高质量数据集建设与运营路径研究

2025 年 12 月 30 日

前 言

在数字经济迅猛发展的时代背景下，数据已成为驱动社会进步和产业变革的核心生产要素。随着人工智能、大数据、云计算等新一代信息技术的广泛应用，高质量数据集作为支撑算法训练、模型优化与智能决策的关键基础，其重要性日益凸显。无论是科学研究、智能制造，还是智慧城市、医疗健康等领域，高质量数据集的质量直接决定了技术应用的精度与效能。然而，当前数据资源的“量大质低”问题依然突出，数据冗余、标注不规范、来源不可靠、更新滞后等问题制约了数据价值的充分释放。在此背景下，系统性地研究高质量数据集的建设与运营路径，具有重要的现实意义和战略价值。

本白皮书立足于国家政策导向与产业发展实践，旨在为高质量数据集的规划、建设与运营提供一套完整的方法论框架与实践指南。白皮书首先厘清了高质量数据集的概念内涵与多维分类体系，系统梳理了全球与我国高质量数据集的发展现状、典型模式与面临的共性挑战。核心部分聚焦于建设与运营实践，详细阐述了从建设模式选择，到覆盖“需求-规划-采集-治理-标注-验证”的全生命周期核心环节，再到构建“资源管理-价值转化-生态共建”三位一体的运营体系，并结合浙江电信的实践案例进行说明。最后，结合前沿趋势，

提出了涵盖系统能力建设、长效运营机制及基础制度保障的发展建议。

我们希望本白皮书能够为政府部门、行业企业等各类参与主体提供有价值的参考，共同推动我国高质量数据集建设迈向体系化、规范化、生态化的新阶段，夯实人工智能发展的数据根基，为发展新质生产力、建设数字中国注入强劲动力。

目 录

一. 高质量数据集概述

（一） 高质量数据集政策背景

1. 国家持续完善高质量数据集顶层设计

近年来，国家高度重视人工智能与数据要素发展，将高质量数据集建设视为夯实人工智能发展数据基础的核心抓手，密集出台系列政策，系统构筑了从战略规划到落地指引的顶层政策法规体系。

从长期布局看，国务院办公厅早在 2021 年印发《要素市场化配置综合改革试点总体方案》，提出建立公共数据共享协调机制，优先推进企业登记监管、卫生健康、交通运输、气象等高价值数据集向社会开放，为数据要素流通与数据集建设筑牢制度根基；2023 年 12 月，国家数据局等部门发布《“数据要素×”三年行动计划（2024-2026 年）》，聚焦科学数据开放共享，提出构建科学知识资源底座，建设高质量语料库与基础科学数据集，支持开展人工智能大模型开发和训练；2024 年 1 月，工信部等九部门出台《原材料工业数字化转型工作方案（2024-2026 年）》，明确建设适用于生成式 AI 的行业数据集，基于现有通用大模型技术底座进行定制化开发训练，构建细分行业大模型。2025 年 1 月，国家

发改委、国家数据局、工信部联合发布《国家数据基础设施建设指引》，明确提出要建设“数据高效供给体系”，支持在农业、工业、交通、金融、科技等行业领域打造高质量数据集，从国家数据基础设施层面，为各行业数据集建设提供方向指引。2025 年 5 月，《数字中国建设 2025 年行动方案》再次强调要加强交通、医疗、制造等重点领域数据标注，建设行业高质量数据集。

在国家政策协同驱动下，多部门合力推进高质量数据集建设。2025 年 2 月，国家数据局在北京召开高质量数据集建设工作启动会，国家发展改革委、教育部、科技部、工业和信息化部等 27 个部门参会。同月，国资委“AI+”行动讲话精神要求，分批构建重点行业数据集，建设好通用基础数据集，鼓励龙头企业与链主企业牵头建设，做强做优数据产业。

2. 地方多措并举推动高质量数据集建设落地

在中央政策引领下，各地积极推进高质量数据集建设。多地结合地方产业特色，围绕技术创新、生态培育、制度保障等关键环节出台了一系列政策举措，着力推动高质量数据集建设从政策规划加速应用落地。

多地以政策文件形式明确数据集建设的量化目标与重点领域，强化“数据+产业”联动。上海提出，2025 年建成

1000 个高质量数据集，覆盖金融、医疗、航运等领域；广东计划 2027 年落地 50 个以上行业数据集，聚焦制造、教育等典型行业；江苏要求面向制造、教育、建筑、交通、文旅、医疗、金融、政务等重点领域，到 2027 年建设 30 个以上行业高质量数据集，支撑人工智能大模型应用；天津以“算力+数据+算法”为路径，计划开发公共数据集、行业数据集、场景化数据集，到 2026 年推动打造 2000 个高质量行业数据集，重点覆盖工业制造、港口物流、生物医药领域。

多地通过激励机制激活市场活力。如贵州设立高质量数据集奖励机制，每年安排资金总额不超过 500 万元，对训练使用量、数据质量等综合排名前 10 的市场主体给予奖励；武汉设立人工智能政策专项资金，支持企事业单位建设人工智能产业高质量数据集，按建设投入成本的 30%给予不超过 200 万元的奖励；浙江支持建设行业级可信数据空间、高质量数据集，支持建设高端数据标注平台，鼓励打造产学研用联动的创新载体，建设一批成效明显、特色鲜明的数据标注基地，开发高质量数据集。

具体政策如下:

表 1-1 各地高质量数据集建设相关政策

省市	政策文件	发布单位	发布年份

上海市	《立足数字经济新赛道推动数据要素产业创新发展行动方案（2023-2025 年）》	上海市人民政府办公厅	2023
广东省	《广东省关于人工智能赋能千行百业的若干措施》	广东省人民政府办公厅	2024
江苏省	《数字经济高质量发展三年行动计划（2025-2027 年）》	江苏省人民政府办公厅	2025
天津市	《天津市算力产业发展实施方案（2024-2026 年）》	天津市人民政府办公厅	2024
贵州省	《关于促进全国一体化算力网络国家（贵州）枢纽节点建设的若干激励政策》	贵州省数据局、发改委等 8 部门	2023
武汉市	《<武汉市促进人工智能产业发展若干政策措施>相关支持高质量数据集建设和数据产品利用资金管理办法（试行）》	武汉市数据局	2025
浙江省	《关于支持人工智能创新发展的若干措施》	浙江省人民政府	2025

（二） 高质量数据集定义

根据《高质量数据集建设指引》的定义，高质量数据集是指经过采集、加工等数据处理，可直接用于开发和训练人工智能模型，能有效提升模型表现的数据的集合。

高质量数据集覆盖多个重点行业领域，凭借高技术含量、高知识密度、高价值应用三大核心特征，在数据要素生态中占据重要地位。高技术含量体现在数据集质量提升已进入技术驱动阶段，依托自动化标注、AI 辅助标注等先进技术，增强对业务场景支撑作用。高知识密度体现在通过构建跨领域的知识图谱，推动多学科知识融合与深度利用。高应用价值反映为能够切实解决产业发展中的实际问题，如在提升决策效率、优化资源配置、改善用户体验等方面发挥重要作用。

（三） 高质量数据集分类

从数据用途来看，高质量数据集包括通识数据集、行业通识数据集和行业专识数据集。通识数据集是面向社会公众的通用知识，具有广泛性、通用性等特点，覆盖多个领域，主要用于通用大模型的训练。行业通识数据集是面向特定行业或领域的通用知识，具有行业普适性、共识性等特点，主要用于行业大模型的训练。行业专识数据集是根据行业企业

自身业务场景和需求收集的专识数据集，具有场景针对性、定制化等特点，能够为行业企业提供高度个性化训练数据资源，主要用于业务场景大模型的训练。

从数据模态来看，高质量数据集包括文本、图片、音频、视频等单模态数据集及多模态数据集。单模态数据集中，文本数据集是指以书面语言为主要形式的数据集合，涵盖自然语言、符号序列等，用于支撑自然语言处理任务及语言模型的训练等；图像数据集是计算机视觉的核心资源，包括照片、绘图和数字生成的图像，多应用于医学诊断、工业检测、交通识别等领域；音频数据集则由声音信号组成，涵盖语音、音乐等多种声学形式，支持语音识别、情感分析等场景；视频数据集注入时空维度信息，支撑从通用动作识别到自动驾驶、机器人交互等应用场景的广泛研究；多模态数据集则指整合两种及以上模态的数据资源，用于支撑复杂任务中的跨模态感知与理解，如图文生成、人机对话、视频理解等应用场景。

