测模型输出,并通过严格的验证流程确保模型在部署生命周期内性能符合预期。安全性与鲁棒性则要求模型能够抵抗攻击和异常输入,这需要模型在被输入噪声或恶意干扰时仍需安全退化而非进行危害生产。公平性则要求模型决策不能引入偏见,因为工业环境中不同的群体对系统的影响不同,故系统应确保AI系统不传播或放大偏见,对所有群体产生公平的结果。可

解释性则有助于在生产环境中建立用户信任,生产线工程师需要了解模型为何作出某个预测,才能采纳并排查故障。相关研究指出,保障工业AI解释性尤其重要,应采用适当的可解释性工具提供决策依据^[87]。综合来看,可靠性、安全性、公平性、鲁棒性和可解释性构成工业大模型的核心信任属性,工业企业需在系统设计和验证过程中逐项评估和强化这些特性,以确保模型决策可控、可审计并符合伦理法规。

4.2 工业大模型应用展望

针对以上问题,本文提出未来工业大模型具有较大发展潜力的研究方向如下:

(1)轻量化模型架构与压缩技术。工业大模型的训练和推理需要巨大的算力支持,这导致了高昂的硬件投入和持续的运维开支。通过模型剪枝、量化、知识蒸馏等技术,将大模型的核心能力迁移到参数量更小的模型中,显著降低对高性能硬件的依赖和运维成本。同时,研发专门面向工业场景的模型压缩算法,在保持任务精度的前提下,实现边缘端设备的快速响应。此外,探索模型即服务的部署模式,通过云原生架构和容器化技术,降低企业的基础设施投入门槛,提升模型部署的灵活性和可扩展性。

(2)超长上下文与实时多模态交互支持。构建高效的长序列建模算法,如稀疏注意力机制、分层记忆网络等,突破传统上下文窗口限制,使模型能够处理数百万Token的工业文档和历史数据。在多模态融合方面,研究跨模态语义对齐技术和端到端的多模态理解架构,实现文本、图像、音频、传感器数据等异构信息的实时融合处理。重点攻克多模态数据的时空同步问题,开发低延迟的实时推理算法,满足工业场景对快速响应的严格要求。

(3)可信大模型体系构建。在鲁棒性方面,研发对抗样本防御和异常检测技术,使模型在面对噪声干扰或恶意攻击时能够安全退化,避免错误决策对生产造成损失。在

可解释性方面,开发面向工业场景的模型解释框架,提供决策路径可视化和因果关系分析,帮助工程师理解和信任模型输出。在安全性方面,建立模型行为监控和异常预警机制,确保模型在全生命周期内的稳定可靠运行。同时,研究公平性评估和偏见消除技术,确保 AI 系统的决策过程符合伦理规范和法规要求。

(4)虚实融合仿真与具身智能。探索工业大模型与数字孪生、具身智能的深度融合技术。将工业大模型作为数字孪生体系的“智能大脑”,通过理解物理世界的复杂机理和实时数据,实现高保真度的仿真建模和预测性维护。研发虚实映射算法,确保数字孪生体与物理系统的实时同步和一致性。在具身智能方面,构建从感知、理解到执行的闭环智能系统,使工业大模型具备自主操作和环境适应能力,推动工业生产向完全自主化方向发展。

5 结语

LLMs 技术作为当前 AI 领域最为重要的研究方向之一,已经在各领域中展现出巨大潜力。工业大模型作为 LLMs 技术与工业场景深度融合的产物,正加速推动制造业向智能化、绿色化、高端化方向演进。本文回顾了 LLMs 的架构、技术及最新应用进展,结合工业应用场景,从底层基础设施到顶层应用部署进行了深入探讨。本文对工业大模型架构进行了全新定义,提出了工业大模型五层架构体系,并揭示了当前工业大模型技术面临如数据偏差、行业知识建模不足和算力资源受限等多重挑战。

