

工业高质量数据集研究报告

中国工业互联网研究院

2025 年 09 月

声 明

本报告所有材料和内容的知识产权归中国工业互联网研究院所有（注明是引自其他地方的内容除外），并受法律保护。任何单位和个人未经中国工业互联网研究院授权，不得使用或转载本研究报告中的任何部分。授权后转载、摘编或利用其它方式使用本报告文字或者观点的，应注明“来源：中国工业互联网研究院”。违反上述声明者，本院将追究其相关法律责任。

前 言

人工智能作为引领新一轮科技革命和产业变革的战略性新兴产业技术，是新型工业化的重要推动力量。党的十八大以来，以习近平同志为核心的党中央把发展人工智能提升到战略高度，强调加快发展新一代人工智能是我们赢得全球科技竞争主动权的重要战略抓手，是推动我国科技跨越发展、产业优化升级、生产力整体跃升的重要战略资源。随着大模型技术的迅猛发展与规模化应用，人工智能发展范式从传统以算法模型优化为导向逐步转向以高质量数据集构建为支撑。特别是在工业领域，人工智能、工业机理等模型的构建，对数据集的质量和规模提出更高要求。工业高质量数据集作为支撑各类工业智能模型落地应用的“核心燃料”，是人工智能技术在工业领域应用发展的关键要素。加快构建工业高质量数据集，是我国抢占全球工业智能制高点、应对国际产业竞争的重要抓手，对加速推进新型工业化具有重大战略意义。

当前，工业高质量数据集概念界定不清晰，内涵特征不明确，分级分类标准不完善。不同行业的高质量数据集建设水平和需求存在较大差异。数据获取、专业标注、多模态融合、数据集质量评估、数据集流通与共享等关键环节缺乏系统规划和指引，建设实施过程面临多重挑战，亟须以场景需求为导向，分级分类推进数据集建设，构建高质量数据集建设运营体系，为推动人工智能赋能新型工业化夯实数据基础。

为指引和推动工业领域高质量数据集建设，在工业和信

息化部信息技术发展司指导下，中国工业互联网研究院牵头组织产业界、学术界进行深入研讨交流、凝聚行业共识，联合相关研究机构 and 行业龙头企业开展了深入实践探索，编制了《工业高质量数据集研究报告》。本报告围绕工业高质量数据集的建设与应用展开深入研究，旨在厘清其内涵特征，明晰分级分类标准，探索数据集建设、评估、流通应用全链路有效路径，为业界推进工业高质量数据集建设、评估及应用提供可参考的理论支撑与发展指引。

编写组

指导单位：

工业和信息化部信息技术发展司

组织单位：

中国工业互联网研究院

参编单位（排名不分先后）：

北京大学

北京理工大学

北京航空航天大学

西安交通大学

东北大学

北京邮电大学

北京科技大学

西安电子科技大学

华北电力大学

北京东方国信科技股份有限公司

联通（辽宁）产业互联网有限公司

北京博华信智科技股份有限公司

上海新唐利企业发展集团有限公司

深圳市杉岩数据技术有限公司

北京顺鑫福通大数据集团有限公司

中电科普天科技股份有限公司

矩质算能（上海）科技有限公司

卡奥斯物联科技股份有限公司

南京钢铁集团有限公司

联通数据智能有限公司

西安塔力科技有限公司

蔚来控股有限公司

北京领邦智能装备股份公司

中国电信集团有限公司

广东汉数科技有限公司

沈阳立知科技有限公司

蓝卓数字科技有限公司

无问智行科技有限公司

北京光辉世联科技有限公司

北京工业大数据创新中心

青岛兮易信息技术有限公司

无锡雪浪数制科技有限公司

合肥莱斯博科技有限公司

北京疆来能源管理有限公司

重庆机电智能制造有限公司

目 录

一、工业高质量数据集的内涵意义	1
(一) 相关概念	1
(二) 工业高质量数据集的内涵与特征	2
(三) 工业高质量数据集建设的重大意义	4
二、工业高质量数据集的分级分类	5
(一) 分级体系	5
(二) 分类体系	13
三、工业高质量数据集的建设路径	17
(一) 建设规划	18
(二) 研发实施	21
(三) 应用验证	29
(四) 典型案例	30
四、工业高质量数据集的评估体系	38
(一) 评估框架	39
(二) 评估指标	40
(三) 测评方法	44
五、工业高质量数据集的流通应用	45
(一) 面临问题	46
(二) 解决路径	47
(三) 实现方案	50
参考文献	53

图目录

图 3-1 工业高质量数据集建设流程.....	18
图 3-2 工业高质量数据集建设技术流程.....	22
图 4-1 工业高质量数据集评估框架.....	40
图 4-2 工业高质量数据集评估流程.....	45
图 5-1 工业高质量数据集可信数据空间总体功能架构.....	48
图 5-2 工业高质量数据集可信数据空间部署架构.....	49

表目录

表 2-1 工业高质量数据集分级方法.....	6
表 2-2 工业高质量数据集分类方法.....	13
表 3-1 典型案例：中厚板轧制力实时智能预测数据集建设要求.....	31
表 3-2 典型案例：矿山行业高质量数据集覆盖要素与场景.....	37
表 3-3 典型案例：矿山行业高质量数据集典型场景应用成效.....	38
表 4-1 工业高质量数据集评估指标.....	42

一、工业高质量数据集的内涵意义

工业高质量数据集作为工业领域人工智能技术应用的关键要素，相较原始工业数据或高质量数据集，场景导向性更强、数据价值更高，能更为充分地满足工业机理模型、工业智能模型等工业模型的建模需求。加快建设工业高质量数据集，对我国抢占全球工业智能制高点、应对国际产业竞争、加速推进新型工业化具有重大战略意义。

（一）相关概念

工业数据作为工业领域认知与改造活动的直接产物，是指在工业生产、制造、创新、运营及相关活动中产生、采集、处理和使用的各类数据的总和，它贯穿于工业产品全生命周期以及工业企业运营管理全过程，是工业数字化、智能化转型的核心要素。在国家标准《智能制造 工业数据 分类原则》^[1]中，定义工业数据为“在工业领域中，涉及企业的所有生产活动和服务所产生的数据”。国际标准化组织指出，工业数据“可被视为某种工业流程的产物，受制于一般性的生命周期活动”^[2]，并在 ISO 8000 系列数据质量国际标准中做出定义“工业数据涵盖产品与生命周期流程，包括制造、分销和维护，生命周期流程所使用的设施，数字孪生，产品几何形状、拓扑结构和可视化、技术词典以及零部件目录”^[3]。

狭义的工业数据是指工业设备与软件运行过程中沉淀的数字化信息，包括生产过程中的传感器数据、设备的运行参数与日志等，是工业生产的直接记录。**广义的工业数据**更强调由原始数据经融合加工所沉淀的工业机理与规律认知，

不仅包含狭义定义中的原始数据，还涵盖了对这些原始数据进行清洗、挖掘、分析后形成的与工业生产相关的各类认知成果，例如通过对设备运行数据的长期分析总结出的设备故障预警模型所蕴含的故障发生机理。**工业数据的特征包括领域特定性、多源异构性、高时序性、强关联性等。**领域特定性指数据产生于工业环境，与设备、产品等物理实体和业务流程紧密相关，需结合行业知识才能理解和应用。多源异构性指工业数据来源广泛，涵盖从设备传感器、执行器、工业控制系统、制造执行系统、企业资源计划到供应链管理、客户关系管理等各个层面，来源设备、协议与格式各异，数据模态各异。高时序性指生产流程的连续性与实时性要求工业数据具备精确的时间戳与序列性。强关联性指工业数据点之间存在复杂的关联关系，反映生产流程中的关联关系。

高质量数据集通常指经过采集、加工等数据处理，可直接用于开发和训练人工智能模型，能有效提升模型性能的数据的集合^[4]。数据集质量包括完整性、准确性、现时性、无偏性、相关性、安全性、数据集共享等通用要求，面向人工智能应用的拓展要求还包括，训练数据集的多样性、代表性、可审核性，测试数据集的均衡性、有效性、易用性等^[5]。

（二）工业高质量数据集的内涵与特征

工业是国民经济的主导产业，相较其他产业类型，工业场景众多、机理复杂，包含几何模型、仿真模型、机理模型、算法模型等诸多工业模型。因此，我们认为，**工业高质量数据集**是从研发、生产、供应、销售、服务等全生命周期各环

节产生和采集，经过清洗、标注等专业化处理，用于分析、建模，以及训练工业模型的数据集合。

工业高质量数据集具有完整度高、质量达标、场景明确和时效性强四个显著特征。**完整度高**即数据集基本覆盖各类工况，且包含齐全的说明文档；**质量达标**即数据集历经完整且充分的采集、预处理、标注、增强、合成与质量评测环节，满足工业数据质量要求；**场景明确**即数据集可直接用于工业领域特定应用场景，能有效提升对应场景模型性能；**时效性强**即数据集具备完备的定期更新计划以及自动化更新机制，实时保障数据集在模型训练、微调等阶段的有效性。

工业高质量数据集蕴含工业生产、运营、管理等各环节的关键信息，是工业智能模型实现预测、推理、生成等的核心要素，可直接决定模型的性能上限与应用价值。**一是数据集规模是模型能力突破的“助推器”**。模型参数量与训练数据量的指数级增长，有效增强了新一代算法模型对复杂系统与认知模式的建模和学习能力，提升泛化性能，突破能力瓶颈。特别是人工智能大模型，通过对海量数据的学习，突破小模型“见过才会”的局限，实现了智能涌现与跨场景的高级认知能力。**二是数据集质量是模型训练的“校准仪”**。数据集质量直接决定了模型的有效性，是训练出可靠、精准模型的前提。准确全面的高质量数据集，能引导模型快速、准确掌握本质规律。反之，错误标注、冗余重复或带有偏见的数据，会让模型在错误路径上“越走越远”。**三是多模态数据是模型认知力提升的“催化剂”**。多模态数据通过融合文

本、图像、音频等不同模态的信息，弥补单一模态的局限性，使模型具备更全面的理解能力，从而显著提升模型对复杂任务的处理水平和与真实世界的交互能力。

（三）工业高质量数据集建设的重大意义

工业高质量数据集建设是推进新型工业化的现实要求。

推进新型工业化，关键在于以科技创新引领产业变革，推动制造业高端化、智能化、绿色化发展。工业高质量数据集作为数据基础设施的重要组成部分，是打通“数据-知识-决策”闭环的基础前提。当前，我国制造业数字化转型进入深水区，大量企业面临“有数据、无价值”的困境，根源之一在于数据质量不高、难以支撑复杂分析与智能决策。建设工业高质量数据集，有助于统一数据采集标准、提升数据治理能力、完善数据标注体系，推动工业数据从“碎片化记录”向“结构化资产”转变。系统性推进工业高质量数据集建设，是破解转型瓶颈、提升全要素生产率、实现可持续发展的现实路径，对构建现代化产业体系具有基础性、先导性作用。

工业高质量数据集建设是抢占工业智能制高点的重要抓手。工业高质量数据集是训练和优化工业人工智能模型的“核心燃料”和“知识底座”。在全球工业智能竞争白热化阶段，科技强国正加速布局覆盖关键工艺、核心设备及典型制造场景的高质量数据资源体系。率先建成系统化数据集的国家，将牢牢掌控工业知识沉淀、复用与创新的主动权，主导智能检测、预测性维护、柔性生产等核心环节的技术突破与标准制定。深化工业高质量数据集建设，是打通智能制造

全链条、破解数据孤岛、激活 AI 赋能的关键路径，是驱动制造业智能化升级的坚实支撑。

工业高质量数据集建设是应对全球产业竞争的战略要求。当前，高端制造领域的国际竞争，正逐步演变为基于数据驱动的产业生态体系竞争。工业高质量数据集不仅承载着核心工艺参数、设备运行规律和生产优化经验，更是推动产业链协同升级、价值链高端跃迁的关键要素。如果缺乏自主可控的高质量数据资源，我国在高端工业软件、智能控制系统等领域将始终面临“不敢用”国外模型又“用不好”国产工具的双重困境，陷入“数据依附”风险，导致技术升级受制于人。唯有加快建设自主、安全、可信的工业高质量数据集，才能从根本上突破“卡脖子”难题，增强产业链供应链的韧性与安全性，构筑面向全球竞争的新优势。