从训练阶段来看，高质量数据集包括预训练数据集、微调数据集、评估数据集。预训练数据集是以无监督或自监督学习方式让模型获取通用特征与知识的数据集，具有规模庞大、无需标注和领域广泛等特点，并涵盖网页、书籍、社交

媒体与百科全书等多种来源。微调数据集是为优化模型特定任务的处理能力，专门用于对模型进行微调的数据集，具有规模较小、标注质量高、任务特定等特点，通常由一系列的问答对组成。评估数据集是专门用于验证和衡量模型性能和泛化能力的数据集，通常需要人工高精度标注，以确保测评结果可信，具备独立性、代表性、时效性等特点。

表 1-2 高质量数据集分类

分类维度	类别名称	应用场景
数据用途	通识数据集	面向社会公众的通用知识，用于通用大模型训练
	行业通识数据集	面向特定行业或领域的通用知识，用于行业大模型训练
	行业专识数据集	根据行业企业自身业务场景和需求收集的专识数据集，用于场景模型训练
数据模态	单模态数据集	语言模型训练、医学诊断、自动驾驶等
	多模态数据集	图文生成、人机对话、视频理解等
训练阶段	预训练数据集	通过无监督和自监督学习，让模型学习通用特征与知识
	微调数据集	对大模型进行特定任务性能的微调
	评估数据集	验证和衡量模型性能和泛化能力

二. 高质量数据集发展现状

（一） 国外高质量数据集发展现状

1. 多元主体共建，开源生态驱动市场质效提升

近年来，美国、欧盟等全球主要经济体加快培育高质量数据集，已形成成熟的供应链条和市场生态。人工智能发展对数据“海量、多模态、高质量”的需求，使国外形成政府机构、高校和科研机构、非营利组织和企业协同建设的格局，市场呈现多类型主体协同参与、开源共享为主导的商业模式。市场多主体协同下，国外高质量数据集市场在规模、质量、模态等方面持续提升:

政府机构层面主导公共数据开放，如美国构建了覆盖联邦、州及市三级政府的数据开放网络，以美国政府开放数据平台 Data.gov 为标杆，截至 2024 年已累计整合发布 29 万余个多领域数据集，覆盖农业、气候、教育等领域，通过统一门户实现跨部门数据共享；欧盟开放数据平台则汇聚了欧盟机构及成员国的农业、能源等跨领域数据资源，形成区域性数据协作网络。

高校和科研机构层面聚焦文本型为主的知识数据集，如宾夕法尼亚大学打造美国国家语料库（ANC），沉淀约 9 万

个文本样本及 1 亿的单词；牛津大学打造英国国家语料库（BNC），沉淀超 900 个数据集及 1 亿单词，为语言学研究及大模型基础训练提供专业知识型数据支撑。

非营利组织建设以爬取网页数据为主的通用数据集，如 Common Crawl 等通过系统性爬取网页数据，每年汇集 200-300TB 通用数据，以免费开放模式，为大模型训练提供支撑。

企业层面依托自有数据深耕垂直领域，如谷歌等科技企业基于业务积累数据，经脱敏、标注等处理后形成标准化数据产品，通过 API 接口或订阅服务实现数据资产的价值转化。Google 依托自有业务场景，推出 Open X-Embodiment（100 万数据集），用于机器人任务训练等垂直场景，填补行业数据空白。

表 2-1 高质量数据集建设与应用需求

序号	建设方	数据集名称	数据集类型	数据规模	数据来源	收费情况	需求领域
1	美国政府	美国 Data.gov	行业数据集	293,063 个数据集	公共数据、行业数据和个人数据	免费	应用于农业、经济、医疗、教育等领域
2	宾夕法尼亚大学等多个机构合作开发	美国国家语料库（ANC）	通用数据集	约 90000 个文本样本，约 1 亿个单词	网页数据、期刊数据、书籍	免费	语言学研究
3	牛津大学出版社	英国国家语料库（BNC）	通用数据集	超过 900 个数据集、超过 1 亿单词	网页数据、期刊数据、书籍	免费	语言学研究
4	布朗大学	OpenWebText2	通用数据集	190 亿字节	网页数据	免费	大模型语言训练
5	开源组	The Pi	通用数	825.18G	公共数	免费	大模型语

	织 Eleuther AI	le	据集	B	据、网页数据		言训练
6	非营利组织	Common Crawl	通用数据集	200-300 TB	网页数据	免费	大模型语言训练
7	Google 公司	Open Images	通用数据集	900 万张图像	网页数据	免费	视觉研究
8	Google 公司	YouTube-8M	通用数据集	800 万个 YouTube 视频链接	网页数据、自有数据	免费	计算机视觉、分类识别
9	Google DeepMind 联合 33 家学术实验室	Open X-Embodiment	行业数据集	100 万数据集	自有数据	免费	机器人特定任务训练
10	Reddit 公司	Reddit	通用数据集	/	自有网页数据	收费	大模型训练

此外，在市场机制方面，以开源为核心的组织与协作模式已成为驱动数据集高效流通与持续增值的关键力量。以GitHub为代表的开源平台，汇聚了政府、企业、科研机构发布的数据集与工具链，形成了全球化的协作网络。例如，以Google DeepMind为代表的企业研究机构，通过开源Open X-Embodiment数据集，吸引了33家科研实验室联合补充数据、验证模型，共同推动RT-X模型性能迭代；以Eleuther AI为代表的非营利性开源组织，则以社区协作方式整合科研数据、公共数据及其他公开网络数据，构建并开源了825GB的The Pile数据集，使其成为大模型训练的重要通用语料库。这种开源协作模式不仅降低了数据获取成本，还通过社区反馈持续优化数据质量，形成了共建、共享、迭代的市场闭环。

2. 基础技术领先，构建完整技术生态体系

国外在高质量数据集建设技术上生态体系较为完善。

在数据采集环节，欧美企业在高精度传感器技术和智能爬虫系统方面保持领先优势，如Google开发的分布式爬虫框架能够高效处理PB级网页数据，同时确保数据采集的合规性。

数据治理技术方面，国外研究机构在基础算法层面持续创新。斯坦福大学提出的自动数据清洗框架支持智能异常检

测和修复，微软研究院开发的数据增强算法能够保持语义一致性的同时生成多样化样本。特别是在数据合成技术领域，NVIDIA 的生成式 AI 技术能够创建高度逼真的合成数据，有效解决训练数据不足问题。

数据标注技术呈现出高度自动化特征。Scale AI 等公司开发的标注平台集成了最先进的计算机视觉和自然语言处理模型，在保证质量的前提下实现标注效率的显著提升。多模态标注技术尤其成熟，OpenAI 开发的跨模态对齐工具能够精准实现图文、音视频等多模态数据的关联标注。

质量评估技术体系方面，建立了系统的评估标准和工具链。Google 的 TensorFlow Data Validation 库提供完整的数据质量监控方案，Hugging Face 牵头制定的数据集评估标准已成为行业共识。这种完善的技术生态体系为国外高质量数据集建设提供了坚实基础。

（二） 国内高质量数据集发展现状

1. 数据集供给规模快速扩展，类型持续丰富

我国高质量数据集供给能力显著增强，逐渐形成以国家基地为枢纽，覆盖多行业的规模化供给体系。在政策引导与市场需求的双重拉动下，数据集在体量上实现跃升，类型结构也持续优化。

国家数据局统筹推进数据资源体系建设，截至 2025 年 3 月，全国已建成 7 个数据标注基地，构建涵盖医疗、工业、教育等关键领域的 335 个高质量数据集，标注总规模达 17,282TB，有效支撑 121 个国产大模型的研发与迭代，带动数据标注相关产业产值超过 83 亿元。2025 年 4 月召开的第八届数字中国建设峰会上，国务院国资委集中发布涵盖智慧能源、工业制造、绿色低碳、金融服务等 10 余个行业的 30 项高质量数据集建设成果，进一步丰富了行业级数据资源供给。2025 年 9 月，国家数据局通过组织开展高质量数据集典型案例征集工作，累计遴选出 104 个典型案例，地域分布覆盖全国 59.3%的地区（19 个区市），并广泛渗透至智慧能源、科学研究、交通运输、工业制造等 17 个重点行业领域与智能驾驶、低空经济等 5 个创新领域。此外，各地数据交易平台也着力推进高质量数据资源的市场化供给与流通应用。例如，截至 2025 年 4 月，贵阳大数据交易所累计发布 939 个多模态数据集，覆盖金融、工业、医疗、商贸等重点领域，为大模型厂商提供体系化、合规化的数据资源支持。