未来,工业大模型的发展将呈现出“通用能力与行业定制并进、云边协同与智能体融合互补、多模态感知与知识增强深度融合”的演化趋势。构建“技术先进、机制完备、生态开放”的工业大模型体系是推动中国工业 4.0 向中国工业 5.0 跃升的关键抓手,也是实现国家制造强国战略目标的重要支撑。通过构建工业大模型,进一步提高国内工业的核心竞争力,为中国在新一轮工业革命中占领先机。

参考文献:

- [1] Ministry of Industry and Information Technology, National Development and Reform Commission, Ministry of Education of the People's Republic of China, et al. Development plan for intelligent manufacturing during the 14th Five-Year Plan period [EB/OL]. https://www.gov.cn/zhengce/zhengceku/2021-12/28/content_5664996.htm.
工业和信息化部、国家发展和改革委员会、中华人民共和国教育部,等.“十四五”智能制造发展规划[EB/OL]. https://www.gov.cn/zhengce/zhengceku/2021-12/28/content_5664996.htm.
- [2] The State Council of the People's Republic of China. Made in China 2025 [EB/OL]. https://www.gov.cn/zhengce/content/2015-05/19/content_9784.htm.
中华人民共和国国务院. 中国制造 2025[EB/OL]. https://www.gov.cn/zhengce/content/2015-05/19/content_9784.htm.
- [3] 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.
- [4] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report [DB/OL]. <https://arxiv.org/pdf/2303.08774>.
- [5] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019: 4171-4186.
- [6] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. *Nature*, 2021, 596(7873): 583-589.
- [7] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models [DB/OL]. <https://arxiv.org/abs/2108.07258>.
- [8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//*Long Beach: 31st Conference on Neural Information Processing Systems*, 2017.
- [9] KAGERMANN H, HELBIG J, HELLINGER A, et al. Recommendations for implementing the strategic initiative INDUSTRIE 4.0: securing the future of German manufacturing industry; final report of the Industrie 4.0 Working Group[R]. *Forschungsunion*, 2013.
- [10] CAYLARP L, NOTERDAEME O, NAIK K. Digital in industry: from buzzword to value creation[Z]. *McKinsey & Company*, 2016, 2(1): 1-9.
- [11] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. *ACM Computing Surveys*, 2023, 55(9): 1-35.
- [12] ZHANG Q T, WANG Y C, WANG H X, et al. Comprehensive review of large language model fine-tuning[J]. *Journal of Computer Engineering and Applications*, 2024, 60(17): 17-33.
张钦彤,王昱超,王鹤羲,等. 大语言模型微调技术的研究综述[J]. *计算机工程与应用*, 2024, 60(17): 17-33.
- [13] LUO J Z, SUN Y L, QIAN Z Z. Overview and prospect of artificial intelligence large models[J]. *Radio Engineering*, 2023, 53(11): 2461-2472.
罗锦孙,孙玉龙,钱增志. 人工智能大模型综述及展望[J]. *无线电工程*, 2023, 53(11): 2461-2472.
- [14] QIN X L, GU X, LI D C, et al. Survey and prospect of large language models[J]. *Journal of Computer Applications*, 2025, 45(3): 685-696.
秦小林,古徐,李弟诚,等. 大语言模型综述与展望[J]. *计算机应用*, 2025, 45(3): 685-696.
- [15] REN L, WANG H, WANG Y, et al. Foundation models for the process industry: challenges and opportunities[J]. *Engineering*, 2025, 52: 53-59.
- [16] CHEN X H, FU W R, LIU C M, et al. Application of artificial intelligence large language model in power equipment operation and maintenance[J]. *Engineering Science of China*, 2025, 27(1): 180-192.
陈晓红,傅文润,刘朝明,等. 人工智能大模型在电力设备运维场景中的应用探讨[J]. *中国工程科学*, 2025, 27(1): 180-192.
- [17] LI G, FANG H, LIU Y P. Large-model drive technology in new power system: status challenges and prospects [J]. *High Voltage Engineering*, 2024, 50(7): 2864-2878.
- [18] ZHAO J H, WEN F S, HUANG J W, et al. Prospect of artificial general intelligence for power systems based on large language model: Theory and applications [J]. *Automation of Electric Power Systems*, 2024, 48(6): 13-28.
- [19] JIA J, YANG Q, FU H, et al. Research on pre-training language model construction and text semantic analysis based on power equipment big data[J]. *Proceedings of the CSEE*, 2023, 43(3): 1027-1037.
贾骏,杨强,付慧,等. 基于电力设备大数据的预训练语言模型构建和文本语义分析[J]. *中国电机工程学报*, 2023, 43(3): 1027-1037.
- [20] ZHANG H, SEMUJUU S D, WANG Z, et al. Large scale foundation models for intelligent manufacturing applications: a survey[J]. *Journal of In-*