二、工业高质量数据集的分级分类

工业场景存在从“设备、产线、工厂、企业”到“产业生态”的多重层级，不同层级数据集的融合程度与决策价值差异显著。为有效释放数据价值潜能、提升工业智能决策水平，应遵循分级分类原则推进工业高质量数据集建设。

（一）分级体系

参照工业互联网技术架构体系，梳理工业数据集的数据来源、数据类型和应用场景，形成工业高质量数据集五级架构，即设备级（L1）、产线级（L2）、工厂级（L3）、企业级（L4）和生态级（L5）。层级表征数据集赋能应用场景的广度，层级越高，数据聚合程度越高，应用场景范围越广，

决策价值密度越高，数据支撑产生的经济价值越大^[6,7]。具体分级方法如表 2-1 所示。

表 2-1 工业高质量数据集分级方法

归属层级	数据集级别	具体定义
基础层	设备级（L1）	指源自物理设备或生产单元，获取设备运行、制造过程状态或工装类的高质量数据集，可应用于设备或生产单元故障诊断、工艺优化和寿命预测等应用场景。
	产线级（L2）	指源自产线或车间，获取控制系统、网络参数类的高质量数据集，可应用于产线异常检测、能耗预测、产线参数寻优与协同控制等应用场景。
核心层	工厂级（L3）	指源自工厂的生产制造环节，获取工厂级管理数据、工艺数据、计划数据、安全环保等类高质量数据集，可应用于生产调度动态优化、安全风险智能监控与预警等应用场景。
	企业级（L4）	指源自企业的经营管理和运维服务环节，获取企业级人力资源、技术、成本、财务及供应链等类高质量数据集，可应用于经营利润预测、智能排产、供应链优化、财务管理等应用场景。
生态层	生态级（L5）	指源自产业生态，获取协同策略和协同管理类高质量数据集，可应用于供需智能匹配、产业风险监测预警、区域产业大脑、区域能源与碳排放协同管理等场景。

设备与产线级作为基础层级，从业务执行的最小单元出发，汇聚设备运行、工艺参数等高质量数据，实现故障诊断、能耗优化等局部效率提升，为数字化转型奠定坚实的数据基石。**工厂与企业级**作为核心层级，聚合生产计划、成本投入、经营管理等数据，驱动生产调度、资源配置等核心业务流程的系统性优化，提升企业整体运营水平。**产业生态级**作为生态层级，通过聚合跨主体、跨领域的的数据，赋能供应链柔性、韧性、产品全生命周期绿色低碳、产业风险监测预警等应用，创造生态协同的增量价值。

1.设备级高质量数据集

从**数据来源**分析，主要源自传感器、仪器仪表、数控装备、智能机床、工业机器人、仓储设备、加工单元、装配单元、条码标签等感知终端。

从**数据类型**分析，主要涉及设备数据、状态数据和工装数据。**设备数据**是指描述设备基本信息，反映设备运行工况和健康状态，影响设备性能和运行状态的环境信息等数据。**状态数据**是指在制造过程、控制过程中反映各种不同状态的物理量信号，如刀具磨破损、机床热变形、切削负荷、振动变形等数据。**工装数据**是指描述工装基本信息，以及采用条码标签等采集的器件、辅具环境、工况数值等数据。

从**数据特征**分析，主要存在实时性强、动态变化大、物理意义明确、多源异构、价值密度低等特点。**实时性强**是指数据产生与传输速度快，需快速响应。**动态变化大**是指设备状态频繁切换，数据模式快速演变。**物理意义明确**是指数据与设备物理状态直接相关，具有明确的业务解释。**多源异构**是指数据来自不同传感器、系统和设备，格式和标准多样。**价值密度低**是指海量数据中真正有价值的信息占比低，需要高效挖掘。

从**应用场景**分析，主要包括设备故障诊断、生产工艺优化、寿命预测、预测性维护等场景。例如，国家电力投资集团有限公司构建了光伏逆变器健康诊断数据集，涵盖光伏逆变器多维度实时运行数据、设备信息及环境数据，可用于开发智能诊断模型，实现逆变器故障的精准诊断与预测性维护，缩短故障修复时间，减少停机损失，降低运维成本 10%，提

升电站运营盈利能力。

2.产线级高质量数据集

从**数据来源**分析，产线级数据主要源自可编程逻辑控制器（PLC）、分布式控制系统（DCS）、人机界面（HMI）、智能控制器、网关、远程终端单元（RTU）、数据采集与监视控制系统（SCADA）等控制与监控系统，承担工艺执行、数据采集与状态监控等功能。

从**数据类型**分析，主要涉及经过处理的设备级数据、控制系统数据和网络参数数据。**控制系统数据**是指 PLC 包含的伺服参数、G 代码、系统操作日志等。**网络参数数据**是指通过联网系统采集的控制层网络参数、信号等。

从**数据特征**分析主要表现为实时性强、数据体量大、数据质量低、多源异构、关联度高等特点。**实时性强**是指数据产生与传输速度快，需快速响应。**数据体量大**是指单条产线每秒可产生数千条数据点，车间层数据量呈指数级增长。**数据质量低**是指设备通信与网络产生的噪声与异常值较多。**多源异构**是指数据来源多样，格式复杂、通信协议多样。**关联度高**是指数据间存在时空关联，如某一设备的伺服数据可能影响另一设备的相关参数。

从**应用场景**分析主要包括产线异常检测、能耗预测、产线参数寻优与协同控制等场景。例如北京首钢股份有限公司构建热轧带钢控制模型训练数据集，该数据集包含热轧产线工艺参数、过程数据、设备状态数据及仿真数据，支撑公司研发“粗轧-精轧-层冷”全流程过程控制在线仿真系统，实

现生产过程核心模型与关键参数的静态/动态精准模拟和实时调优，提升控制模型在线优化迭代效率与安全性。

3.工厂级高质量数据集

从**数据来源**分析，主要源自计算机辅助制造系统、计算机辅助设计系统、计算机辅助工程系统、计算机辅助工艺设计、制造执行系统、设备管理系统、质量管理系统、仓库管理系统、实验室管理系统、计量系统、健康安全与环境管理系统、产品生命周期管理系统等。

从**数据类型**分析，主要涉及经过处理的产线级数据、管理数据、工艺数据、技术数据、作程数据、物料数据、计划数据、安全环保数据、质量数据和计量数据。**管理数据**是指描述人员、部门、岗位、班组的基本信息，人工成本、技能等级等相关属性，以及车间地理位置、设备布局、人机交互硬件等数据。**工艺数据**是指在产品工艺设计过程中产生的二维、三维工艺路线及装配作业指导书等与产品工艺直接相关的数据信息，以及 BOM 信息、零部件信息、数控程序等数据。**技术数据**是指企业根据自身特点开展的产品研发、技术资料、工艺资料等设计数据。**作程数据**是指工位考勤状态，派工、完工信息等作业过程数据。**物料数据**是指产成品的包装运输，在制品的跟踪，工装的申请、使用、报废、出入库等物料数据。**计划数据**是指根据生产计划下达的生产指令、机台计划、生产备料等计划数据。**安全环保类数据**是指涉及企业生产安全和环保相关的数据。**工厂质量数据**是指报检、检验、试验、质量指导书、返工返修等质量数据。**计量数据**

是指检验检测数据。

从**数据特征**分析，主要存在流程关联性强、多模态异构、时空复杂度高、业务逻辑深度耦合等特点。**流程关联性强**是指数据围绕产品设计、工艺、制造、质量等环节产生，各系统数据高度关联。**多模态异构**是指数据来源庞杂，包括CAD/CAE的设计模型文件、MES的结构化报表、HSE的非结构化文本报告、质量检测图像等。**时空复杂度高**是指数据在时间和空间维度上具有复杂关系，如生产批次的顺序、物料在仓库和产线的流转路径等。**业务逻辑深度耦合**是指数据与工厂排产、工艺规程、质量标准等核心业务逻辑紧密绑定，数据价值高度依赖于业务场景解释。

从**应用场景**分析，主要包括车间资源协同分配、供应链协同调度、生产智能排产、安全风险监控预警、物料搬运仓储优化与排放监控控制等。例如，国家电力投资集团有限公司构建光储充一体化综合能源系统高质量数据集，涵盖分布式光伏设备参数及出力、储能系统状态监测、充电桩负荷交互、设备资料图纸、运维工单等数据，可用于光储充一体化综合能源管理系统的智能运维与精准调控，实现分布式光伏发电精准预测与运维评估，提升综合能源效率和“源-网-荷-储”协同水平，提高能源利用效率，降低运营成本并保障电网安全稳定。

4.企业级高质量数据集

从**数据来源**分析，主要源自企业资产管理系统、企业资源规划系统（ERP）、供应链管理系统、物流信息系统、财

务系统、人力资源系统、客户关系管理系统（CRM）、网络安全系统和实验室管理系统等。

从**数据类型**分析，主要包括经过处理的工厂级数据、人力资源数据、财务数据、采购销售数据、成本数据、供应链数据、调度数据、生产计划数据、质量数据和售后服务数据。**人力资源数据**是指职员基本信息如人员薪酬、合同、绩效、奖惩、考勤、加班、年休假、离退休等跟职员有关的人力资源数据。**财务数据**是指描述会计科目、薪酬管理、账户、合同等的有关财务数据。**采购销售数据**是指采购、销售、市场、客户等市场相关数据。**成本数据**是指工艺员、部门在生产制造中涉及的设计费用、加工费用、材料费用等成本数据。**供应链数据**是指融合采购、供应商等一体的供应链数据。**调度数据**是指能源产耗数据、调度指令等调度数据。**生产计划数据**是指与生产相关的项目计划、需求计划、协作计划、工艺变更等生产计划数据。**企业质量数据**是指采购检验信息、第三方检验，不合格品审理、客户反馈和索赔等质量数据。**售后服务数据**是指售后服务信息反馈、服务处理、改进等售后服务数据等。

从**数据特征**分析，主要存在强业务规则驱动、强战略导向、全域覆盖价值链和高安全合规要求等特点。**强业务规则驱动**是指数据严格遵循财务准则、人力资源政策、销售合同条款等企业规章制度，对准确性和一致性要求高。**强战略导向**是指数据多为跨部门、跨周期的汇总统计结果，服务于企业高层战略决策，如财务报表、市场占有率等。**全域覆盖价**

价值链是指数据贯穿从供应商、生产、客户的整个价值链，涉及企业内外部多个协作方。**高安全合规要求**是指包含大量财务、人力、客户隐私等敏感信息，必须满足网络安全法、数据安全法等法律法规要求。

从**应用场景**分析，主要包括经营利润预测、智能财务管理、人员技能任务匹配、经营风险监控与战略决策支持等场景。例如，中国联通构建高质量经营决策推理数据集。该数据集面向智慧经分场景，汇聚企业全量经营分析场景真实数据，包含具有意图实体标签的经营管理数据（5.2 亿 tokens），用于支撑市场策略模拟推演与风险预判，提升大模型经营分析语义理解能力（准确率从 81%提升至 98%），提高企业经营决策效率。

5.生态级高质量数据集

从**数据来源**分析，主要源自客户关系管理、远程运维、协同设计研发、协同采购、协同生产、资源共享、供需对接、共性应用和供应链协同等。

从**数据类型**分析，主要包括协同策略数据和协同管理数据。**协同策略数据**是指协同计划分解、协同目标规则、资源能力目录、任务调度、风险与应急策略等数据。**协同管理数据**是指协同组织信息、协同流程定义、协同任务与进度、协同资源调度、协同绩效与评估、协同审计与合规等数据。

从**数据特征**分析，主要表现为跨主体、产业化和合作博弈等特点。**跨主体**是指数据的所有权和使用权分布在不同的法律实体。**产业化**是指数据应用旨在提升整个产业链的效率