从数据集体量看，国内主流大模型预训练数据集普遍达到 10-18TB 量级，指令微调、强化学习等后训练数据规模亦显著提升。以 DeepSeek V3 为例，其指令微调数据规模已达

150 万条，显示出国内高质量数据集供给已逐步对接大模型训练的实际需求。

在数据集供给类型结构方面，数据集建设呈现出专业化的特征。除通用和行业数据集外，面向复杂场景的“推理数据集”逐渐成为发展重点。数学推导、代码生成、科学计算、逻辑推理等思维链数据被广泛构建并应用于模型训练中。此类数据显著提升模型的认知与推理能力，也体现了国内数据集建设正从“规模扩张”向“能力导向”转变。

2. AI 场景驱动数据集需求激增，规模快速扩张

国内人工智能产业逐步成熟，AI 模型在算法性能、泛化能力及多模态处理方面持续突破，推动其在多个行业场景拓展落地。据艾瑞咨询数据，2023 年，中国 AI 基础数据服务市场规模已达 45 亿元，预计将以 30.4%的年复合增长率持续扩张，至 2028 年整体规模将突破 170 亿元，显示出强劲的市场活力和发展潜力。

这一进程中，高质量数据集的需求呈现爆发式增长，应用场景已从互联网、安防等传统领域，快速纵深渗透至能源、交通、工业、医疗等国民经济关键行业。

当前，高质量数据集需求呈现出显著的行业化、场景化特征。根据中国信通院资料，目前我国已涌现出上百个高质

量数据集典型案例与先行先试项目，横跨 20 余个行业。如在能源领域，国家管网集团的天然气管网运行数据集、南方电网的配电网智能规划多模态数据集，支撑着能源调度的智能化与安全运行；在材料与工业领域，中铝集团的铝合金材料金相组织图片数据集、中国电信的工业纺织线路检测数据集，服务于生产质检与流程优化；在交通领域，中国交通的交通基础设施安全监测数据集助力智慧交通建设；在医疗领域，中国联通的肺结核影像标注数据集则直接赋能临床辅助诊断。此外，在应急管理、金融风控、公共服务等多个领域，高质量数据集都已成为驱动 AI 赋能、解决行业痛点的核心燃料。这些典型案例表明，高质量数据集的需求已不再局限于技术研发层面，而是深度融入各行各业的生产运维、智慧交通、灾害预警、临床医疗等具体业务场景，呈现出规模化、纵深化的全面发展态势。

3. 应用技术突破，技术生态体系加速构建

我国高质量数据集建设技术发展坚持市场需求导向，在产业应用方面取得显著进展，技术路径注重实用性和规模化落地。

数据采集技术紧密结合国内网络环境特点。如百度、阿里巴巴等企业开发的爬虫系统针对中文互联网生态进行了

深度优化，能够高效采集并处理中文特色数据内容。在传感器数据采集方面，华为、海康威视等企业在图像和视频采集技术上达到国际先进水平。

数据治理技术注重实用性和效率。如 DeepSeek 开发的数据清洗工具针对中文文本特点进行了专门优化，在错别字纠正、语法纠错等方面表现优异；商汤科技开发的图像增强算法在医疗影像、工业质检等专业领域取得良好效果。

数据标注技术形成独特的人机协同模式，形成了“机器预标注-人工精细校对-质量反馈优化”的闭环流程，有效提升标注效率并保障数据质量。百度数据标注平台在中文 NLP 任务中标注准确率及成本控制效果显著。在多模态标注方面，腾讯开发的跨模态标注工具（基于腾讯混元多模态 AI 的智能视频分析与创作助手）支持大规模视频内容分析，为短视频行业发展提供技术支持。

质量评估技术注重标准化和流程管控。随着数据标注基地的规范化建设，逐步形成统一的质量评估标准和操作规范。在评估指标和自动化评估工具方面，中国信息通信研究院人工智能所开发 60 个数据质量检验算子工具包，指标自动化评测率达到 75%，形成了数据质检工具链。在质量评估技术

方面，阿里巴巴开发的数据质量评估平台 DataWorks 支持实时质量监控和预警，确保数据集质量符合模型训练要求。

（三） 我国高质量数据集发展难点堵点

当前，我国高质量数据集建设在政策驱动与场景牵引下已取得显著进展，仍面临数据供给不足、技术不成熟、数据来源单一和运营不完善等突出问题。

1. 数据供给不足，高质量数据稀缺

当前，我国高质量数据集发展仍然面临供给侧产量有限、质量参差的挑战。

数据规模方面，国内中文数据集虽已初步积累，但与国际先进水平相比，在规模体量、覆盖面和精细化程度等方面仍存在明显差距。以自然语言处理（NLP）为例，英文领域有 Common Crawl、Wikipedia 等大规模开源数据集，而中文领域现有资源在规模、多样性和更新频率仍难以满足需求。据 AI 应用开放社区 Hugging Face 统计，中文开源数据集数量仅为英文开源的 11%，且多集中于基础文本领域，缺乏高质量的多模态标注数据。

数据质量方面。现有数据质量评估大多考察的是完整性、一致性等基础维度，缺乏对于行业知识内涵和业务场景的深度挖掘，许多数据集存在语义一致性差、时效滞后等问题。

尤其在工业、医疗、法律等专业领域，需深度融合行业知识的高质量标注数据极为稀缺。

2. 技术不成熟，关键环节存在短板

在高质量数据集建设的技术链条中，我们在标准体系、工具链完备性和先进性等方面仍存在突出瓶颈，制约了数据集生产的效率、规模与质量。

在数据治理环节，自动化与智能化水平不足，面对海量多源异构数据，尤其是非结构化数据，自动化清洗、去重、质量分级等基础处理仍依赖传统方法，效率与精度有限。在数据标注环节，针对医疗影像、法律文书等专业强、细粒度高的复杂场景，机器预标注质量难以满足要求，对专业人工依赖度高，且支持复杂逻辑推理与多模态语义对齐的工具链尚不成熟。在质量评估环节，缺乏能动态感知数据分布对模型训练影响、识别隐性偏差的评估体系，导致数据集建设与模型优化脱节。这些技术短板共同影响了数据生产的效率、规模与最终质量，阻碍了数据要素价值的充分释放。

3. 数据来源单一，开源生态培育不足

相较于国外依托 GitHub 等平台形成的政府、企业、高校、非营利组织及开发者社区多元协同、共建共享的成熟开源生态，我国高质量数据集建设仍主要依赖于互联网企业与

科研机构，资源整合效率较低，数据集来源单一导致无法满足大模型多样化训练需求。

一方面，公共数据开放受限。政府部门、科研机构虽掌握大量高质量数据，但公共数据受制于数据安全与管理机制等原因，开放共享进展缓慢，而高校和科研机构的数据集多聚焦科研场景，向产业界开源共享的比例不足 30%。另一方面，企业数据壁垒较高。出于商业竞争考虑，企业间数据孤岛现象严重，难以汇聚形成大规模行业数据集。

同时，社会力量参与不足。类似国外 Common Crawl、Eleuther AI 的非营利开源组织在国内稀缺，现有平台（如数据交易所、标注基地）也因企业数据封闭倾向，难以有效整合资源。国内开源文化尚未深度渗透，多数机构更注重数据“私有化”保护，难以形成像国外 The Pile 等的社区迭代效应，阻碍了高质量数据集的规模化、可持续发展。

4. 运营不完善，制约数据价值释放

当前，我国高质量数据集建设在政策驱动与场景牵引下已取得显著进展，但运营能力不足正成为制约其价值释放的核心瓶颈。一方面，多数主体存在“重建设轻运营”倾向，缺乏全生命周期质量管控与动态优化机制，数据质量依赖人工检查、安全风险防控薄弱，难以通过模型反馈实现“数据

-模型-业务”闭环迭代；另一方面，运营管理体系不完善，用户需求响应滞后，且数据资产化与产品化路径模糊，对内难以支撑战略决策，对外无法释放要素价值，导致85%的交易所挂牌数据集“有货无市”，最终制约了数据要素从“资源”向“资产”的实质性转化。

三. 高质量数据集建设路径

（一） 高质量数据集建设流程

高质量数据集的建设是一个覆盖数据从“产生”到“应用”全生命周期的系统工程，主要包括数据规划、数据采集、数据预处理、数据标注和模型应用等环节。

图 3-1 高质量数据集建设流程

需求分析是数据集建设生命周期中的关键阶段，其核心目的是确保所采集和处理的数据能够严格满足前期明确的数据需求，并为实现人工智能应用目标提供系统性支持。根据不同的建设模式进行差异化规划，企业可构建更高效、精准的数据集，支撑业务创新与技术落地。