- telligent Manufacturing, 2025: 1–52.
- [21] LI M, AN C F, GAO J, et al. Industrial big data analytics practice based on data lake platform: an example of intelligent oilfield energy efficiency analysis[J]. Data Governance, 2024, 43(11): 75–84.
李满, 安创锋, 高静, 等. 基于数据湖平台的工业大数据分析实践: 以智能油田能效分析为例[J]. 信息安全与数据治理, 2024, 43(11): 75–84.
- [22] REN L, WANG H T, DONG J B, et al. Industrial foundation model: architecture, key technologies, and typical applications[J]. Scientia Sinica Informationis, 2024, 54(11): 2606–2622.
任磊, 王海腾, 董家宝, 等. 工业大模型: 体系架构、关键技术与典型应用[J]. 中国科学: 信息科学, 2024, 54(11): 2606–2622.
- [23] Siemens and Microsoft. Siemens and Microsoft scale industrial AI [EB/OL]. <https://news.microsoft.com/source/2024/10/24/siemens-and-microsoft-scale-industrial-ai/>.
- [24] SiemensGlobal. Unlocking the power of generative AI: siemens industrial Copilot–SiemensGlobal [EB/OL]. <https://www.siemens.com/global/en/company/insights/unlocking-the-power-of-generative-ai-siemens-industrial-copilot.html>.
- [25] IBM. Realize the promise of AI with Watsonx [EB/OL]. https://www.ibm.com/products/watsonx?utm_source=chatgpt.com.
- [26] NVIDIA DEVELOPER. NVIDIA ACE for games [EB/OL]. https://developer.nvidia.com/ace-for-games?utm_source=chatgpt.com.
- [27] RockwellAutomation. Maximize productivity with FactoryTalk Design Studio [EB/OL]. https://www.rockwellautomation.com/en-gh/products/software/factorytalk/design-studio.html?utm_source=chatgpt.com.
- [28] ABB. ABB Genix™ Copilot [EB/OL]. https://new.abb.com/process-automation/genix/abb-genix-copilot?utm_source=chatgpt.com.
- [29] Honeywell. Honeywell Unveils innovative AI assistant for industrial operators in Honeywell Forge production intelligence [EB/OL]. <https://www.honeywell.com/us/en/press/2025/02/honeywell-unveils-innovative-ai-assistant>.
- [30] Schneider Electric Canada. Schneider Electric drives Generative AI productivity and sustainability solutions by integrating Microsoft Azure OpenAI [EB/OL]. <https://www.se.com/ca/en/about-us/newsroom/news/press-releases/schneider-electric-drives-generative-ai-productivity-and-sustainability-solutions-by-integrating-microsoft-azure-openai-655f799c5279d6c16108ba9d>.
- [31] Huawei Cloud. Pangu large models [EB/OL]. https://www.huaweicloud.com/product/pangu.html?utm_source=chatgpt.com.
华为云. 盘古大模型 [EB/OL]. https://www.huaweicloud.com/product/pangu.html?utm_source=chatgpt.com.
- [32] ERNIE (Encyclopedia of Representations from Advanced Intelligence) Large Language Models (LLMs). The ERNIE big model 4.5 series is open-sourced [EB/OL]. <https://wenxin.baidu.com/>.
文心大模型. 文心大模型 4.5 系列开源 [EB/OL]. <https://wenxin.baidu.com/>.
- [33] AlibabaCloud. Tongyi Large Model: the top choice for enterprises to embrace the AI era [EB/OL]. <https://www.aliyun.com/>.
阿里云. 通义大模型企业拥抱 AI 时代首选 [EB/OL]. <https://www.aliyun.com/>.
- [34] China Industrial Internet Research Institute. DeepSeek's "going viral" is expected to accelerate the popularization of AI [EB/OL]. <https://www.china-aii.com/xyxz/7140803.jhtml>.
中国工业互联网研究院. DeepSeek“出圈”有望加速 AI 普及 [EB/OL]. <https://www.china-aii.com/xyxz/7140803.jhtml>.
- [35] China Industry Internet (Beijing) Technology Group Co., Ltd. Smart Work + DeepSeek = China's first Industrial Foundation Models (IFMs) combo, accelerating the implementation of industrial intelligence [EB/OL]. <https://zhuanlan.zhihu.com/p/21301141904>.