和竞争力。**合作博弈**是指各合作参与方有强烈的价值主张，数据共享面临信任、定价和利益分配挑战。

从**应用场景**分析，主要包括供需智能匹配、产业风险监测预警、区域产业大脑、区域能源与碳排放协同管理、设计协同等场景。例如，新疆汇通互联公司构建辐射“一带一路”的多源物流体系数据服务体系。该体系基于区块链技术，包含1个专用数据集、7个功能数据集和1个数据模拟训练数据集，囊括煤炭、煤化工、农业、新能源等12个行业。该数据集体系可用于构建交通流量、事故风险、价格指数等预测模型，以多源数据融合促进多环节贯通、多主体协同和跨行业赋能，降低整体物流成本10%，累积服务企业超1000家。

（二）分类体系

按照产品全生命周期（研发设计、生产制造、运维服务、经营管理）对工业数据集进行分类，具体分类方法如表2-2所示。

表 2-2 工业高质量数据集分类方法

数据集类别	具体定义
研发设计	指源自产品研发设计全流程，涵盖从需求输入到样品定型验证的高质量数据集，主要来源于产品生命周期管理、计算机辅助设计、计算机辅助工艺过程设计等研发设计系统，可应用于产品结构优化、工艺方案迭代等应用场景。
生产制造	指源自产品生产执行全流程，涵盖从原料投入到成品产出的高质量数据集，主要来源于制造执行系统、可编程逻辑控制器、分布式控制系统、监控与数据采集及生产管理系统等，可应用于生产进度调度、产品质量管控等应用场景。
运维服务	指源自设备全生命周期维护过程，涵盖从客户使用反馈到产品报废回收的高质量数据集，主要来源于客户关系管理、远程运维系统、售后管理系统、设备传感器等运维服务系统，可应用于产品故障预警、客户满意度提升、产品迭代

数据集类别	具体定义
	改进、资源循环利用等应用场景。
经营管理	指源自企业整体经营运营过程，涵盖采购、销售、财务、人力资源、供应链、网络安全等核心管理环节的高质量数据集，主要来源于企业资源计划、客户关系管理、供应链管理、财务系统、人力资源系统、网络安全系统等，可应用于企业成本控制、供应链优化等应用场景。

1.研发设计类高质量数据集

从**数据类型**分析，主要涉及设计规划数据、概念设计数据、技术设计数据、详细设计数据、物料清单数据、设计仿真数据、工艺仿真数据和样品制作数据。**设计规划数据**指围绕设计目标、技术路线、资源分配、时间节点制定的规划类信息。**概念设计数据**指基于市场与客户需求形成的产品初步构想、功能框架、外观草图、核心原理方案等早期设计数据。**技术设计数据**指明确产品技术参数、性能指标、核心部件选型、接口标准、关键工艺要求等数据。**详细设计数据**指对产品各部件、组件的尺寸、材质、加工精度、装配关系、公差要求等进行细化描述的数据。**物料清单数据**指描述产品组成结构，包含各物料规格、数量、层级关系、供应商信息等数据。**设计仿真数据**指通过仿真软件对产品结构强度、力学性能、运行工况、环境适应性等进行模拟计算产生的数据。**工艺仿真数据**指对产品加工工艺、装配流程、工装适配性等进行仿真模拟产生的数据。**样品制作数据**指样品生产过程中的加工记录、检验结果、尺寸测量数据、装配调试反馈等数据。

2.生产制造类高质量数据集

从**数据类型**分析，主要涉及生产计划管理数据、生产技术准备数据、生产订单管理数据、生产领料管理数据、生产

过程管理数据、生产外协管理数据、生产质量管理数据、生产完工管理数据、生产仿真管理数据、设备管理数据、原料配比数据、生产外协管理数据、生产能耗数据、生产安全数据、环保排放数据和运行人员数据。**生产计划管理数据**指基于订单需求或市场预测制定的生产任务分配、产能规划、生产批次安排、交付周期规划等数据。**生产技术准备数据**指生产前的工艺文件编制、工装夹具调试、设备参数设定、技术交底记录等数据。**生产订单管理数据**指记录生产订单编号、产品型号、生产数量、交付日期等数据。**生产领料管理数据**指生产过程中物料领用的申请、审批、发放记录、物料消耗统计等数据。**生产过程管理数据**指生产各工序的加工记录、工序流转状态、工时统计、异常停机原因、过程调整记录等数据。**生产外协管理数据**指委托外部单位完成部分工序加工的外协订单、质量要求、交付进度、费用结算、外协质量验收等数据。**生产质量管理数据**指生产过程中的原料检验报告、工序检验记录、成品检测数据、不合格品判定与返工返修记录等数据。**生产完工管理数据**指产品完成生产后的入库记录、成品检验总报告、生产产能统计、生产能耗汇总、生产异常总结等数据。**生产仿真管理数据**指对生产流程优化、设备布局调整、产能匹配验证等进行仿真产生的数据。**设备管理数据**指生产设备的实时运行状态、故障报警信息、维修记录、保养计划与执行情况等数据。**原料配比数据**指流程工业生产中各类原料的混合比例、投放量、配比调整记录等数据。**工艺运行数据**指生产过程中关键工艺参数的实时监测与历史

记录数据。**生产能耗数据**指生产过程中水电汽等能源的消耗记录、能耗峰值、能耗分析报告等数据。**生产安全数据**指生产现场的安全检查记录、隐患排查结果、安全培训记录、安全事故处理报告等数据。**环保排放数据**指生产过程中废水、废气、废渣等污染物的排放指标监测数据、环保处理设施运行状态等数据。**运行人员数据**指生产现场操作人员的考勤记录、岗位分配、操作流程执行记录、技能培训情况等数据。

3.运维服务类高质量数据集

从**数据类型**分析，主要涉及维修服务数据、维护服务数据、增值服务数据、报废/回收数据、客户满意度数据、产品使用反馈数据。**维修服务数据**指产品售后出现故障后的维修申请单、故障诊断结果、维修方案制定、维修过程记录、维修费用结算、维修效果验证等数据。**维护服务数据**指产品定期保养计划、预防性维护执行记录、维护耗材使用统计、维护后设备性能检测数据等数据。**增值服务数据**指为客户提供的超出基础维修维护的服务相关的服务记录、客户需求反馈、服务评价等数据。**报废/回收数据**指产品达到使用年限或因故障报废后的回收计划、回收过程记录、拆解分类结果、资源再利用评估报告等数据。**客户满意度数据**指通过问卷调研、电话回访、在线反馈等方式收集的客户对产品性能、服务响应速度、维修质量等方面的满意度评价数据。**产品使用反馈数据**指客户在产品日常使用中发现的问题、功能优化建议、特殊使用场景描述等数据。

4.经营管理类高质量数据集

从**数据类型**分析，主要涉及采购管理类数据、销售管理类数据，财务类数据、人力资源类数据、供应链数据和网络安全数据。**采购管理类数据**包括采购计划数据、采购方案数据、采购订单数据、采购合同数据、采购执行过程数据、供方管理数据、退/换货管理数据。**销售管理类数据**包括市场预测数据、客户管理数据、销售计划数据、销售报价数据、销售订单数据、销售合同数据、发票管理数据、应收款管理数据、退/换货管理数据。**财务类数据**包括成本数据、会计科目数据、薪酬管理数据、账户数据、合同财务数据。**人力资源类数据**包括职员基本信息、薪酬数据、合同数据、绩效数据、奖惩数据、考勤数据、年休假数据、离退休数据。**供应链数据**指融合采购、生产、库存、物流、销售等环节的供应链全流程数据，包括库存水平实时统计、物流运输路线与时效记录、供应链节点协同进度、供应链风险预警信息等。**网络安全数据**指企业网络安全管理平台、态势感知系统采集的安全漏洞扫描结果、网络攻击行为记录、安全防护措施执行情况、数据泄露风险预警等数据。

三、工业高质量数据集的建设路径

工业高质量数据集建设应以研发设计、生产制造等场景需求为导向，从场景、技术、管理等维度系统规划，明确建设流程、技术路径，形成规范化的工业高质量数据集建设实施路径，打造匹配工业智能化需求的高质量数据集。建设流程如图 3-1 所示，涵盖建设规划、研发实施、应用验证三大核心环节，实现数据集从前期规划到持续优化的完整闭环。

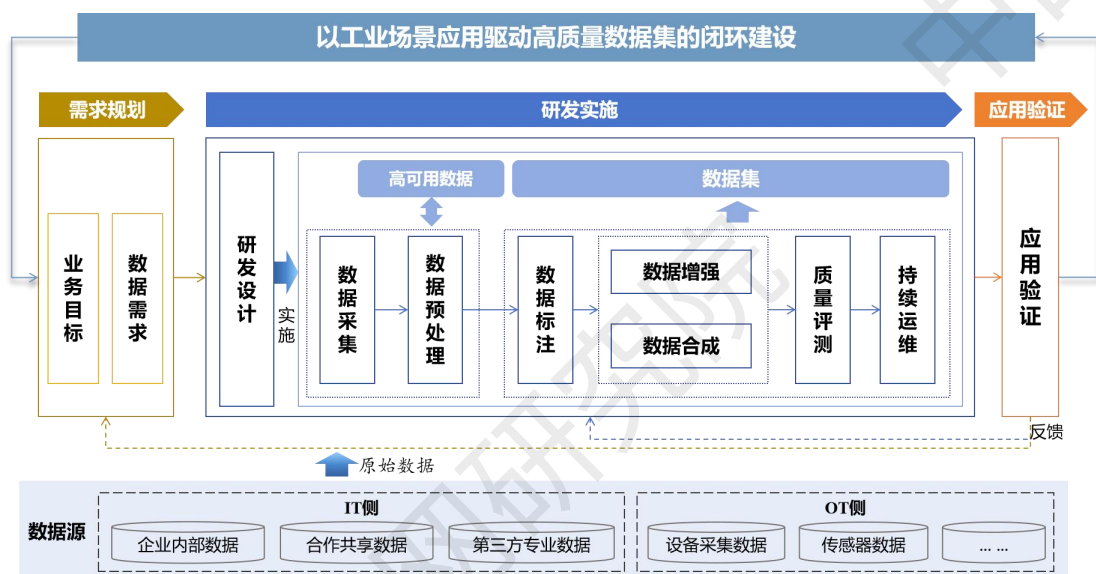


图 3-1 工业高质量数据集建设流程

（一）建设规划

建设规划作为数据集建设的前置环节，是决定数据集能否贴合业务场景、释放实际价值的“源头锚点”。建设规划需要具体考量以下三个方面的因素：**一是要有效规避数据集建设资源错配**。工业数据具有规模庞大、维度复杂、关联性强、采集成本高等特征。通过建设规划，能够精准定义数据集建设目标，为数据集的范围、维度、精度及更新频率等核心参数划定边界，避免陷入“为建而建”的资源浪费，保障资源配置效率；**二是要保障数据集建设适配工业场景差异性**。不同业务目标对数据需求差异显著，通过构建“业务痛点-数据需求-应用场景”映射关系，确保数据集建设贴合真实场景需求，避免采集冗余数据或遗漏关键特征，从源头防止数据集与业务“两张皮”。例如，针对焊接车间飞溅缺陷问题，需明确缺陷图像的采集角度、标注标准及关联工艺参数，确保数据集服务于缺陷识别模型训练；**三是要为数据集建设全流程提供统一目标基准**。数据采集阶段明确采集对象和方式，

标注阶段定义数据标注规范，应用迭代阶段根据业务目标变化动态调整数据维度。

以场景应用为导向的建设规划，通过构建“业务目标-数据需求-应用效果”的闭环链路，为数据集建设提供精准的方向指引。建设规划具体步骤如下。

一是锚定业务目标，反向推导数据集建设要求。以“场景具象化、环节清晰化、指标可量化”为原则，聚焦工业领域的生产工艺参数优化、产品缺陷检测、生产过程机理建模等特定应用场景，明确具体业务目标，包括业务覆盖的关键环节、验收指标等，反向推导各场景数据集建设目标。例如，在产品缺陷检测场景中，需界定检测覆盖的产品类型、缺陷类别，并明确缺陷识别准确率、误判率、检测效率提升幅度等可量化的验收指标。