数据规划环节将数据需求的输入，转化为指导后续所有工作的具体方案。其核心目的是确保所采集和处理的数据能够严格满足前期明确的数据需求，并为后续的数据采集和处理提供基础。数据规划的难点在于需求目标的拆解、多源数据的融合、以及需求动态变化等问题，因此需保持规划的弹

性，在实施中逐步细化调整，并建立贯穿项目始终的反馈优化机制。

数据采集是高质量数据集建设的基石，其核心作用在于为整个数据价值链提供高质量、多样化且合法合规的原始原料。在数据采集过程中，通过从内部系统、外部合作、公开数据集等多源渠道，系统性地汇集文本、数字、音频、视频等多模态数据，确保所获数据能够精准锚定核心业务场景。需要注意的是，数据采集环节必须严守安全与合规的红线，严格遵循相关法律法规，对涉及个人隐私的数据进行脱敏与匿名化处理，确保数据来源与获取方式的合法性。数据采集的广度、精度与合规性，直接决定了后续数据预处理、标注及模型训练的效果上限。

数据预处理涵盖数据转换与清洗、增强与合成，以及脱敏处理，是构建高质量数据集的关键。数据清洗通过对缺失值、异常值和重复值的系统处理，并统一数据格式和类型，提升数据准确性、完整性和一致性，为后续分析奠定基础。数据增强和合成技术通过变换现有数据或生成新数据来弥补数据稀缺和不平衡问题，但需基于对原始数据分布和业务需求的理解进行评估。数据脱敏则通过替换、泛化等手段保护隐私，确保数据可用性与安全性的平衡，尤其在金融、医疗领域，企业需根据业务特性构建合适的脱敏体系。这些步骤共同作用，以确保数据集的质量和合规使用。

数据标注是指对原始数据进行筛选、分类、标记和注释等加工处理，将其转化为机器可读的标准化格式的过程。相关组织需制定统一的标注规范与流程，进行变量赋值，再进行数据标记注释，经过数据质检，最终形成可用于机器学习的高质量数据集。在高质量数据集建设中，数据标注通过建立统一的数据语义框架，将原始数据转化为机器可理解的标准化格式，为模型训练提供精准、可操作的数据输入，从而提升数据的有序化程度和场景适配性，是激活数据要素价值、支撑人工智能应用落地的关键基础。

模型验证是承上启下的关键环节，其核心作用在于检验数据集的质量是否足以支撑模型达到预期目标，并为上游数据工作的迭代优化提供精准反馈。模型验证是数据质量的“试金石”，直接反映了数据集在规模、多样性、标注准确性等方面的质量水平。如果模型性能未达预期，往往意味着上游的数据采集、清洗或标注环节存在缺陷。其次，它构成了质量闭环迭代的反馈枢纽。验证中发现的问题会被精准反馈至数据规划、预处理、标注等上游环节，驱动数据团队有针对性地进行数据增强、偏差校正或重新标注，从而在迭代循环中不断提升数据集的整体质量与应用价值。

（二） 高质量数据集建设模式

随着人工智能技术在各行业逐步渗透，数据集的构建模式也呈现出多元化发展趋势，不同的建设路径直接影响着数据资产的价值实现方式。高质量数据集建设模式主要有四类：一是政府牵引模式，借助政府数据开放，促进高质量数据集流通；二是需求拉动模式，发挥行业企业主体作用，深化数据开发与应用；三是服务供给模式，依靠数据服务企业，增加高质量数据集供给；四是生态协同模式，依托开源生态力量，推动高质量数据集迭代。

政府牵引模式。该模式下政府引导公共数据开放，推动公共领域数据流通共享，高效汇聚形成基础性、通识类数据集。比如各地目前建设公共数据开放平台，汇聚了大量数据集。目前各地公共数据平台上开放的有效数据集总数逐年增长，从 2017 年的八千多个数据集增长到 2025 年的 48 万多个，复合年均增长率为 66%[1]。

服务供给模式。数据服务商依托其专业的采集标注能力和成熟的生产流程，为市场提供数据集供给服务。这类企业以大量、多源异构数据为基础，通过主动的数据探索、关联分析与价值挖掘，反向发现潜在的业务需求或优化方向。该模式的优势在于能快速形成大规模数据资产，为后续模型探

[1] 数据来源：《2025 中国地方公共数据开放利用报告——省域》

索提供丰富素材，特别适合通用大模型、预训练模型等需要海量多样化数据的任务。

需求拉动模式。其核心特征在于以终端行业或企业的实际应用需求为根本驱动力，通过解决特定业务场景中的真实问题，逆向牵引数据资源的规划、采集与治理，能够有效避免数据冗余或缺失问题。需求拉动模式最显著的优势是其高度的针对性和实用性，特别适用于垂直行业应用，能确保数据集的建设与业务价值直接挂钩。同时，该模式高度依赖领域专业知识，可能导致建设成本和周期相对较高。例如，在开发医疗影像诊断系统时，通过组织专业医生团队进行精细标注，构建具备临床价值的高质量数据集。

生态协同模式。依托开放社区和协作机制，汇聚广泛的社会力量共同建设和完善数据集。这类模式主要适用于通识类高质量数据集的构建，具有成本低、覆盖广、迭代快的突出特点。然而，开源社区数据集也存在数据质量参差不齐、数据重复严重、数据安全性较差等问题。因此，该模式更适合对数据安全性要求不高但需要广泛覆盖度的基础性数据集建设，如自然语言处理、计算机视觉等领域的通用训练资源。

图 3-2 高质量数据集建设模式

（三） 高质量数据集建设核心环节

1. 规划先行：定义数据集的建设方向与范围

高质量数据集的建设，从需求出发进行精准设计，明确为什么建、建什么。数据需求一般包含业务驱动和数据驱动两种类型业务驱动需求从具体的业务需求出发，以解决实际问题为目标。数据驱动需求则以现有数据资源为核心，通过系统性挖掘与整合释放潜在价值。

（1）业务驱动聚焦解决业务痛点

业务驱动的需求分析始于业务痛点的深度拆解，随后围绕这些需求设计数据采集方案，明确所需的数据类型、质量标准和标注规则。在数据采集阶段，会优先从内部业务系统（如交易日志、用户行为记录）提取高价值信息，同时结合外部数据补充，并通过闭环反馈机制持续验证数据对业务指标的提升效果。

因此，业务驱动的数据规划强调数据与业务需求的深度

结合，以解决具体业务问题为核心，紧密围绕业务场景展开。这类需求通常需要数据部门与业务部门密切协作，明确核心问题及所需的数据类型，进而梳理业务场景对应的数据源及其质量要求，例如从CRM系统或用户行为日志中提取关键信息，并在此基础上开展针对性的数据清洗和标注工作，最终形成可直接支撑业务场景应用的专项数据集。

（2）数据驱动聚焦挖掘数据价值

数据驱动的需求分析通常从多源数据的汇聚与探索性分析开始，例如对海量用户行为日志进行聚类或关联规则挖掘，识别出隐藏的模式或趋势，再据此设计数据增强与标注策略。

数据规划则以数据资源本身的质量和可用性为导向，重点关注数据的内在特性和潜在价值。该模式通常从数据资源的全面盘点和评估入手，筛选具有潜在利用价值的内部或外部数据，绘制系统化的“数据资源地图”，再通过规范的数据治理流程，逐步将原始数据转化为标准、可靠的数据资产，为后续的探索性分析和模型构建提供支持。数据驱动型项目往往需要建立自动化流水线来处理异构数据，并通过众包或合作生态扩大数据覆盖范围。

2. 数据采集：确保数据来源的合规性与多样性

通过整合多元化数据来源，并采用API接口、ETL工具、实时流处理技术等技术进行采集，同时在数据收集前实施合

法性审查，以系统性地保障数据的合法合规性与多样性。

（1）自有数据采集

自有数据采集是指企业或组织利用自身业务系统、设备或渠道，主动、直接地获取和生成原始数据的过程。数据直接产生于内部的生产、运营、服务等环节，如用户交易记录、设备传感器日志、应用程序操作行为等，因此与核心业务场景的关联性最强。自有数据的优势是来源可控、与业务高度相关、格式相对规范、且通常无需担心外部合规风险，是构建高质量、高价值数据集的核心和首选方式。