中工互联(北京)科技集团有限公司. 智工+DeepSeek=? 中国首个工业大模型组合拳, 工业智能落地加速 [EB/OL]. <https://zhuanlan.zhihu.com/p/21301141904>.
- [36] Tsinghua Zhipu AI. Z Zhipu·one-stop large model development platform [EB/OL]. <https://bigmodel.cn/>.
清华智谱 AI. Z 智谱·一站式大模型开发平台 [EB/OL]. <https://bigmodel.cn/>.
- [37] YUAN X Q. Five essential issues of Industrial Foundation Models (IFMs) [EB/OL]. <https://13115299.s21i.faiusr.com/61/1/ABUIABA9GAAgnOCXsQYotuzJvAU.pdf>.
袁晓庆. 工业大模型的五个基本问题 [EB/OL]. <https://13115299.s21i.faiusr.com/61/1/ABUIABA9GAAgnOCXsQYotuzJvAU.pdf>.
- [38] RONG K, KANG Z Y, LUO Y N. Large Language Models (LLMs) empowering all industries: an ecosystem-based business model [EB/OL]. https://www.ccsn.cn/dkzgx/zgxp_zgshkxpj/2023nd4q_131885/202403/t20240308_5737565.shtml.
戎珂, 康正瑶, 罗怡宁. 大模型赋能万行万业: 生态型商业模式 [EB/OL]. https://www.ccsn.cn/dkzgx/zgxp_zgshkxpj/2023nd4q_131885/202403/t20240308_5737565.shtml.
- [39] BI K, XIE L, ZHANG H, et al. Accurate medium-range global weather forecasting with 3D neural networks[J]. Nature, 2023, 619: 533–538.
- [40] Microsoft Research Lab-Asia. Towards industrial foundation models: integrating large language models with industrial data intelligence [EB/OL]. <https://www.microsoft.com/en-us/research/articles/towards-industrial-foundation-models-integrating-large-language-models-with-industrial-data-intelligence/>.
- [41] Volcengine & IDC. White paper on Large Models application implementation: an action guide for enterprise AI transformation [EB/OL]. https://pdf.dfcfw.com/pdf/H3_AP202501091641873256_1.pdf?1736454894000.pdf.
火山引擎&IDC. 大模型应用落地白皮书企业 AI 转型行动指南 [EB/OL]. https://pdf.dfcfw.com/pdf/H3_AP202501091641873256_1.pdf?1736454894000.pdf.
- [42] LEE J, SU H, JI D Y, et al. Engineering artificial intelligence: framework, challenges, and future direction [DB/OL]. <https://arxiv.org/abs/2504.02269>.
- [43] LEE J, SU H. Rethinking industrial artificial intelligence: a unified foundation framework [DB/OL]. <https://arxiv.org/abs/2504.01797>.
- [44] BAYAR A, ŞENER U, KAYABAY K, et al. Edge computing applications in industrial IoT: a literature review [C]//International Conference on the Economics of Grids, Clouds, Systems, and Services, 2022: 124–131.
- [45] GrayStan. 5G and IIoT: emerging use cases in industrial manufacturing [EB/OL]. <https://www.telit.com/blog/use-cases-5g-iiot-manufacturing/>.
- [46] WANG C, HASLER S, TANNEBERG D, et al. LaMI: Large Language Models for multi-modal human-robot interaction [C]//Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024: 1–10.
- [47] Siemens and Microsoft. Siemens and Microsoft drive industrial productivity with generative artificial intelligence [EB/OL]. <https://newsroom.sw.siemens.com/en-US/siemens-microsoft-generative-ai/>.
- [48] REN L, WANG H, LI J, et al. AIGC for industrial time series: From deep generative models to large generative models [DB/OL]. <https://arxiv.org/abs/2407.11480>.
- [49] DUAN Y, HONG Y, NIU L, et al. Few-shot defect image generation via defect-aware feature manipulation [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 571–578.
- [50] MOENCK K, THIEU D T, KOCH J, et al. Industrial Language-Image Da-