二是形成数据集需求，明确数据集建设目标。以业务目标为核心牵引开展需求分析，厘清数据集需求规格，具体包括：定义数据集特征维度、数据类型、样本规模、更新机制、质量基线等关键参数；明确交付数据集的格式标准、存储形态与验证方式，确保能够反向对应业务目标的支撑价值，形成可追溯、可量化、可执行的《数据集建设需求清单》，为数据集研发提供明确依据。

三是收集多维数据源，构建数据源匹配矩阵。以《数据集建设需求清单》为基准，构建“需求-数据源”匹配矩阵，系统整合 IT 侧与 OT 侧数据资源，从字段完整性、时效性、样本充足性等维度综合评估数据源。针对评估不达标项，制

针对性优化策略，确保数据源覆盖需求清单的各项要求，形成“需求拆解→数据源匹配→综合评估→优化补全”全流程体系。例如，在风电设备故障诊断数据集建设中，针对风机罕见故障类型（如齿轮断裂）样本不足问题，通过回溯历史设备维修日志，提取齿轮断裂相关历史数据，补齐样本量。

四是多维度论证可行性，夯实数据集建设基础。以数据集建设项目落地可行性为目标，围绕技术可行性、成本可行性等核心维度，论证数据集建设的可行性。**技术可行性论证层面**，针对数据采集、标注、增强等研发实施环节，评估技术难度与潜在风险。**成本可行性论证层面**，构建成本测算模型，测算数据集建设全流程成本构成。基于技术与成本可行性论证结果，建立数据集需求清单动态修正机制，确保需求清单与技术能力、成本预算相适配。

五是构建系统化工作计划，保障建设实施全程可控。以数据集建设“过程可控、结果达标”为目标，遵循“责任可追溯、进度可管控、风险可预判”原则，制定工作计划框架，明确实施路径与责任边界，形成覆盖数据集建设全周期的行动指南，包括：组建专项团队，明确权责分工，夯实“责任可追溯”执行基础；制定精细化进度计划，构建“进度可管控”阶段机制；围绕技术落地、进度延误与成本超支等风险，制定分级分类风险预案，实现“风险可预判”前置防控。

缺乏科学的建设规划，将导致数据集建设偏离业务实际，引发多重风险。**一方面，数据集建设易陷入“重规模轻实效”误区。**例如，盲目采集海量冗余数据，导致数据存储、清洗、

标注等环节资源浪费；因关键特征缺失，数据集无法支撑核心业务场景的模型训练，导致“数据建成即无用”。另一方面，数据集往往面临“建成即落后”困境。例如，因数据维度与业务迭代脱节而被闲置，因标注标准不统一导致数据集跨场景复用率低下。

（二）研发实施

研发实施阶段是工业高质量数据集从“概念”走向“实用”的“炼金过程”。围绕数据集建设需求清单，针对性制定技术路径与方案，通过多源异构数据采集、数据预处理、数据标注、增强与合成等技术手段，将分散无序的“原始素材”转化为可支撑工业场景应用的高质量、高可用工业数据集，实现“从0到1”质的飞跃。同时，经质量评测验证的数据集可突破单一场景限制，实现跨域复用，推动数据集价值从单点向规模化应用辐射，完成“从1到N”的持续价值延伸。

工业高质量数据集的构建需包括研发设计、数据采集、数据预处理、数据标注、数据增强、数据合成、质量评测及持续运维八个核心环节，技术流程如图3-2所示，实现从碎片化、低价值工业原始数据到标准化、高质量数据集的转化。



图 3-2 工业高质量数据集建设技术流程

一是研发设计指引实施技术路径。作为数据集研建的顶层规划环节，研发设计需紧扣数据集需求，制定覆盖数据采集、处理、标注等全流程的技术方案与标准规范，为研发实施过程提供行动指南。具体包括：明确工业数据采集接口协议与格式要求、数据处理流程，规划数据存储架构及灾备策略，制定融合领域知识的数据标注技术规范、增强策略等，保障数据集建设按既定技术路径有序推进。

二是数据采集提供工业数据“原料”。以“全量覆盖、质量可控、场景适配”为原则，按照既定数据采集方案，通过 IT 与 OT 两侧协同，实现原始数据全面采集。IT 侧对企业内部信息系统、合作共享数据，以及供应链上下游等外部数据，按数据源类型实施分类采集。OT 侧依托工业传感器、边缘计算设备等硬件设施，多维度实时采集生产环境参数等原始数据。同时，在采集流程中预设数据清洗规则，最终形成覆盖全需求场景、质量达标的原始数据集。

示例：某数控机床生产车间数据采集

为提升数控机床生产车间的生产线智能化水平，以数控机床数据采集为核心，按照数据源实施分类数据采集，实现生产全流程的数据贯通。

一是针对企业内部信息系统，通过 API 接口方式对接 ERP、MES 等系统，定时采集生产订单数据（如订单数量、交付周期）、质量检测数据（如零件尺寸合格率）。

二是针对物流企业等外部系统，采用 API 接口方式或约定的数据交换协议，获取零部件运输轨迹等物流数据。

三是针对生产车间现场的数据采集，首先，硬件方面主要借助车间安装的振动传感器、视觉传感器等设备，以及边缘计算设备；其次，通信协议适配方面，针对不同设备的通信接口与协议完成协议转换适配；最后，根据参数变化特性设置采集频率（如关键动态参数每 100ms 采集一次），按照采集频率采集主轴转速、扭矩、工件定位精度等数据。其中，传感器负责温度、振动等数据感知，边缘计算设备通过预设“阈值过滤+趋势判断”清洗规则实现数据本地化预处理。

通过采集覆盖“订单-生产-质检-物流”全流程数据，为后续数控机床故障预警模型训练提供完整的原始数据支撑。

三是数据预处理提供原始“高可用数据”。数据预处理是针对工业原始数据中普遍存在的噪声干扰、信息缺失等质量问题，通过实施数据清洗、转换和整合等操作，从根源提升数据可用性。数据清洗阶段采用规则过滤、异常检测等手段剔除噪声、补全缺失值，确保数据样本的准确性与完整性。数据转换阶段通过对原始数据实施格式转换、多模态数据语义映射与对齐，结合量纲统一、数据归一化或标准化处理，

解决多源数据异构性、数值尺度差异等问题，实现工业数据的标准化整合。

示例：某数控机床车间数据预处理

基于各类传感器采集的数控机床车间原始数据存在数据缺失、噪声干扰、数据不一致等问题，亟需进一步通过数据预处理提升数据质量，关键环节如下。

1.数据清洗：对于温度等数据缺失问题，利用线性插值法进行填补；针对振动数据中的噪声问题，采用小波去噪等技术完成噪声过滤。

2.数据转换：利用数据转换脚本，将不同机床的位移、速度等数据格式统一为 CSV 格式，并将单位全部转换为国际标准单位（如将英寸转换为毫米）。通过归一化处理（如采用 min - max 归一化方法，将数据映射到[0, 1]区间），消除量程差异。进一步对数据进行标准化处理，保证不同机床、不同工况下的振动数据处于同一尺度，有助于模型更有效地学习数据特征。

3.特征提取：例如，从振动、电流等时序数据中提取均值、方差、峰值、峰峰值、均方根值等时域特征。

四是数据标注赋予工业数据“可理解性”。数据标注是通过融合工业机理与领域知识，依托专业标注工具对文本、图像、视频、时序信号等原始数据添加标签，使数据可被机器理解和使用。针对工业数据集需求，按照既定标注方案，对不同模态数据实施图像边界框标注、文本分类等差异化标注流程，形成场景适配的标注数据集。

示例：面向数控机床加工过程的时序振动数据标注

为提升数控机床加工过程的故障预警能力，针对主轴系统的时序振动数据，明确标注精度要求等，选择适配的标注工具与协作模式（如人机协同），开展数据标注工作，关键路径如下。

示例：面向数控机床加工过程的时序振动数据标注

- 1.标注数据源：某型号的数控机床主轴在不同加工工况下的振动数据，同时，确保数据已经过预处理，质量达标。
- 2.数据标签设计：结合机床运行机理，设计“启动加速”“稳定加工”“减速停机”“异常振动”四类时序标签，对每类标签定义阈值范围，并关联每段时序数据对应的工艺参数（如主轴转速）。
- 3.数据标注实施：**一是**采用时序分段标注与特征关联标注，结合自动化标注平台完成预标注，由具备机床运维经验的工程师进行分段修正标注。**二是**标注质量核验，通过“标注员交叉校验+算法自动质检+领域专家抽样审核（抽样比例 $\geq 5\%$ ）”多层核验，确保标注数据集质量，形成“主轴振动-工况关联标注数据集”。
- 4.标注数据集应用：标注数据集可用于训练故障预警模型，提升模型对“轴承磨损”“主轴不对中”等典型故障的提前识别能力。

五是数据增强缓解数据稀缺难题。数据增强通过对原始数据进行合理变换与扩展，在不改变数据核心特征的前提下扩充数据集规模，有效解决工业场景故障样本少、极端工况数据不足等难题，推动数据集从“小而精”向“大而全”演进。增强方式主要包括以下三种：**面向领域特性的可控增强**，结合工业场景数据的物理约束与场景关联性定制化数据增强策略。例如，对轴承缺陷图像采用角度受限的旋转增强，避免过度变换导致缺陷形态失真；**基于模型反馈的智能增强**，通过构建“识别-增强”闭环机制提升数据增强的精准性。例如，先训练轻量教师模型识别难例与偏差，再对难例区域做亮度调整、局部遮挡等针对性增强；**基于质量感知的增强过**

滤，对增强后的样本进行置信度评分与误差检测，剔除低置信样本与高误差样本，形成“增强-评估-淘汰”闭环机制，确保增强数据的质量。

示例：汽车零部件表面缺陷检测图像数据增强

某汽车零部件制造商通过工业相机采集发动机缸体、轮毂等零部件表面图像，用于缺陷检测。原始数据集含 3000 张图像，其中缺陷样本仅占 20%（600 张），正常样本占 80%（2400 张），存在严重的类别不均衡问题。为平衡数据分布，提升模型对缺陷的识别能力，采用数据增强策略如下。

1.数据变换快速生成新数据样本

通过对原始数据实施几何变换、颜色调整、噪声添加等操作，生成新数据样本，提升数据集多样性。

（1）几何变换：对缺陷图像实施随机旋转（如 $\pm 15^\circ$ ）、水平和垂直翻转、缩放（如 0.8-1.2 倍）等操作，生成新缺陷图像。

（2）颜色变换：调整图像亮度（如 $\pm 10\%$ ）、对比度（如 $\pm 15\%$ ）、饱和度（如 $\pm 10\%$ ）等，增加图像多样性。

（3）噪声添加：例如，向图像中添加高斯噪声，模拟工业现场光照不均、设备抖动等干扰因素，增强模型鲁棒性。

2.生成对抗网络（GAN）生成高质量复杂数据样本

针对某些细微、罕见的缺陷，难以大量采集，GAN 通过学习真实数据分布，生成全新且逼真的缺陷图像，弥补原始数据不足。

（1）数据准备：将原始 3000 张图像按 7:3 比例划分为训练集（2100 张）与测试集（900 张），训练集中缺陷样本 420 张，正常样本 1680 张。对图像进行归一化处理，将像素值缩放到 $[-1, 1]$ 区间，适配模型输入要求。

（2）模型训练：生成器与判别器交替训练。生成器先输入随机噪声生成

示例：汽车零部件表面缺陷检测图像数据增强

图像，判别器对生成图像与真实图像进行判别，根据判别结果反向传播更新判别器参数；随后固定判别器，生成器根据判别器反馈更新自身参数，使生成图像更难被判别器识破。在训练发动机缸体缺陷图像时，经过多轮迭代，生成器生成的缺陷图像纹理、色泽与真实缺陷愈发相似，判别器准确率提升至 85%。