（2）数据交易共享

数据交易作为一种高效的数据采集方式，其核心在于通过市场化机制，将经过加工和标准化的数据产品作为商品进行流通。它不仅是获取外部数据资源的重要途径，更在连接数据供给与需求、激活数据价值方面发挥着关键作用。当前，我国高质量数据集交易呈现爆发式增长态势，已成为驱动人工智能产业发展的重要引擎。据国家数据局统计，截至 2025 年 6 月，全国各地高质量数据集全国各地交易机构挂牌 3364 个高质量数据集，总规模达到 246PB，累计交易额接近 40 亿元。

（3）网络数据爬取

网络数据爬取作为一种高效的数据采集技术，其核心内涵是通过编写自动化程序，模拟人类浏览网页的行为，按照

预设的规则从互联网上系统地抓取公开可访问的信息。网络数据爬取的优势在于通过自动化与高效率的数据获取，能够不间断地处理海量信息，将分散在不同网站、格式各异的非结构化网络内容，如新闻、商品价格、社交媒体帖子等转化为集中、规整且可分析的数据集，从而极大地弥补了单一机构内部数据在广度和多样性上的不足。但在使用网络爬取技术时，必须高度重视其法律和伦理边界。如中国《个人信息保护法》则要求即便对公开个人信息进行处理也须取得单独同意，数据的合规处理成本大幅增加。

（4）数据合成增强

合成数据通过生成式算法模拟现实数据分布，在缓解数据获取瓶颈与隐私约束间构建技术通路。目前已在自动驾驶、医疗及金融等领域实现工业级应用,例如在自动驾驶领域，合成数据可用于模拟各种复杂路况，帮助训练自动驾驶模型。同时，合成数据也存在真实性和算法偏见等问题，亟待构建数据治理体系。我国《生成式人工智能服务管理暂行办法》要求合成数据服务提供者存留完整技术日志，且不得使用未完成安全评估的基础模型构建金融、医疗领域标注数据集。

3. 数据标注：连接数据与应用场景的关键桥梁

在高质量数据集构建中，数据标注是将原始数据转化为机器可理解信息的关键环节，直接影响模型性能。通过分类、标记等操作建立统一语义框架，数据标注不仅提升数据的结

构化程度和准确性，还能针对特定场景定制标签体系，解决数据与需求的适配性问题。随着大模型发展，数据标注正从人工密集型向 AI 辅助的智能化转型，成为释放数据要素价值的基础。

企业根据自身的数据特点和能力，通过自建团队、外包、众包等方式完成数据标注。

图 3-3 数据标注服务模式

（1）转包模式

转包模式指公司在数据采集、加工等环节，将任务外包给第三方数据标注公司和团队，由其负责组建团队、完成标注任务，并按照合同约定交付标注数据。

与自建团队模式相比，转包给第三方无需管理和培训成本，能以较低价格实现标注任务，同时可以通过商业合同约定交付质量和验收标准，降低标注风险。但转包模式无法保障企业数据安全，存在数据外泄风险，比较适用于标注任务

量较大、对标注质量有一定要求但企业自身资源有限的企业，包括中小型科技企业、互联网创业公司等。

（2）众包模式

众包模式是数据标注常见的生产模式之一，主要指企业借助众包平台，将数据标注任务分解为众多小任务，分发给大量分布在不同地域的众包参与者进行标注。企业根据参与者完成任务的数量和质量支付相应报酬。

众包模式的优势在于灵活响应标注需求，且企业无需承担众包人员培训、办公场地等费用，成本较低。然而，该模式也存在一些局限性。由于众包参与者大多缺乏专业标注知识，导致标注质量难以控制，往往需要通过多级审核和动态评级机制来提升标注质量，同时标注工期难以保障，数据安全性也较小。

综合来看，众包模式适用于对标注质量要求相对不高、数据量巨大且需要快速完成标注的企业，例如品牌化的专业性服务商和科技互联网巨头等。

（3）自建团队

自建团队模式主要指企业依靠自身的资源和能力，组建直属的标注团队，从人员招聘、培训，到标注流程的制定、管理以及标注工具的研发或采购，均由企业内部完成。团队成员通常是企业的正式员工，在企业内部的办公环境中开展标注工作。

该模式具备多方面优势，如标注团队易管理、标注质量有保证、标注工期可控及数据安全性高等，且内部标注团队可以通过知识沉淀，提高数据标注效率。然而，其不足之处也较为明显，前期投入成本较高，且团队建设难度较大。医疗、国防、金融等对数据安全要求极高的行业通常倾向于自建团队。同样，大型互联网科技公司由于对标注质量和定制化程度要求高，也会选择自建标注团队。

4. 质量测评：确保数据集安全可靠的关键标尺

开展系统性、规范化的数据集质量评价，是判断数据集是否达到模型训练或场景应用标准的基本路径，也是推动高质量数据资源建设的核心抓手。通过数据集的质量测评，可以有效倒逼数据生产和管理环节提质增效，全面提升数据资源的可用性、可信度与应用价值。

（1）数据质量检测流程

评估准备阶段是评价工作的基础，重在明确"为何评、评什么、怎么评"。首先需清晰界定目标数据集的基本信息、业务应用场景及核心评价目标，例如是为支撑精准营销还是信贷风控，这直接决定了后续评估的侧重点。随后，需划定评价的范围和具体对象，是评估单个数据表还是一整套数据产品。在此基础上，制定出具体的评价策略与技术规范，同时组建具备业务理解、数据技术和统计知识的专业团队，并配备必要的自动化评价工具与数据环境，为后续工作奠定规范

性与一致性的基础。

核心环节是质量评估指标体系构建与实施。通常需围绕完整性、准确性、一致性、时效性、唯一性等核心维度设计分层分级的指标，并为每项指标明确定义计算规则、测量方法和评分标准。在实施过程中，可结合自动化检测工具进行大规模扫描，并针对复杂业务逻辑辅以人工抽样核查，从而确保评估工作能够全面、系统且高效地执行，真实反映数据质量现状。

最终环节是问题分析与改进。该阶段需要将各项指标的评测结果进行多维度加权汇总，形成量化的质量总分与等级，从而对数据资产健康状况形成整体判断。更重要的是，需深度分析质量问题的根因，并据此提出具体、可操作的改进建议，例如修复数据采集接口、优化清洗规则或完善管理规范。通过定期回顾质量改进效果，优化质量规则和流程，形成一个可持续优化的闭环管理过程。

（2）数据质量评价标准

根据《高质量数据集 质量评测规范》高质量数据集的质量要求应覆盖说明文档、数据质量及模型应用三个维度。

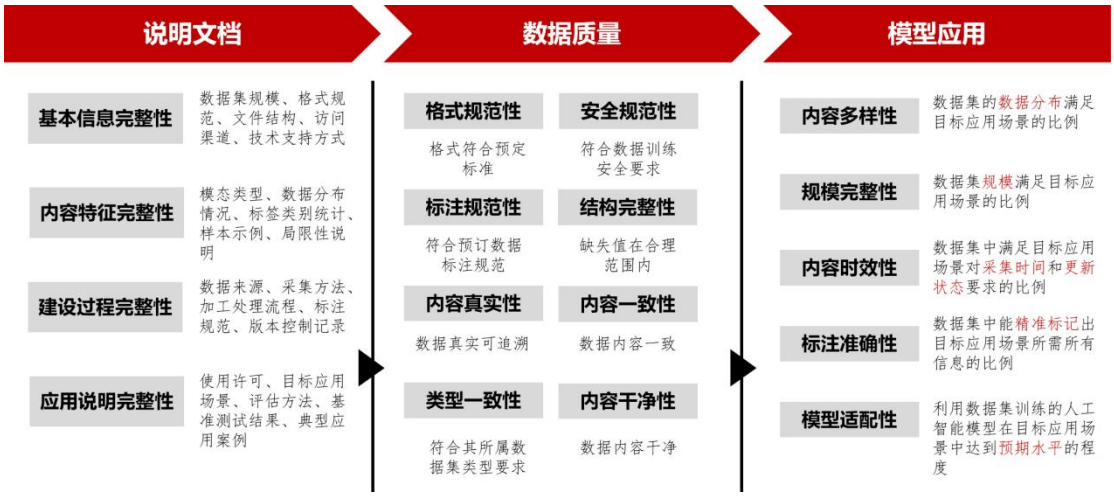

图 3-4 数据质量测评规范

（四） 高质量数据集建设典型案例

作为数字经济赋能实体经济的重要践行者，中国电信股份有限公司浙江分公司（以下简称“浙江电信”）依托云网融合与数据治理优势，联合某集团，针对传统水泵安装数据碎片化、标准不统一、质量追溯难等痛点，严格遵循“需求分析—数据规划—数据采集—数据预处理—数据标注—模型验证”的六步建设流程，打造覆盖安装全流程的高质量数据集，为智慧泵房建设提供坚实数据支撑，也为工业装备安装类数据集建设提供可复用经验。