- taset (ILID): adapting vision foundation models for industrial settings [J]. *Procedia CIRP*, 2024, 130: 250–263.
- [51] ZHOU J, CAO Y, LU Q, et al. Industrial large model: a survey [C]//21st International Conference on Manufacturing Research (ICMR2024), 2024: 10009.
- [52] HUANG X, YANG C. Pretrained language – knowledge graph model benefits both knowledge graph completion and industrial tasks: taking the blast furnace ironmaking process as an example [J]. *Electronics*, 2024, 13 (5): 845.
- [53] WEN S, WU H, NIU Y, et al. CIPR: multimodal industrial anomaly detection via latent bridged cross-modal prediction and intra-modal reconstruction [J]. *Advanced Engineering Informatics*, 2025, 65: 103240.
- [54] WANG Y, PENG J, ZHANG J, et al. Multimodal industrial anomaly detection via hybrid fusion [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 8032–8041.
- [55] GAO Y, XIONG Y, GAO X, et al. Retrieval-augmented generation for large language models: a survey [DB/OL]. <https://arxiv.org/abs/2312.10997>.
- [56] 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.
- [57] LI Z, WANG Z, WANG W, et al. Retrieval-augmented generation for educational application: a systematic survey [J]. *Computers and Education: Artificial Intelligence*, 2025, 8: 100417.
- [58] GUPTA S, RANJAN R, SINGH S N. A comprehensive survey of retrieval-augmented generation (rag): evolution, current landscape and future directions [DB/OL]. <https://arxiv.org/abs/2410.12837>.
- [59] WAN Y, CHEN Z, LIU Y, et al. Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing [J]. *Advanced Engineering Informatics*, 2025, 65: 103212.
- [60] SOUDANI H, KANOULAS E, HASIBI F. Fine tuning vs. retrieval augmented generation for less popular knowledge [C]//Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, 2024: 12–22.
- [61] ZHU D, YANG B, LIU Y, et al. Energy management based on multi-agent deep reinforcement learning for a multi-energy industrial park [J]. *Applied Energy*, 2022, 311: 118636.
- [62] JANG J, KLABJAN D, LIU H, et al. Scalable multi-agent reinforcement learning for factory-wide dynamic scheduling [DB/OL]. <https://arxiv.org/abs/2409.13571>.
- [63] UCAR A, KARAKOSE M, KIRIMÇA N. Artificial intelligence for predictive maintenance applications: key components, trustworthiness, and future trends [J]. *Applied Sciences*, 2024, 14(2): 898.
- [64] WU S, KHASAHMADI A, KATZ M, et al. CAD-LLM: large language model for CAD generation [C]//New Orleans: The 37th International Conference on Neural Information Processing Systems, 2023.
- [65] LI Y, ZHAO H, JIANG H, et al. Large language models for manufacturing [DB/OL]. <https://arxiv.org/abs/2410.21418>.
- [66] LIU M, KIRBY RENE T D, et al. Chipnemo: domain-adapted llms for chip design [DB/OL]. <https://arxiv.org/abs/2311.00176>.
- [67] YIN Y, WANG Y, XU B, et al. ADO-LLM: analog design bayesian optimization with in-context learning of large language models [C]//New York: the 43rd IEEE/ACM International Conference on Computer Aided Design, 2024.
- [68] LI P. LLM-USO: large language model-based universal sizing optimizer [DB/OL]. <https://arxiv.org/abs/2502.02764>.
- [69] FAKIH M, DHARMAJI R, MOGHADDAS Y, et al. LLM4PLC: harnessing large language models for verifiable programming of plcs in industrial control systems [C]//Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice, 2024: 192–203.