（3）数据增强：训练完成后，使用生成器对原始缺陷样本进行扩充。针对轮毂表面腐蚀缺陷，生成器根据学习到的真实腐蚀缺陷特征，生成新的腐蚀缺陷图像，且在缺陷形状、大小、分布密度等方面呈现丰富多样性。

六是数据合成填补数据供给缺口。数据合成利用生成式 AI 等技术模拟生成与真实场景数据特征高度吻合的模拟样本，弥补数据供给总量不足、结构性缺失及获取成本高昂等短板，提升数据集场景覆盖度与特征完备性。在生成阶段，明确待合成数据的类型、规模及特征约束，采用生成对抗网络（GAN）、扩散模型等生成式模型定向训练，产出高保真模拟样本，确保其在特征分布等方面与真实数据保持对齐。在验证阶段，将合成数据与真实数据按多比例混合进行测试，检验合成数据对模型性能的提升效果及是否存在偏差，确保合格的合成数据纳入主数据集。

示例：汽车零部件缺陷图像数据合成

某汽车零部件企业生产发动机缸体，真实缺陷样本仅 300 张（含裂纹 120 张、气孔 80 张、砂眼 100 张），导致缺陷检测模型准确率不足，需通过 GAN 合成 1500 张高质量缺陷图像，弥补数据集不足，关键路径如下。

1.数据准备：对原始 300 张原始缺陷图像经过清洗、标注、增强等处理后，输出 560 张带标注的标准化图像缺陷数据集，保证图像无黑边、无像素溢出，

示例：汽车零部件缺陷图像数据合成

数据分布均匀，形成适配 GAN 模型的标准化数据集。

2.模型搭建：生成器与判别器设计，用“标准化数据集”训练 GAN 模型，构建能生成逼真缺陷的 DCGAN 模型架构。

3.训练优化：使 GAN 模型通过对抗训练达成纳什平衡，能生成高相似度缺陷图。

4.生成控制：生成器明确裂纹、气孔、砂眼等缺陷需求类型，基于训练成熟的 GAN 模型批量生成特定类型的缺陷图像，确保每张缺陷形态不同。

5.质量筛选：将生成的图像与真实缺陷图对比，剔除低质量生成图像，保留可用样本。

七是质量评测筑牢数据集高质量“防线”。数据质量评测是推动工业数据集从“建成”到“建好”的关键环节。围绕准确性、完整性、一致性、可靠性、时效性等核心指标，构建多维度数据集质量评测体系，采用自动评估与人工评估协同方式，全面掌握数据集质量状况。通过量化指标度量与定性研判，精准定位极端工况数据缺失、设备异构数据冲突等质量问题，为工业数据集迭代优化提供清晰的技术路径。

八是持续运维保障数据集长效可用。工业数据集的“高质量”需依托长效运维机制实现持续保障，通过定期更新数据集内容，淘汰过时数据，补充新增场景数据，确保数据集与业务需求动态适配。主要运维方式包括：**领域适应性更新**，根据工业应用场景变化及时更新数据集特征维度，确保数据集的场景适配性；**持续优化更新**，依托用户反馈机制与自动化监测工具，实时捕获数据错标等质量偏差，触发重标注、

增补样本等迭代流程，推动数据集在反馈与调整中完成自我进化；**动态维护更新**，实时监控数据源，及时解决设备漂移、新工况数据缺失等问题，确保数据集持续可用。

研发实施环节缺失或执行不规范，将导致工业高质量数据集建设陷入“需求空转、应用失效”困境，从而引发全链条风险。**一是数据集沦为“低价值数据堆积”**，例如，因缺乏标准化采集和质量校验机制，数据集无法满足模型训练需求，将沦为复用率极低的“数据垃圾”。**二是数据集与实际需求适配度不足**，例如，在汽车焊接场景中，因未规范采集电流、电压等动态参数，数据集难以支撑质量检测模型训练，直接影响缺陷识别精度。**三是应用验证阶段陷入低效循环**，因研发实施不规范导致的数据集质量缺陷会直接传导至应用验证环节，导致企业反复返工，模型验证周期显著延长。

（三）应用验证

因工业场景的动态性与复杂性，工业数据集普遍存在标注逻辑偏差、场景覆盖盲区等隐性问题，难以通过研发阶段的质量评测完全暴露，须将其置于工业场景实际应用中，通过场景的实际验证来检验数据集的可用性。

应用验证环节通过构建“数据集优化→模型提升→数据集再优化”迭代闭环机制，实现模型性能与数据集质量的双向提升，进而推动数据集建设全流程的持续改进，是工业高质量数据集建设的“试金石”与“推进器”。**一方面，数据集通过模型反馈优化**。将构建完成的数据集接入工业预测模型、故障诊断系统等实际应用载体，通过模型表现反向验证

数据集的完整性、样本分布合理性、时效性以及场景适配性，精准定位数据集中存在的标注偏差、特征维度不足、时效性滞后等问题，并反馈至建设规划和研发实施环节。另一方面，**模型性能提升反推数据集新需求**。随着数据集的不断优化，模型性能同步提升，其适用场景随之扩展，进一步对数据集的覆盖范围与特征维度提出新要求，实现“模型能力边界扩展→数据集范围扩展”的正向循环。

若跳过应用验证，将导致“看似合格”的数据集向全流程传导，形成“数据质量不足→模型性能失效→业务损失→成本返工”的恶性循环。**一是数据集质量缺陷隐性化引发“带病上线”风险**。例如，生产线故障诊断数据集未经实际工况验证，可能遗漏特定负载下的设备异常特征，导致数据集中“正常样本”与真实场景“故障样本”边界模糊，严重影响模型训练效果。**二是模型性能失控导致业务风险传导**。基于“表面合格”数据集训练的模型，预测精度、诊断准确性等指标存在隐性缺陷，实际预测或诊断中可能出现高误报率或漏报率，影响生产决策。**三是迭代返工引发后期成本失控**。数据集缺陷暴露后需回溯至数据采集、清洗、标注等环节进行重新优化，导致时间、人力等成本指数级增长。

（四）典型案例

1.南京钢铁集团有限公司：中厚板轧制力实时智能预测数据集

（1）建设背景与目标

南京钢铁集团有限公司（以下简称“南钢”）中厚板产

线依赖进口 TMEIC 轧制力机理模型，存在预测精度不足（ $\pm 10\%$ 误差内准确率约 84%）和自主可控缺失的瓶颈，存在三大核心业务痛点：**一是**非稳态工况下，传统机理模型预测失准率达 20%，换辊、换钢种等环节需频繁人工干预，影响产线连续作业效率；**二是**小批量特种钢材生产时，现有模型自学习能力不足，轧制力预测波动大，板形缺陷率上升，制约高端特种钢材国产化；**三是**中厚板生产中，现有模型预测精度不足致 AGC 系统调节偏差，同板差合格率低，影响厚板成材率。

为突破传统机理模型预测失准、现有模型预测精度不足等瓶颈，南钢亟须开发融合冶金知识与人工智能算法的**自主高精度轧制力预测模型**，构建覆盖炼钢、连铸、加热、轧制全工序链的高质量结构化数据集。

（2）数据集建设规划

从样本范围、质量指标、安全合规等方面进行数据集建设规划，明确数据集建设要求，确保数据集的结构化、完整性、准确性、可复用性与时效性等，满足高精度轧制力预测模型的训练与验证需求，如表 3-1 所示。

表 3-1 典型案例：中厚板轧制力实时智能预测数据集建设要求

维度	建设要求
样本范围	覆盖不同钢种（如普钢、特种钢）、不同规格、稳态与非稳态工况（如换辊后、换钢种初期）以及小批量钢种
	涵盖炼钢/连铸、加热、轧制等全工艺流程的关键参数。例如，炼钢/连铸阶段的化学成分和连铸拉速，加热阶段的温度曲线和出炉温度，以及轧制阶段的入口厚度、温度、速度等
质量指标	完整性： 覆盖完整生产流程，关键参数缺失率需极低
	准确性： 关键输入参数（如化学成分、拉速、温度、厚度、辊径）的实际测量值需严格遵循生产实际，与真实值偏差控制在 $\pm 5\%$ 范围内

维度	建设要求
	异常值控制： 全数据集异常值比例须<5%，并经过严格识别与处理
	时序对齐： 需确保不同频率（高频传感器 vs 低频系统）的数据点在时间维度上精确对齐，以反映真实工艺状态
	目标值： 包含精确测量的实际轧制力值作为模型训练与验证的目标变量
安全合规	重要数据保护： 针对核心工艺数据，实施物理隔离存储策略，在独立安全域部署高敏感数据存储节点。访问控制采用系统管理员与安全管理员双管模式。
	隐私合规： 订单号、供应商代码等信息进行脱敏，钢种映射为随机编码隐匿真实牌号，在消除敏感信息泄露风险的同时，维持工艺参数与材料性能的关联逻辑
	操作跟踪： 记录字段级变更（包括操作时间、账号、IP、数据 ID 及值变更轨迹），对化学成分修改等敏感操作自动标记告警，且全链路数据操作日志保留≥3 年

（3）数据集建设实施

南钢通过数据采集、处理及增强等技术手段，构建了涵盖 122 个维度、500 余万条的高质量数据集。在研发规范方面，例如，应实现高频传感器数据与低频系统数据的时序对齐；以板坯 ID 为主键，结构化存储全链路数据。在数据处理方面，采用“冶金规则+统计检测+无监督学习”三级协同清洗体系，从基于冶金物理约束的粗筛，到利用数理统计方法识别显性异常值，再到通过无监督学习识别隐性异常三个层次进行数据清洗。针对多系统数据异构问题，利用物料跟踪技术实现跨工序数据动态关联与时空对齐，解决数据在时间、空间维度的匹配难题。在数据增强方面，深度融合材料科学理论与冶金先验知识构建知识特征库，开展特征增强工作，强化特征判别能力与模型可理解性，有效弥补传统数据驱动模型工况适应性不足的缺陷。

（4）数据集应用验证

南钢结合数据集与模型训练的协同优化机制，将高质量数据集应用于轧制力预测模型的训练与验证，以持续提升模型预测精度。

一是模型训练驱动数据迭代。深入分析模型训练结果（如特征重要性、预测误差分布），识别影响模型性能的关键数据项或潜在数据质量问题。

二是数据项针对性调整优化。根据模型分析反馈，动态调整数据集中的数据项构成。例如，可能增加关键过程的高频传感特征、补充缺失工艺参数，或剔除冗余/低贡献度特征。

三是数据集同步优化。数据项调整后，对数据集进行同步清洗、验证与更新，确保调整后的数据集质量不降低。

四是闭环反馈验证。使用优化后的数据集重新训练与验证轧制力预测模型，评估精度提升效果。

通过“模型训练反馈 - 数据项调整 - 数据集优化 - 模型再训练”的闭环迭代过程，不断精炼数据集的核心特征与信息密度，支撑轧制力预测模型实现更高的准确性与泛化能力。

（5）数据集应用成效

在南钢中厚板生产过程中，轧制力的精准预测是确保产品质量与生产效率的关键环节。通过将高质量数据集与人工智能模型深度融合，成功构建了轧制力实时智能预测系统，并取得显著应用成效。

一是质量稳定性显著提升。轧制力预测相对误差控制在10%范围内的准确率 $\geq 95\%$ ，优于传统模型的预测精度，有效减少因轧制力波动导致的厚度超差、板形失稳等质量问题；

非计划率降低 3%，提升产品合格率。

二是生产效率与产能提升。板材成材率提升 0.2%，降低材料消耗，提高轧制过程稳定性。充分发挥轧机最大能力，在设备安全范围内逼近理论最大压下量，单道次形变量增加，降低轧制总道次数，最终提升轧机规程产能。

三是数据驱动工艺优化。系统积累的实时生产数据为工艺参数优化提供数据支撑，形成“预测－反馈－优化”闭环，持续提升模型泛化能力。未来可拓展至多钢种、多规格板材的智能轧制策略优化，进一步挖掘降本增效潜力。