案例建设聚焦传统水泵安装的核心短板：一是人工审核效率低下，面对日均数百单的安装量，工单积压严重拖慢结算进度与客户回访时效；二是审核标准缺乏统一规范，不同审核人员的主观判断差异导致漏判误判频发，直接推高返工成本并影响客户体验；三是虚假报工行为难以识别，部分安

装人员存在上传非现场照片、未完工即标记“完工”等违规操作，缺乏有效监管手段。这些问题给服务质量与品牌声誉带来潜在风险，也制约了安装服务的数字化升级，亟需通过高质量数据集建设破解人工管控的效率、标准与监管难题。

在具体建设流程中，需求分析阶段通过业务调研，明确水泵安装审核的效率、标准与监管痛点，锁定智能化审核核心需求；数据规划阶段梳理安装全环节数据，划分影像、参数、标签等数据类别，制定统一的采集范围、格式规范与质量评估标准；数据采集阶段通过传感器自动采集设备运行与安装过程数据，结合标准化表单补录工艺关键信息，整合形成原始数据集；数据预处理阶段采用自动化工具过滤重复数据、修正异常值，完成数据格式标准化与对齐；数据标注阶段采用 AI 预标注与行业专家复核结合的方式，围绕审核核心维度构建标签体系，确保标注准确性；模型验证阶段将标注数据输入智能审核模型，测试适配性并根据运行效果优化数据，结合多级质检机制，保障数据集质量达标。

表 3-1 水泵安装高质量数据集建设流程

阶段	子任务	具体工作内容
数据需求	确定数据集内容和范围	明确具体需要哪些数据，包括数据格式、数据规模和数据范围等。
数据采集	制定采集规范	制定数据采集标准
	执行数据采集	按照采集流程、数据格式要求、权限规

		范等进行数据采集操作，确保采集过程合规且数据可用。
数据预处理	数据清洗	对采集到的数据进行去重、缺失值处理、异常值剔除等操作，提升数据质量。
	数据标准化	将数据转换为统一的格式、单位、尺度等，便于后续标注和模型使用。
数据标注	确定数据标签	基于业务需求和模型目标，确定需要标注的标签体系（如分类标签、标签选项等）。
	组建与培训标注团队	招募标注人员，开展标注规则、工具使用等培训，确保标注人员理解标注要求。
	执行数据标注	标注人员按照流程和规则对数据进行标注操作。
数据集质量评估	设定质量评估指标	细化质量评估指标，明确各指标的计算方法和达标阈值。
	建立质量检查机制	制定日常检查、阶段性检查、最终检查的具体流程、抽样比例、检查人员分工等。
	执行质量检查并处理问题	按照质量检查机制对标注后数据集进行检查，对发现的问题组织标注人员修改完善，直至达标。
模型应	高质量数据集	用于人工智能模型开发和训练，对模型

用	模型应用验证	性能是否达到预期进行评估，以验证数据集是否满足要求

该数据集的建设可精准破解传统水泵安装人工审核效率低、标准不统一、虚假报工难识别的核心短板，通过支撑智能审核模型实现整体安装状态判定准确率≥90%、关键接口识别准确率≥92%、误判率≤3%的目标，同时依托可追溯的评测流程与“评测-反馈-优化”闭环机制，既提升审核效率与规范性，又筑牢服务质量与品牌声誉的保障防线。

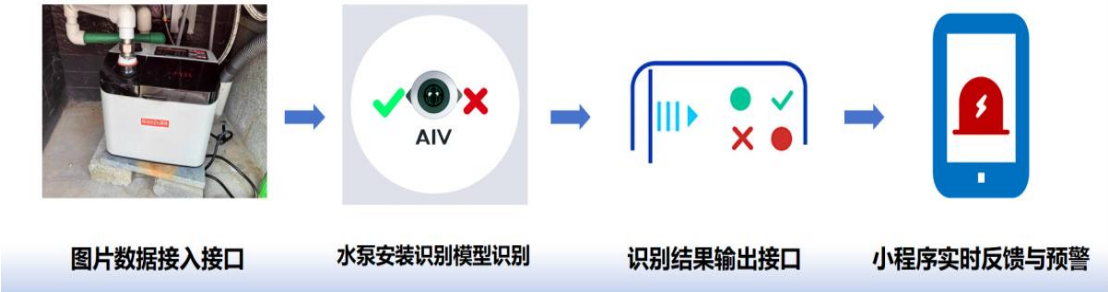

图 3-5 水泵安装高质量数据集应用整体流程

四. 高质量数据集运营体系

高质量数据集建设仅为起点，而运营则是释放其核心价值的关键。高质量数据集运营需以建立数据集管理体系为底座，以构建内外双循环价值体系为引擎，以打造协同发展生态体系为保障。其中数据集管理体系是基础支撑层，通过建立标准化的管理机制，实现数据资源的规范化管控与持续优化；内外双循环价值体系是价值实现层，在数据集管理体系的基础上，对内推动数据集转化为战略资产，对外通过产品化流通促进跨组织的合规有序价值交换；协同发展生态体系是生态扩展层，推动数据要素在更广范围内的高效配置与创新协同，最终实现从内部管理到内外循环、再到生态共荣的体系化演进。

图 4-1 高质量数据集运营体系

（一）建立数据集管理体系

高质量数据集运营的首要任务是围绕制度构建、目录管理、发布管理、质量监控及更新管理等关键环节，建立完善的数据集管理体系，全面实现数据集的可视、可控、可信、可用、可追溯。

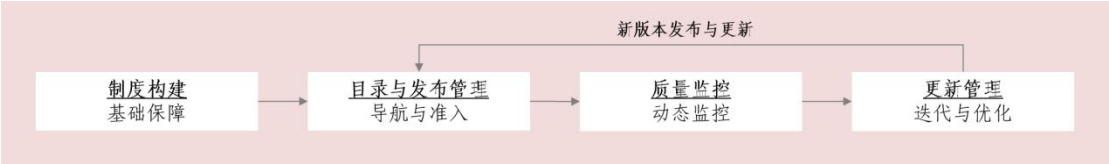

图 4-2 数据集管理体系

高质量数据集管理制度的构建是推动数据从资源向资产转化的基础保障，为后续价值评估、资产入表与流通交易等奠定制度基础。一是建立统一的数据集标准规范，明确数据定义、业务口径和质量要求，为后续价值评估提供标准化基础。二是构建协同开发机制，制定从需求分析、联合开发到验收上线的全流程协作规范，确保数据集建设与业务场景深度耦合，提升数据资产的实际适用性。三是制定数据共享和流通交易规则体系，规范数据申请、审批、使用及流通等全链条环节，推动数据集实现高效流通与价值转化。四是建立数据考核与激励制度，将数据质量及数据应用成效等关键指标纳入相关部门与个人的绩效考核体系，为后续价值评估与资产转化提供持续动力。如国家数据局公布的高质量数据集典型案例显示，中国电信在网络大模型高质量数据集运营中提出集团+省多级协同机制，通过职责分工、质量追溯、

周期联动等创新管理，保障数据集动态更新。

高质量数据集目录和发布管理旨在通过构建科学规范的分类目录与标准化发布机制，建立面向业务场景的数据服务门户，实现数据集的高效流通。一是建立多维分类体系，基于数据领域、业务主题、来源类型和安全等级等特征构建分类框架，确保分类系统既符合业务认知需求又满足管理规范要求；二是打造智能化服务门户，提供智能检索、分层浏览等功能，提升数据集的可发现性和可理解性。三是建立标准化审核机制，对数据来源合规性、数据质量完备性及数据规范性进行审查，从源头保障数据集的合法性与发布质量；四是制定完备的技术文档体系，涵盖数据基础描述、数据字典要素、标注规范、版本信息及使用指南等，形成可追溯的数据集档案。

高质量数据集质量监控是对数据集质量状态实施动态管控的核心手段，依托自动化工具对完整性、准确性、时效性等关键质量指标进行实时追踪与波动预警，建立从问题发现、定位分析到处置验证的闭环管理机制，实现对质量异常与潜在风险的快速响应，保障数据在生产与应用过程中的稳定性与可靠性。