- [70] JEONG Y. Digitalization in production logistics: how AI, digital twins, and simulation are driving the shift from model-based to data-driven approaches [J]. *International Journal of Precision Engineering and Manufacturing-Smart Technology*, 2023, 1(2): 187–200.
- [71] PALMA G, CECCHI G, RIZZO A. Large language models for predictive maintenance in the leather tanning industry: multimodal anomaly detection in compressors [J]. *Electronics*, 2025, 14(10): 2061.
- [72] HADI M U, QURESHI R, SHAH A, et al. A survey on large language models: applications, challenges, limitations, and practical usage [DB/OL]. [TechRxiv TechRxiv.org](https://arxiv.org/abs/2312.10997), 2023.
- [73] ZHOU J, LI C Z, SHI C L. A review of the application and development of numerical and intelligent techniques in materials science [J]. *Material Sciences*, 2025, 15(1): 131–139.
- 周娇, 李朝忠, 石从黎. 数值化与智能化技术在材料科学中的应用与发展综述 [J]. *材料科学*, 2025, 15(1): 131–139.
- [74] MENON D, RANGANATHAN R. A generative approach to materials discovery, design, and optimization [J]. *ACS Omega*, 2022, 7(30): 25958–25973.
- [75] NIYONGABO R A, ARNOLD C, RAND B P, et al. LLM-Prop: predicting the properties of crystalline materials using large language models [J]. *npj Computational Materials*, 2025, 11(1): 186.
- [76] JIANG X, WANG W, TIAN S, et al. Applications of natural language processing and large language models in materials discovery [J]. *npj Computational Materials*, 2025, 11(1): 79.
- [77] DING Q, MIRET S, LIU B. Matexpert: decomposing materials discovery by mimicking human experts [DB/OL]. <https://arxiv.org/abs/2410.21317>.
- [78] JIA S, ZHANG C, FUNG V. Llm4design: autonomous materials discovery with large language models [DB/OL]. <https://arxiv.org/abs/2406.13163>.
- [79] MASLEJ N, FATTORINI L, PERRAULT R, et al. Artificial intelligence index report 2025 [DB/OL]. <https://arxiv.org/abs/2504.07139>.
- [80] AHANGAR M N, FARHAT Z A, SIVANATHAN A. AI trustworthiness in manufacturing: challenges, toolkits, and the path to industry 5.0 [J]. *Sensors*, 2025, 25(14): 4357.
- [81] FU C, LIN H, WANG X, et al. Vita-1.5: towards gpt-4o level real-time vision and speech interaction [DB/OL]. <https://arxiv.org/abs/2501.01957>.
- [82] REDDI V J. Generative AI at the edge: challenges and opportunities: the next phase in AI deployment [J]. *Queue*, 2025, 23(2): 79–137.
- [83] ZHANG M, SHEN X, CAO J, et al. Edgeshard: efficient LLM inference via collaborative edge computing [J]. *IEEE Internet of Things Journal*, 2024, 12(10): 13119–13131.
- [84] LI C, CHEN Q, LI L, et al. Mixed distillation helps smaller language model better reasoning [DB/OL]. <https://arxiv.org/abs/2312.10730>.
- [85] SARKAR S, BABAR M F, HASAN M, et al. LLMs as on-demand customizable service [DB/OL]. <https://arxiv.org/abs/2401.16577>.
- [86] CHEN H, DENG W, YANG S, et al. Towards edge general intelligence via large language models: opportunities and challenges [DB/OL]. <https://arxiv.org/abs/2410.18125>.
- [87] DIMITRAKOPOULOS G, VARGA P, GUTT T, et al. Industry 5.0: research areas and challenges with artificial intelligence and human acceptance [J]. *IEEE Industrial Electronics Magazine*, 2024, 18(4): 43–54.

(责任编辑:孙娟)