2.蔚来控股有限公司：新能源汽车电驱研发大模型数据集

（1）建设背景与目标

蔚来控股有限公司（以下简称“蔚来”）在新能源汽车电驱系统研发测试环节，发现传统模式存在显著短板。一是测试项评估过度依赖个人经验，导致决策缺乏数据支撑，易造成资源浪费；二是测试规格说明编写工作量大、重复性高，历史案例复用难度大，且因人员对需求理解差异，出现用例描述模糊、执行标准不统一等问题；三是电驱系统测试环节具有强跨学科特性，涉及电磁兼容性（EMC）测试、NVH（噪声、振动与声振粗糙度）优化、多合一集成系统开发等多个领域；四是数据来源分散，涵盖高校研究成果、主机厂实测报告与第三方检测机构数据等，结构化整合难度极高。

为解决上述短板，蔚来围绕电驱测试场景，构建新能源汽车电驱研发大模型高质量数据集，支撑电驱测试的测试

SPEC 和脚本自动生成。一方面，提升测试 SPEC 编写效率，目标效率提升 10 倍以上。另一方面，提高测试覆盖度，目标覆盖度达到 95%，同时解决传统测试模式中经验依赖、资源浪费、标准不统一等问题，为电驱系统研发提供数据驱动的支撑体系。

（2）数据集建设实施

蔚来在数据集建设过程中，攻克了电驱系统测试数据跨学科整合的复杂性、数据隐私安全风险等问题，建立覆盖数据“采集 - 处理 - 分析 - 建模 - 验证”的闭环系统，形成性能优化、可靠性提升的完整数据闭环，已成功构建了 5TB 规模的文本类新能源汽车电驱研发大模型数据集。在多源数据采集环节，从数据库、API、传感器、企业文档、专家知识库等多渠道获取原始数据，基于 datafabric 思想，实现数据连接。在数据清洗环节，对于多源数据开发转换器和处理器，处理缺失值、噪声数据及重复项，保证数据质量与一致性。在数据标注环节，利用大模型能力对非结构化数据进行标注，与结构化数据共同构建领域数据集。在数据划分环节，将领域数据集划分为训练集、验证集和测试集，用于模型开发与评估。在模型训练与评估环节，利用数据集对模型进行在线和监督学习，提升模型的效果，并利用测试集、专家知识和大模型验证性能。

（3）数据集应用成效

围绕电驱测试场景需求，新能源汽车电驱研发大模型数据集在基础研发效率、AI 模型性能、研发模式变革及行业价

值四个方面取得显著应用成效。

一是基础训练支撑。在测试用例生成领域，采用经严格清洗和标注的数据集进行微调训练，结合专家规则的验证流程，可使系统生成测试用例的准确率提升 10% 以上。这种专家知识校验的训练方式，确保了模型在特定领域具备专业级的推理能力。

二是调优与泛化能力提升。通过构建包含用户显式反馈和隐式修正注意力特点，能够采集多维度反馈的数据集，实现模型持续优化。实践表明，这种结合用户显式反馈和隐式反馈的调优方式可使模型在真实场景的适应能力提升 15%。

三是 RAG 优化实践。基于知识图谱的检索增强生成 (RAG) 系统，通过构建领域知识图谱和关联性校验机制，实现知识获取的高准确率。该系统从多源异构数据中提取实体和关系，构建结构化知识图谱；在检索阶段，利用语义相似度计算和路径推理，确保检索内容的相关性；再通过一致性校验模块，过滤矛盾或过时信息。这种双重校验机制使知识获取的正确率达到 98.5% 以上。

3. 西安塔力科技有限公司：矿山行业全场景要素数据集

(1) 建设背景与目标

当前矿山行业智慧化建设面临“烟囱效应”突出、数据孤岛严重、系统协同困难、传统运维效率低下等痛点，导致数据价值难以释放，安全管理与生产运营成本高企。西安塔力科技有限公司（以下简称“塔力科技”）以矿山数字化转型为核心目标，构建“人-车-设备-环境”的全场景高质

量数据集，覆盖矿山全场景关键要素，如表 3-2 所示，破解矿山数据利用低效、安全预警滞后、人工依赖度高的难题。

表 3-2 典型案例：矿山行业高质量数据集覆盖要素与场景

要素	数据集覆盖场景
人员	人员装备状态、违规行为（如危险区域闯入、未戴防护装备）
车辆	车辆分类、停放规范、行车轨迹、人车交互风险
设备	设备运行状态（如液压支架伸缩、采煤机截割）、防护完整性、标识有效性
环境	异物占道、物料堆放隐患、极端工况（烟雾、低光照、粉尘）

（2）数据集建设实施

在矿山图像数据集建设实践过程中，塔力科技通过数据采集、精准标注、数据增强等技术手段，构建了涵盖 350 多座煤矿数据、57 个煤矿场景及 1.1 亿张要素图片的高质量矿山数据集。在数据采集方面，涵盖每矿约 11.2 万张现场图像，包含烟雾、黑暗等 11 类极端工况。在数据标注方面，塔力科技引入全场景语义分割支持复杂交互识别，精准标注了 53 子类目标标签，保证标注一致性波动 $\leq 5\%$ ，分割掩码误差 $< 2\%$ 。在数据增强与小样本策略方面，实施亮度扰动等常规增强，采用过采样与合成样本补偿稀缺样本（如瓦斯突出前兆、烟雾、违规行为等），结合迁移学习与 MixUp/CutMix 提升小样本泛化能力。

（3）数据集应用成效

围绕矿山核心生产与安全场景，选取采煤工作面、井下运输巷道、主副井口等关键区域开展数据集应用验证，实现设备状态、人员行为、设备异常等情况的实时监测与智能预警，具体场景应用效果如表 3-3 所示。

表 3-3 典型案例：矿山行业高质量数据集典型场景应用成效

场景	维度	应用成效
采煤工作面智能监测	设备状态识别	通过全场景分割对液压支架、采煤机等设备进行像素级标注，实时监测液压支架伸缩状态（识别精度 95.2%）、采煤机截割部运行轨迹，自动识别防护栏缺失（响应时间<1.5 秒），较传统人工巡检效率提升 400%
	人员行为管控	针对采煤面人员违规进入危险区域、未按规定佩戴防护装备等行为，结合骨架提取技术，违规识别精度达 96.7%，误报率降至 3.2%
井下运输巷道安全防控	人车交互管理	采用“车—人”分割算法对矿车、胶轮车与行人进行动态追踪，在巷道交叉点、弯道等盲区实现提前 3 秒预警。系统运行以来，井下运输事故发生率下降 62%，其中，人车碰撞隐患识别准确率达 95.3%
	皮带机异常监测	通过分割模型识别皮带机跑偏（误差<2mm）、煤流堵塞、异物混入等状态，结合红外热成像数据，同步监测滚筒过热现象。某巷道皮带机曾因异物卡堵触发系统预警，避免了设备烧毁事故
主副井口多维度监管	人员出入管控	对井口区域进行语义分割，精准识别未授权人员闯入、作业人员未系安全绳等违规行为，联动门禁系统实现自动拦截，月度违规事件较改造前减少 78%
	极端环境适应	针对井口雾气、粉尘等干扰，采用 CLAHE 对比度增强与合成样本训练，在能见度<5 米的环境下，人员识别准确率仍保持 89.6%，设备状态监测不受影响
其他重点场景应用	人员行为识别	典型场景包括睡岗、玩手机、攀爬不当等，基于全场景分割，识别精度提升至 98%以上
	车辆运行与安全监测	在矿车运行、井下巷道人车交互等场景，“车—人”分割与交互检测结合，实现车辆运行轨迹自动化追踪与危险区域预警
	设备运行状态识别	针对液压支架伸缩状态、防护缺失、传感器异常、带式运输机头/机尾状态等场景，可实时识别皮带开/停机、有煤/无煤、遮挡、防护缺失等状态，减少隐患漏检
	环境与风险预警	在烟雾、火情、煤流异常、异物阻塞等场景，结合时空数据，环境异常识别率提升至 95%以上，提前预警火灾和运输风险

四、工业高质量数据集的评估体系

工业高质量数据集的评估体系是衡量和保障数据集质量、价值与安全水平的核心，对推动数据集从量的积累到质的飞跃具有决定性作用，是确保工业数据要素高效流通、充

分释放数据价值、赋能工业高质量发展的关键基础。通过全面科学的评估体系与测评方法，衡量数据集对建模、预测、诊断、优化等任务的支撑能力，为数据集的持续优化、迭代升级与应用推广提供精准指导和可靠依据。通过不断识别问题、优化流程、提升效率，实现数据集质量的持续改善，为工业智能应用提供可靠数据底座。

（一）评估框架

工业高质量数据集评估应围绕数据质量建立定性与定量相结合的评估体系。一是要重点评估数据集的准确性、完整性、一致性、时效性与标注质量等基础质量特性；二是要考察数据集的元数据完整性、场景适配性、可复用性与可扩展性等应用适配能力；三是要审核数据集的内容合规与信息脱敏等安全合规水平。

数据集的评估过程应结合典型行业与业务场景，选取差异化的评估指标，衡量其对建模、预测、诊断、优化等任务的支撑能力，采用公式计算、模型基准测试、专家评审与场景验证、第三方机构认证等方法开展评估工作，为工业高质量数据集的持续优化与应用推广提供依据。

当前面向通用场景的数据集质量评估已在形成诸多标准^[5,8-10]，但现有数据集评估标准对工业领域人工智能应用中尤为关注的分布、标注、时效性与泛化性等要素覆盖不足，对行业特性的差异化评估标准亦有待细化。现有数据集评估标准多着重数据样本自身的质量，对工业数据集构建与应用中面临的标注基准一致性、元数据信息完整性、数

据精度与分辨率、有效信息密度、合成数据真实性与时间跨度等影响实际应用成效的问题关注不多。

因此，针对工业高质量数据集完整度高、质量达标、时效性强、场景明确的显著特征，应参考现有数据评估标准，设计多层次的工业高质量数据集评估框架并设计具体评估指标，覆盖产品缺陷检测、工艺参数优化、设备预测性维护等常见工业应用场景与机器学习、深度学习、多模态大模型等典型模型构建过程中涉及的各类数据质量要素。

工业高质量数据集评估框架以《信息技术 数据质量评价指标》国家标准^[8]为基础，参考相关行业标准与应用需求，提出规范性、完整性、准确性、一致性、时效性、稠密性、均衡性、专业性、通用性、合规性等 10 个一级指标。每个一级指标下还可细分为若干二级指标，以对元数据、数据分布、数据标注等方面进行具体评估，如图 4-1 所示。上述的一级和二级评估指标，是实施工业数据集质量评价的最大集合，实践中应根据具体应用与数据特点，选取适用于所评估对象的评估指标开展数据集质量评价活动。

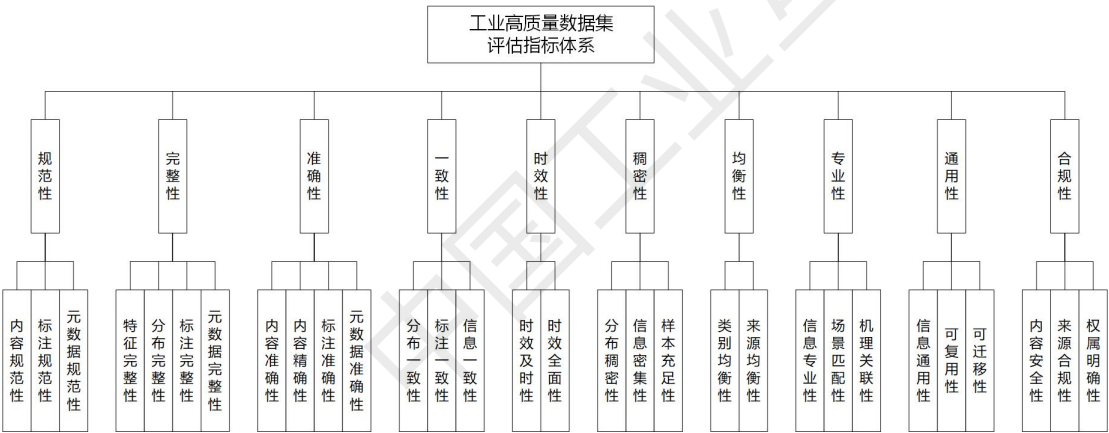


图 4-1 工业高质量数据集评估框架

(二) 评估指标

工业数据集的质量评估既遵循数据质量的一般评估原则，也需考虑工业领域独有特点与服务模型应用的具体需求。工业高质量数据集的评估指标包含 10 个一级指标与 30 个二级指标。各项一级指标内涵与范围概述如下：