高质量数据集更新管理核心在于建立维护与增强机制，确保数据集持续符合业务发展需求。一是建立数据集更新机制，根据业务周期特点制定差异化更新策略，通过规范化流

程实现数据内容的定期补充与动态扩充，确保持续时效性与完整性。二是构建反馈驱动优化机制，以数据集质量、模型训练结果和实际应用效果作为核心反馈，反向优化数据采集策略、处理规则及质量管控规则，形成“数据-模型-数据”闭环飞轮，持续提升数据集多样性与模型泛化能力。三是实施版本控制管理，采用专业化工具管理数据集版本演进，并建立回溯机制，支持历史数据追溯与问题排查，确保数据应用过程稳定可靠。国家数据局公布的高质量数据集典型案例显示，中国南方电网有限责任公司在电网调度负荷预测高质量数据集的运营中，建立了数据飞轮运营机制，通过将实时产生的电网负荷、新能源出力与精细化气象等数据自动加工成新的数据集，持续迭代增量训练 AI 模型，实现数据集与 AI 模型的协同进化与持续增值；中国科学院海洋研究所、中国科学院大气物理研究所、中国科学院南海海洋研究所为推动打造全球海洋环境变化关键参数高质量数据集持续运营，建立了数据集动态更新机制，确保数据的时效性和连续性。

（二）构建内外双循环价值体系

高质量数据集价值实现的核心路径在于构建协同联动的内外双循环体系，对内通过资产化运营赋能企业内部发展，对外通过产品化流通拓展数据价值空间，形成内外联动、相互促进的数据要素价值实现机制。

1.对内资产化运营

高质量数据集对内资产化运营建立在数据集管理体系的基础之上，围绕价值评估、资产入表与融资、内部场景应用三个维度系统展开。其中数据集管理体系所确立的标准规范、质量保障与更新机制，为资产化运营提供可信、可溯、可持续的数据基础。

高质量数据集价值评估为高质量数据集运营提供价值指引。其通常围绕成本、场景、市场、经济、安全五个维度构建完整的价值评估体系。其中成本维度关注数据集全生命周期的资源投入，核心指标包括数据获取、处理、存储与维护成本；场景维度关注数据集与业务场景的适配度、场景应用频率、深度、广度，确保数据集能够有效支撑业务需求；市场维度从供需关系、稀缺性、竞争力等角度评估数据集的市场地位，为商业化决策提供参考；经济维度衡量数据集直接或间接创造的经济效益，包括销售收入、成本节约、风险规避等方面价值贡献；安全维度评估数据全生命周期的安全防护能力，核心指标涵盖数据加密、访问控制、合规性等，为数据价值稳定实现提供安全保障。

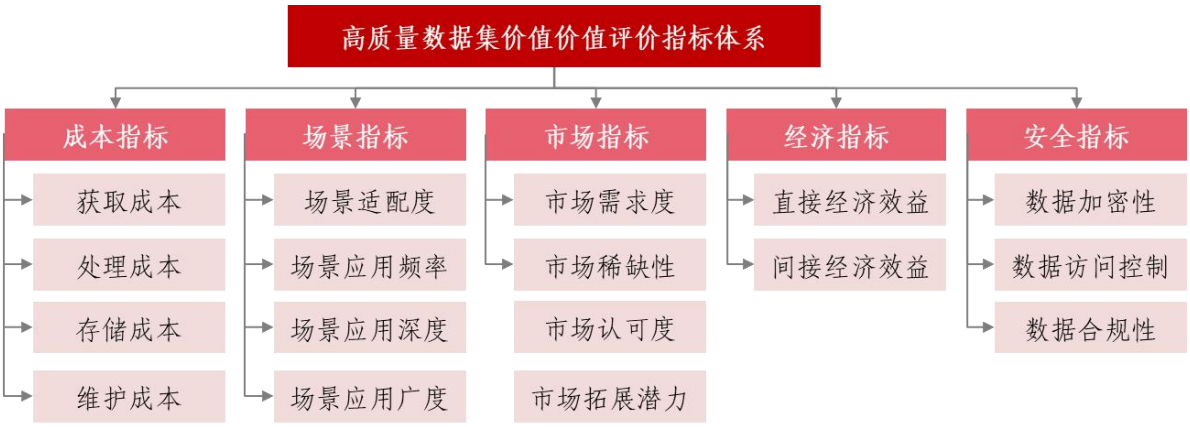

图 4-3 高质量数据集价值评估指标体系

高质量数据集入表及授信融资等的实施，将有力推动数据集从静态资源向动态战略资本转化。数据资产入表不仅为数据集赋予规范的财务属性与显性的资本价值，更通过优化企业资产负债表结构，为后续定价、交易及融资等提供权威的价值基准。具体推进方面，一是建立数据资产入表机制，依据会计准则及相关法规，将符合要求的高质量数据集依法确认为无形资产或存货，完成财务报表层面的确认与计量；二是构建多层次的数据资产融资体系，依托经确认的数据资产提升企业信用水平，创新拓展包括质押融资、证券化在内的多元化融资渠道，增强企业资本运作能力。

高质量数据集内部应用主要围绕流程优化、决策支撑、模式创新方面展开，通过数据赋能，推动业务持续升级与价值释放。流程优化方面，财务、人力资源、法务等高质量数据集，可高效支撑文档自动化处理、流程智能监控等关键场景，实现业务流程的整体优化；决策支撑方面，用户行为、交易日志、供应链数据等高质量数据集，可支撑开展精准营

销、需求预测、反欺诈建模等分析，增强业务洞察与风险识别能力；模式创新方面，用户交互、行业知识库等高质量数据集，可支撑洞察市场动态与用户需求，驱动产品迭代升级与业务模式创新，进而拓展增长路径，构建可持续的竞争新优势。

2.对外产品化流通

高质量数据集对外产品化流通是释放数据集价值的重要路径，需通过构建多元化的服务模式与健全多层次的流通方式，持续推动高质量数据集的市场化配置与价值转化。

当前，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：一是标准化数据集服务，将高质量数据集打包为标准化产品，面向行业客户、科研机构、平台企业开放供给；二是定制化数据集服务，基于客户特定业务场景和需求，提供包括需求分析、数据采集、清洗标注、质量验证在内的端到端定制服务，确保数据集与客户业务场景的深度契合；三是一体化解决方案服务，将数据集与算法模型、软件工具及算力资源深度融合，形成一体化解决方案，通过降低技术门槛实现数据价值的嵌入式转化。多种模式相互补充，共同推动高质量数据集在多层次商业化场景中的价值释放。

高质量数据集价值的高效发挥依托于共享、开源开放及交易三类流通方式的协同推进。从实践来看，不同服务模式

倾向于适配差异化的流通路径，其中标准化数据集服务多依托开源开放与共享模式，降低基础数据的获取门槛，促进数据资源的广泛流通与创新应用；而定制化数据集与一体化解决方案服务，则更侧重于通过规范、可信的交易机制开展，以确保服务深度、合规安全与商业可持续性。

目前三类流通方式各具特点，共同构建起多层次的流通生态，三类流通方式的特点如下:

高质量数据集共享是政府机构、行业联盟及企业内部基于共同认可的合作框架，在特定授权范围内实现数据集按需共享的行为。在运行体系上，需建立涵盖权限管理、流程规范与激励措施的完整制度，确保在严格遵循数据安全与合规要求的前提下，促进数据要素在授权范围内有序流动并实现价值释放。国家数据局公布的高质量数据集典型案例显示，国家石油天然气管网集团有限公司油气调控中心构建的天然气管网运行高质量数据集，通过建立行业数据生态共享机制，不仅支撑内部智能调控，还向国家能源局共享 24 小时实时数据，为城市燃气公司等上下游赋能，推动数据价值从“内部降本增效”延伸至“行业协同创新”，成为能源数字化转型“管网方案”的核心。

高质量数据集开源开放主要采取免费开放策略向社会公众提供数据资源。其价值实现不依赖于直接交易，而是通过生态共建和技术创新等间接方式实现。具体实施包括两种

类型：一是社会数据集开源，依托开源社区平台，由研究人员、科技公司、非营利组织及个人开发者共同推动；二是公共数据集开放，即通过政府建设的公共数据开放平台、城市可信数据空间等基础设施，设立高质量数据集专区，免费向社会公众开放与民生密切相关的非涉密数据，涵盖公共治理、环境资源、交通出行等领域。国家数据局公布的高质量数据集典型案例显示，智元创新（上海）科技有限公司构建的具身智能领域百万真机高质量数据集已实现开源，在魔塔社区、OpenDataLab、Huggingface、GitHub 等国内外平台上线，成功吸引全球研究者的广泛参与，累计下载量达 29K。

高质量数据集交易是指在数据权属清晰界定的前提下，由供方向需方有偿转让数据集使用权或所有权的行为。相较于开放和共享，高质量数据集交易更能有效激发市场参与主体积极性，正成为数据集流通的主要实现形式。目前高质量数据集交易主要在经政府批准设立的数据交易场所开展，数据交易场所在其中承担撮合交易、合规审核与清结算等职能，通过执行统一的交易规则与监管标准，推动形成规范化、标准化的流通交易渠道。与此同时，可信数据空间等新型基础设施以及点对点协商交易等灵活方式，共同为数据集产品提供灵活高效的交易路径，有效补充场内交易的服务范围。

表 4-1 高质量数据集流通方式

	数据集共享	数据集开源开放	数据集交易

收费模式	免费/收费	免费	收费
流通渠道	政府机构、行业联盟及企业内部间	开源社区平台、政府公共数据开放平台、城市可信数据空间等	政府批准设立的数据交易场所、各类可信数据空间、点对点协商交易

浙江电信已在杭州数据交易所完成“网络诈骗网站高置信度预测数据集”产权登记并上架产品进行交易。该数据集是浙江电信为互联网反诈专项治理构建的高质量、高置信度多模态数据资源，其核心在于综合利用前沿技术流程与人工审核保障数据可靠性。数据来源于对浙江省内用户上网流量中可疑网站访问行为的筛选，通过拟人态爬虫同步获取 URL 文本与网页快照，并创新性地引入多模态大模型（融合视觉与文本联合编码器）对图片数据进行深度预处理与特征提取，再经涉诈黑库匹配及专业团队抽样复核，最终形成约 6.5 万个涵盖图片与 JSON 格式的结构化数据。

图 4-4 “网络诈骗网站高置信度预测数据集”产权登记证书

（三）打造协同发展生态体系

构建协同发展的数据集生态体系是推动数据资产可持续运营、实现价值最大化的重要支撑。为系统推进生态建设，需围绕以下四个关键维度统筹布局：一是打造生态合作联盟，通过组建联合实验室、产业联盟等形式，强化与科研院所、

创新机构、跨界企业等的深度协同，推动数据集的协同开发与资源共享。二是建立生态协同规则，制定涵盖生态准入、收益分配、权益保护等规则体系，明确各方权责，打通数据集供给方、需求方与服务方之间的协同链路，营造公平、透明、可信的合作环境。三是提供开放生态服务，开放数据集建设工具、测评工具及数据集等，吸引第三方厂商入驻，推动工具共享与数据资源纳管，为生态伙伴提供业界领先的技术支持与服务。四是打造流通基础设施，加快建设以可信数据空间为代表的新型基础设施，联合产业链各方构建安全可控的数据合作网络，在保障数据安全与隐私的前提下，促进数据要素有序流通、高效配置与价值共享，赋能产业协同发展。通过构建“组织联盟、制度规则、服务开放、基础支撑”四位一体的协同网络，不仅能够持续释放数据价值，更能巩固自身在产业生态中的核心地位，实现从数据运营向生态赋能的战略升级。

五. 高质量数据集实施建议

（一） 聚焦关键环节，打造系统建设能力

为破解高质量数据集规模化生产、高效流通与可信应用的关键瓶颈，企业需着力强化技术自主创新能力，构建先进完备的数据生产基础设施，打造开放协同的服务平台体系，

全面提升数据要素供给质量与支撑水平。

一是构建资源地图。锚定关键智能场景，全面梳理企业内外数据资源，形成涵盖数据源、格式、质量、权属的可视化资源目录，支撑数据集供需匹配与场景化应用。

二是加强技术攻关。聚焦智能标注、多模态融合、合成数据生成等关键环节，组织产学研联合研发，突破数据处理在自动化、精准化与安全可控方面的技术瓶颈，提升数据集生产的效能与质量。

三是部署生产设施。建设专业化、规模化的数据集生产平台，引入自动化与智能化工具链，推动高质量数据集在数据采集、清洗、标注、增强等环节的自动化闭环生产与持续质量演进，实现数据生产流程的标准化、智能化与可追溯。

（二） 建立运营生态，驱动数据价值演进

为推动高质量数据集实现可持续运营与价值最大化释放，行业需积极构建多方参与、机制灵活、生态繁荣的发展环境，通过共建共享凝聚合力，畅通多元转化循环，最终形成数据要素价值充分涌流的良性生态。

构建动态运营管理体系。为实现高质量数据集的持续保值增值，需构建覆盖其全生命周期的运营管理体系。具体而言，通过建立用户需求响应机制、成本精细化管理体系、质量与安全维护体系，并结合制定分级共享策略、推动标准化

流通、建立共建与价值分配机制，实现数据集从静态管理向动态进化转型，使其在使用中持续优化并拓展应用边界。

形成数据价值应用闭环。推动建设集资源目录、质量评测、合规流通等功能于一体的数据集管理与服务平台，提供可视化工具与完整数据字典，降低使用门槛。内部数据集可通过数据门户供团队查询，外部数据集可发布至开源社区，并配套版本控制与自动化监控规则，实现数据集的规范发布、持续更新与质量可控，形成数据供给与价值应用的闭环。

打造开放协同产业生态。通过参与产业共同体、数据标注产业联盟，以及探索委托授权、知识产权保护等制度创新，搭建合作平台与交流机制。鼓励以揭榜挂帅、技术竞赛等形式征集解决方案，并通过数据交易所、可信数据空间等基础设施，在安全合规前提下，推动数据资源跨域融合与高效利用，最终形成多方参与、共建共享、价值共创的可持续发展生态。

（三） 构建制度保障，激发产业整体效能

为破解当前高质量数据集市场存在的供给不足、流通不畅、合规风险等关键瓶颈，政府需超越单一的数据开放者定位，转向承担系统性市场培育者的角色。其核心目标是围绕高质量数据集的供给能力、流通能力与安全合规三大支柱，协同施策，通过完善产业链条来系统性激发市场活力，推动

数据要素市场的健康、高效发展。

一是在供给端，需主动开展行业供需调查与撮合，并重点加强公共数据的系统性梳理与高质量供给，同时通过提供产权登记、纠纷协调等产业服务，降低交易成本。**二是在流通端**，应双管齐下：一方面加强隐私计算等“数据可用不可见”的流通基础设施建设，另一方面通过试点评估、定价、交易等市场机制，并分类扶持服务型、技术型、应用型“数商”，繁荣市场生态。**三是在合规与安全端**，必须筑牢底线，严格落实相关法律法规，建立覆盖供给、流通、应用全过程的动态防护与监管能力，并建立内容审查标准，防范数据伦理与意识形态风险。

最终，通过公共数据供给和基础服务降低市场启动门槛，通过规则与标准建设规范市场秩序、稳定市场预期，通过扶持关键市场主体和基础设施建设弥补市场失灵。以此构建一个供给充分、流通高效、安全可信的高质量数据集市场环境，使数据要素能够合规、顺畅地流向最具创新活力的应用场景，为数字经济发展提供坚实的数据基础。

[1]高质量数据集建设指引.国家数据局等，2025.

[2]人工智能高质量数据集建设指南.中国信息通信研究院，2025.

[3]高质量数据集建设实践指南（1.0）.大数据技术标准推进委员会，2025.

[4]工业高质量数据集研究报告.中国工业互联网研究院，2025.

[5]姜春宇,白玉真,刘渊,王超伦.构建企业级人工智能高质量数据集：方法与路径[J].大数据,2025,11(06):47-56.

[6]吴世忠.大模型时代高质量数据集建设的进展、挑战与治理路径[J].保密工作,2025,(10):4-8.

[7]胡晓女,李涛，李姗姗.人工智能大语言模型数据集现状和充实对策研究[J].大数据,2025,11(06):57-71.

[8]李荪，樊威，曹峰　燕江依　从“经验驱动”向“标准驱动”推动高质量数据集建设[J]　　业和信息化 2025 (08):26-31.

[9]丁浩,张畅,顾乐,等　面向　全生命周期的高质量数据集评测体系研究[J].数字化转型　025,2(08):87-97.

[10]孙健.高质量数据