（1）**规范性**指数数据集符合数据标准、数据模型、业务规则、元数据或权威参考数据的程度，包括数据集的格式、标注、单位和元数据的规范性等方面。

（2）**完整性**指数数据集按照建设规则要求，数据元素被赋予数值的程度，包括数据集的特征完整性、分布完整性、标注完整性、文档完整性等方面。

（3）**准确性**指数数据集准确表示其所描述的真实实体（实际对象）真实值的程度，包括数据的内容、精度、误差、标注信息、文档描述等方面。

（4）**一致性**指数数据集内部及与其他数据源之间的统一程度，包括数据单位与格式的一致性，数据集样本分布与真实世界数据分布的一致性，不同样本间数据标注基准的一致性等方面。

（5）**时效性**指数数据集能够反映当前或指定时间窗口内的真实状态程度，包括是否涵盖了所需的时效区间、是否包含过时信息等方面。

（6）**稠密性**指数数据集单位数据所含信息量的多样化程度，包括数据样本分布密度、信息价值密度与样本数据充足性等方面。

（7）**均衡性**指数数据集数据的均匀分布程度，包括数据

类别分布均衡性和数据来源均衡性等方面。

(8) **专业性**指数据集涵盖的信息内容和工业场景的匹配程度，包括知识的专业性、业务场景的匹配性、工业机理的关联性等方面。

(9) **通用性**指数据集在跨行业、跨场景的适用程度，具体包括知识的适用性、数据集的可复用性与可迁移性等方面。

(10) **合规性**指数据集的内容、来源与权属遵循各类规章制度的合规程度，包括相关法律法规、监管要求、行业标准与伦理规范等方面。

各项二级指标说明如表 4-1 所示。

表 4-1 工业高质量数据集评估指标

一级指标	二级指标	指标说明
规范性	内容规范性	数据集中样本的数据格式、单位应与标准格式保持一致，无空白、缺失与无效数据。
	标注规范性	数据集中需标注的样本，标注的格式应与标准保持一致。 (适用于包含标注样本的情形)
	元数据规范性	数据集元数据应具备规范的标识与格式。数据集元数据中包含的各类元素内容应以规范形式填写。
完整性	特征完整性	数据集中样本包含的模态与特征，应足够描述对象行为特性，或满足预期要求。
	分布完整性	数据集中样本应覆盖对象在所关注范围内所有可能的状态分布范围。
	标注完整性	数据集中需标注的样本，应完整标注出所有需要标注的标签、类别、对象等信息。 (适用于包含标注样本的情形)
	元数据完整性	数据集元数据中应完整包含数据集的基本信息、内容特征、建设过程与应用说明等各类元素内容。
准确性	内容准确性	数据集中样本的内容应与其所描述对象的实际状态、信息、属性等保持一致。对于合成数据，数据的取值、形态、信息等特征应与其所描述的真实对象保持一致。对于语料数据，语料文本应符合

一级指标	二级指标	指标说明
		合正常语言逻辑，包含的信息和知识应正确无误。
	内容精确性	数据集中样本的数值有效数字、图像与视频分辨率，误差范围、噪声占比等精度属性应足够精准刻画对象特征，或满足预期要求。 (适用于对精度指标存在要求的情形)
	标注准确性	数据集中需标注的样本，标注的类型、位置与尺度等应准确无误。 (适用于包含标注样本的情形)
	元数据准确性	数据集元数据中包含的各类元素内容应准确无误。
一致性	分布一致性	数据集中样本所数据样本在空间、时间、属性等维度上的分布形态应与真实对象保持一致。
	标注一致性	数据集中需标注的样本，不同样本间应遵循一致的标注基准与判别标准。 (适用于包含标注样本的情形)
	信息一致性	数据集中样本包含的数值、逻辑和知识等信息应能够自洽，不应存在自相矛盾。
时效性	时效及时性	数据集中包含的内容应符合对象当前的实际状态，不应存在过时数据和信息。 (适用于对时效及时性存在要求的情形)
	时效全面性	数据集中包含的内容应全面覆盖对象在全部时间范围内的状态信息，或覆盖指定的时段区间，不应存在遗漏或缺失。 (适用于对时效全面性存在要求的情形)
稠密性	分布稠密性	数据集中样本在空间与时间维度的样本密度应满足香农采样定理要求，能够无失真地刻画原始对象状态信息。
	信息密集性	数据集中样本应包含有效信息，具备高信息密度，无重复、冗余、噪声、无意义样本。
	样本充足性	据集中包含的样本数量应足够描述对象行为特性，满足模型构建对样本数量的需求。
均衡性	类别均衡性	数据集中样本所属类别应具备均衡的类间分布，涵盖稀缺场景与类型。 (适用于对类别均衡性存在要求的情形)
	来源均衡性	数据集中样本应具备多样化的来源，涵盖有代表性的数据来源。 (适用于对类别均衡性存在要求的情形)
专业性	信息专业性	数据集应包含面向特定工业行业领域的有效信息和知识，可用于解决专业领域问题。 (适用于对信息专业性存在要求的情形)
	场景匹配性	数据集的形式应与相关的业务场景相契合，内容应能够满足相关业务场景实际使用需求。
	机理关联性	数据集中数据的生成、分布、内容等应符合对象

一级指标	二级指标	指标说明
		的自身机理逻辑，具备基于机理的可解释性。 (适用于对机理关联性存在要求的情形)
通用性	信息通用性	数据集应具备跨场景、跨用户、跨行业不等的通识类信息和知识。 (适用于对信息通用性存在要求的情形)
	可复用性	数据集应具备对使用方能够持续复用的能力。
	可迁移性	数据集应具备对相似场景可迁移泛化的能力。
合规性	内容安全性	数据集中敏感信息应完成脱敏处理，不应包含恐怖、偏见、歧视、误导、虚假等有害信息，不得包含中毒与恶意数据，遵循《中华人民共和国数据安全法》与《中华人民共和国个人信息保护法》等相关法规规定 ^[11-13] 。
	来源合规性	数据集中数据的来源与采集过程应合法或得到授权。
	权属明确性	数据集中开源数据应符合开源协议，非公有数据应具备权属证明与使用许可。

(三) 测评方法

工业高质量数据集的评估通常包含评估准备、评估活动和评估结论三个阶段，如图 4-2 所示。**评估准备阶段**包括评估对象分析、评估指标选取、评估指标权重、评估准则制定设定四个步骤。评估准备需根据具体场景和数据特点，对评估指标进行取舍，选取适合的数据集质量指标并按需设定不同的指标权重，形成制定数据集评估的计划。**评估活动阶段**包括评估活动执行、分项指标结果收集、总体质量指标计算三个步骤。评估活动需对选定的各项指标分别按标准化的方法进行评定并收集评估结果；指标计算应基于具体指标特点与业务场景，选取定性或定量方式进行，以合格/不合格、等级或分数形式得出评估结果。**评估结论阶段**包括数据集质量判别、评估报告编制两个步骤。评估结论需根据各项评估指标的等级或得分，结合指标权重综合得出最终的数据集质量判别结果，并编制评估报告。

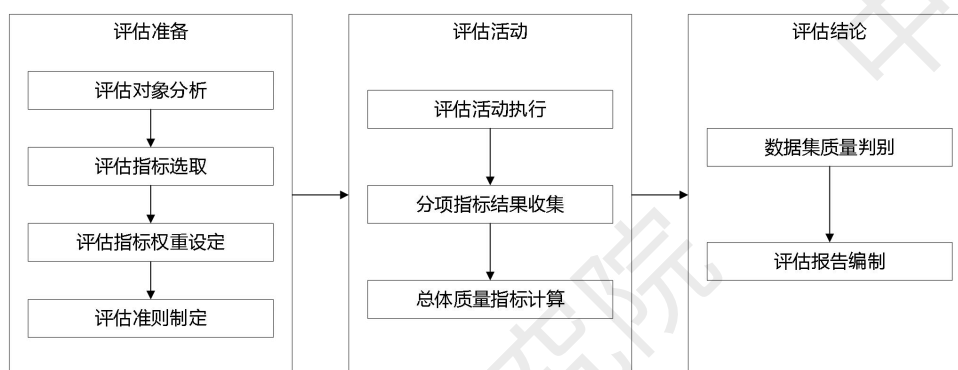


图 4-2 工业高质量数据集评估流程

不同行业不同场景下，对评估指标的选取各有侧重，应结合实际数据特点与模型类型进行分析。例如，**对于基于视觉图像的产品缺陷检测**，需要特别重视标注准确性与标注一致性指标，以保证直接影响模型的训练效果；**对于需要精准建模的过程控制系统辨识**，数据记录的精确性应给予更高的关注。**对于基于超声信号的结构探伤**，需要重点优化采样频率与信噪比指标以提升异常信号的识别效果；**对于设备预测性维护**，因剩余寿命预测模型的构建需要设备全生命周期的工况与失效记录，故而需对时效完整性和分布全面性指标加强关注。

工业高质量数据集的建设不是一劳永逸的，而是贯穿数据集从构建到应用整个生命周期中持续改进的过程。应从质量管理的角度，将数据集视为一种产品，引入系统化的质量管理体系机制，结合“策划—实施—检查—处置”的“PDCA”循环^[14,15]，全方位系统性提升数据质量，确保数据集质量的可控性与可追溯性，推动工业高质量数据集建设可持续发展。

五、工业高质量数据集的流通应用

工业高质量数据集的流通应用是激活工业数据价值、赋能产业提质增效的重要环节。当前，数据孤岛严重、权属不

明晰、流通安全缺乏保障等实际问题阻碍了工业高质量数据集的流通共享，亟需构建可信数据空间等新型数据基础设施，通过区块链、隐私计算等技术，实现工业高质量数据集跨主体协同、供需协作和合规高效利用。

（一）面临的问题

当前，工业高质量数据集流通应用的主要问题集中在供需协作不畅、隐私保护不足、流通管控体系不完善等方面，直接影响数据集价值转化效率。

一是数据集隐私保护与共享需求存在矛盾，共享意愿不足。工业领域蕴含研发设计、生产工艺参数等大量敏感数据，企业为保护商业机密、防范数据泄露、确保合规运营，对数据集共享持高度谨慎态度。然而，工业高质量数据集的开放共享是实现数据价值释放的关键路径，需要各企业共享数据集，形成合力。当前隐私保护技术在实用性上难以兼顾，依然存在数据泄露风险，这种安全不确定性显著削弱了企业共享积极性，“安全顾虑压倒共享动力”问题突显，严重制约数据集的价值释放。

二是数据集权属不明与全链路追溯能力不足，主权缺乏保障。工业高质量数据集流通涉及数据提供方、使用方等众多主体，各主体权属边界模糊、缺乏统一的权属登记与确认机制，易引发数据集使用争议和利益分配纠纷。此外，数据集在传输、使用等环节缺乏唯一标识与完整日志记录，流向追踪与行为审计较薄弱，一旦发生数据集滥用或泄露，难以快速追踪溯源，进一步加剧数据提供方对自身数据集归属权

的担忧，制约数据集的可信共享与流通。

三是数据集流通可信管控与信任机制缺失，阻碍供需协作。工业高质量数据集具有高价值密度属性，且在流通利用环节涉及多主体、多角色参与，不同参与方的访问范围、操作权限存在显著差异，对数据集全流程使用过程的精准管控提出更高要求。由于跨主体可信管控能力不足，对数据集的使用行为和操作目的验证难度较大，数据集质量难以保障，导致企业对数据集共享缺乏足够信任，形成“管控不足→信任缺失”的负面循环。在供需协作方面，信任机制缺失导致供需对接成本高、周期长，进一步推高了数据集的协作成本。

四是数据集建设分散无序，供需匹配低效。从供给端看，数据集建设分散于各企业、各生产环节，呈现“碎片化”分布特征，缺乏跨主体、跨环节的统筹规划，难以形成规范化的高质量数据集。使用方由于缺乏数据集检索与筛选渠道，难以快速定位符合自身业务需求的数据集，增加了需求挖掘成本。而供给方无法通过公共或开放平台触达潜在用户，导致高质量、高价值数据集的市场曝光度不足，难以形成有效供给和精准匹配，整体呈现“信息孤岛与无序供给并存”的局面，降低了数据集的供需匹配效率。

（二）解决路径

依托国家工业互联网大数据中心，构建工业高质量数据集可信数据空间，为供需多方数据主体提供安全、合规、可信的应用环境，确保数据集可信可管可追溯，促进工业高质量数据集的精准管理、安全共享与高效利用，总体功能架构

如图 5-1 所示。



图 5-1 工业高质量数据集可信数据空间总体功能架构

一是技术赋能，夯实数据集流通信任底座。基于隐私计算、安全沙箱、可信认证、数据使用控制及全链路追溯等核心技术，构建“数据集可用不可见、用途可控可追溯”的信任机制，打通数据集供给方、需求方及服务方协同链路，在“数据集物理不出域”前提下，通过“可用不可见”模式满足多主体的数据集使用需求，实现数据集主权不转移、价值可流转。通过身份可信认证与权限管控筑牢“准入门槛”和“权限基石”，利用多主体身份认证、细粒度权限策略、动态权限调整等能力，保障数据集访问“主体可信、按需分配、动态可控”。利用全链路台账与追溯体系打造数据集全生命周期“数字账本”，构建数据集“共建一共享一共用”全流程台账体系，通过记录数据集提供方和接收方、共享时间、使用场景、操作行为、数据集流向等全维度信息，以及全链路关联追溯，实现“数据来源可查、使用去向可追、责任主体可定、共建贡献可溯”。基于全流程安全防护技术构建“安全屏障”，利用数据脱敏与加密、隐私计算等技术能力，在保障数据集隐私与安全前提下，实现数据集价值有效释放。

二是架构支撑，形成规模化层级化管理体系。通过“国家—产业集群—企业”三级互联部署架构，形成规模化覆盖、层级化运营的可信数据空间管理体系，部署架构如图 5-2 所示。**企业层面**，通过链接生产制造、设备运维等数据源头，利用数据采集、清洗及标注等技术手段，构建高质量“数据集原材料”；**产业集群层面**，聚焦先进制造业产业集群，通过数据可信流通与协作，整合优化企业级数据集，形成具备产业特色或行业共性的数据集；**国家层面**，构建全国工业高质量数据集目录体系，实现工业数据集资源的有序整合与优化配置。

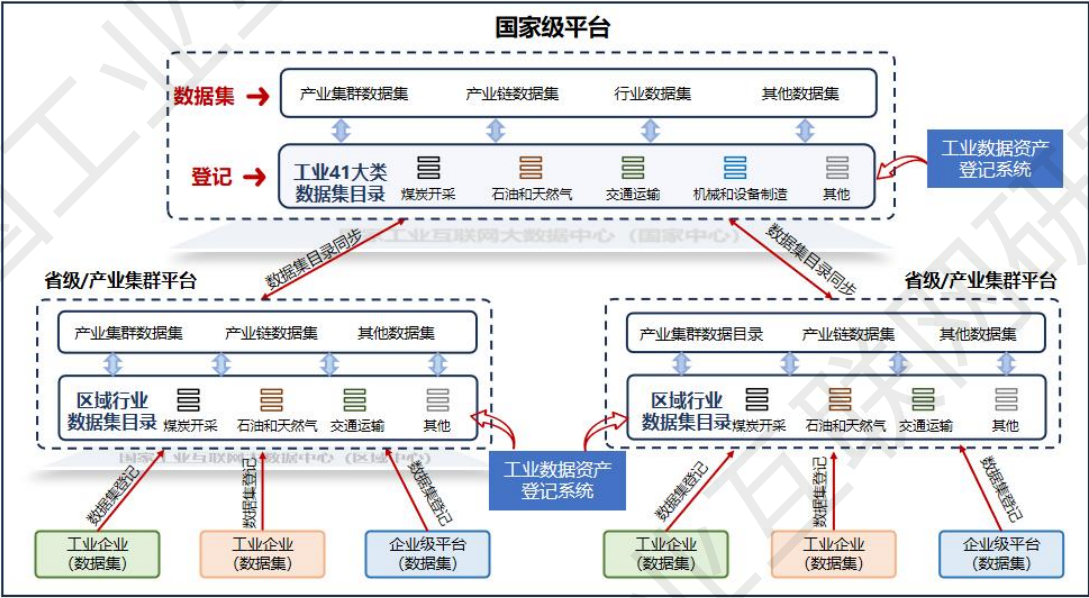


图 5-2 工业高质量数据集可信数据空间部署架构

三是价值转化，精准匹配全层级数据集需求。依托工业高质量数据集可信数据空间，通过构建工业高质量数据集“联合共建、可信共享、高效共用”的全链路闭环体系，支撑设备、产线、生态等全层级工业高质量数据集产品的注册、审核、上架及安全可信共享，有效破解数据集的流通壁垒与

信任障碍。通过数据集需求侧画像与供给侧标签的精准匹配，满足不同层级工业模型对数据集的差异化需求，推动数据集价值从单点应用向全局赋能延伸。

（三）实施方案

以工业高质量数据集的流通应用为导向，依托工业高质量数据集可信数据空间，加快推动数据集的确权登记、目录体系构建和工业数据应用。

1.开展数据确权登记

厘清企业数据底数、明确权属关系，是推动工业数据流通与应用的首要环节。通过建立“国家—产业集群—企业”三级工业数据登记体系，帮助企业梳理数据并完成登记，积累工业数据目录，提升工业数据供给能力。一是**企业级**，帮助企业梳理数据资产。企业通过部署轻量化登记工具，能够在数据不出企业数据库的前提下，生成标准化数据目录，并获得登记证书；二是**集群级**，绘制产业数据资源地图。产业集群、园区或地方政府聚合其辖区内企业数据目录，形成区域性的数据资源地图；三是**国家级**，汇聚企业和集群数据目录，构建覆盖全国的工业数据资源地图。以纺织行业为例，企业在本地部署可信登记软件，自主选择织机运行参数、原材料消耗、库存等数据进行登记，相关信息上传至确权平台。平台审核通过后，为企业颁发登记证书。不同区域的纺织企业数据登记不断积累，将最终形成全国纺织行业数据目录。截至 2025 年 8 月，中国工业互联网研究院工业数据资产登记平台已在苏州、北京、沈阳等地试点，累计注册企业 2200

余家，发放登记证书超 1700 张，为高质量数据集建设提供有效支撑。

2.构建数据集目录体系

以工业高质量数据集可信数据空间为“运行载体”，构建标准化的工业高质量数据集目录体系，并汇聚至国家工业互联网大数据中心，支撑高效检索与供需匹配，为流通应用奠定基础。一方面，推进形成“分业-分级-分类-分任务”多维目录框架，实现工业高质量数据集精准归类。分业维度，依据《国民经济行业分类》（GB/T 4754-2017）^[16]，针对 41 个大类制定差异化规则，归类行业特定数据集。例如，制造业专用设备类重点包含机床加工精度、设备故障诊断等数据集。分级维度以 5 级架构（L1~L5）为核心，构建数据来源、类型、特征、场景的层级映射。例如，用户查询“L3 工厂级安全环保数据集”时，系统可通过映射定位到质量管理体系或健康安全环境管理系统的相关数据集，并关联到安全风险监控预警场景。分类维度按产品全生命周期细化数据类型。例如，研发设计类涵盖产品 BOM 数据、工艺设计数据等；生产制造类涵盖原料配比数据、工艺运行数据等。分任务维度针对训练、微调、评测、仿真任务标注标签。例如，训练数据集强调多样化普适逻辑，仿真数据集突出合成场景定制，确保目录支持模型全链路需求。另一方面，支撑工业高质量数据集检索，促进供需匹配。依托目录体系，融合语义识别、标签匹配（如结合分业“纺织业”与分级“L3 工厂级”）等技术，搭建智能检索引擎，提升需求方检索效率。构建多维

度需求画像（如基于分任务“微调”与分类“研发设计”的精准推荐），实现从“被动检索”到“主动匹配”的模式转变。例如，在可信数据空间中，用户查询“汽车制造业 L4 企业级供应链优化数据集”，系统自动匹配经营管理类采购销售数据，支持联邦学习下的可信共享。

3.推进工业数据应用

基于工业高质量数据集的应用探索正加速演进，逐渐形成三类商业化服务模式：**一是工业场景训练语料服务。**数据服务商通过平台化订阅或私有化定制，向模型厂商提供精加工的工业场景数据集，提升其模型在细分领域的专业精度。如，某数据服务商为头部模型厂商提供的定制化工业语料服务，单个项目金额可达近千万元，体现了高质量工业语料作为“AI 燃料”的巨大商业价值。**二是 AI 一体机定制化服务。**数据服务商结合数据集、模型、算力以一体机部署方式为工业应用企业提供安全、高性价比的智能服务。如，某龙头企业打造的工业 AI 一体机，极大降低了 AI 技术在工业现场的应用门槛，实现了数千万元的产品服务收入。**三是数据处理工具服务。**数据服务商以云化 API 或定制化服务等方式为人工智能应用企业提供数据“采、洗、标、测、用”等一站式数据处理服务。如某企业自研的数据标注平台，最高可提升 6 倍标注效率，平均降低数据处理 40% 成本。随着高质量数据集与人工智能在各垂直细分行业的深入应用，将催生出更多高效、自适应、智能的工业数据应用新模式，加速产业智能化转型升级。

参考文献

- [1] 全国工业过程测量控制和自动化标准化技术委员会. 智能制造 工业数据 分类原则: GB/T 42128-2022[S]. 2022.
- [2] ISO TECHNICAL COMMITTEE ON AUTOMATION SYSTEMS AND INTEGRATION - INDUSTRIAL DATA. ISO/TC 184/SC 4 Industrial Data[EB/OL]. <https://committee.iso.org/home/tc184sc4>.
- [3] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. Data quality - Part 2: Vocabulary: ISO 8000-2:2022[S]. 2022.
- [4] 全国数据标准化技术委员会. 高质量数据集 建设指南: TC 609-5-2025-01[S]. 2025.
- [5] 中国电子工业标准化技术协会. 人工智能 数据集质量评估要求: T/CESA 1381-2025[S]. 2025.
- [6] 重庆市大数据应用发展管理局. 工业数据分类分级导则: DB50/T 1453-2023[S]. 2023.
- [7] 辽宁省工业和信息化厅. 工业数据分类分级管理指南: DB21/T 3867-2023[S]. 2023.
- [8] 全国信息技术标准化技术委员会. 信息技术 数据质量评价指标: GB/T 36344-2018[S]. 2018.
- [9] 全国自动化系统与集成标准化技术委员会. 工业数据质量通用技术规范: GB/T 39400-2020[S]. 2020.
- [10] 中国通信标准化协会. 面向机器学习的电信数据规范 数据质量: YD/T 4522-2023[S]. 2023.
- [11] 中华人民共和国数据安全法[A]. 2021.
- [12] 中华人民共和国个人信息保护法[A]. 2021.
- [13] 国务院. 网络数据安全管理办法[A]. 2024.

- [14] 全国质量管理和质量保证标准化技术委员会. 质量管理体系 要求: GB/T 19001-2016[S]. 2016.
- [15] 全国自动化系统与集成标准化技术委员会. 数据质量 第 63 部分: 数据质量管理: 过程测量: GB/T 42381.63-2023[S]. 2023.
- [16] 中国标准化研究院. 国民经济行业分类: GB/T 4754-2017 [S]. 2017.