CAICT 中国信通院

中国人工智能产业发展联盟
Artificial Intelligence Industry Alliance

集智 专题报告

人工智能高质量数据集建设指南

中国信息通信研究院人工智能研究所
清华大学计算社会科学与国家治理实验室
中国人工智能产业发展联盟数据委员会

2025年7月

版权声明

前　言

党中央和国家高度重视推动数据赋能人工智能高质量发展。2025年4月25日，中央政治局就加强人工智能发展和监管进行第二十次集体学习，习近平总书记指出，要“深化数据资源开发利用和开放共享”，要“全面推进人工智能科技创新、产业发展和赋能应用”。近年来，国家相关部委和地方政府围绕高质量数据集建设和运营、数据标注产业发展等出台系列政策，并通过投资奖补、标准制定和验证、样板案例建设等举措积极推进高质量数据集政策落地。党中央和国家的政策部署为业界推进高质量数据集建设提供了方向指引和根本遵循。

随着大模型技术的迅猛发展，数据集作为人工智能核心三要素之一，在算法趋同、算力普惠的竞争环境中正在构建难以复制的差异化壁垒。人工智能发展正在进入“数据驱动”新阶段，高质量数据集的建设不仅是提升AI模型性能的关键，也是推动“人工智能+”行动落地的重要保障。然而现阶段，大量机构在高质量数据集建设中面临目标定位模糊化、实施路径碎片化与技术底座薄弱化三重挑战，不知道需要什么数据集、如何建设数据集、怎样评估数据集质量，制约了人工智能应用落地。《人工智能高质量数据集建设指南》正是在此背景下启动起草，旨在为业界建设高质量数据集提供有实操价值的指导和参考。

指南从政策、技术、产业层面介绍了当前高质量数据集建设的背景，梳理了高质量数据集的定义、特征、分类、建设主体以及“三大

建设难点”，提出了人工智能数据工程的“五大核心要素”和企业建设高质量数据集“三步走”战略，分析了高质量数据集建设的核心技术，展示了科学、通信、交通、铁塔、医疗、文化等领域高质量数据集建设实践，最后从工程能力、技术创新、质量评估、版权合规、基础制度创新等层面对未来高质量数据集建设的趋势进行了展望，并提出了对政府部门和企业机构的建议，为业界推进高质量数据集建设提供有力支撑。

目 录

图 目 录

表 目 录

一、高质量数据集成为人工智能应用升级的核心要素

数据集是指一组相关数据的集合，是用于分析、建模、训练算法的关键要素。大模型技术的飞速演进，对数据集规模、多样性、质量等提出极高要求，业界迫切需要服务人工智能大模型发展的高质量数据集。高质量数据集是指用于训练、验证和优化人工智能大模型而收集、整理、标注形成的覆盖行业核心专业知识和生产经营活动信息的数据资源集合。

（一）政策层面：“人工智能+数据要素”政策协同布局

近年来，国家和地方政府纷纷出台人工智能和数据要素相关政策，推动高质量数据集的建设、流通和开发应用。

1.国家部委持续完善高质量数据集建设顶层设计

一是加快出台人工智能和数据要素政策。近年来，国家层面政策普遍强调数据集建设是人工智能发展的核心内容，人工智能是发挥数据要素乘数效应的重要手段。2023 年 12 月，国家数据局等 17 个部门联合发布的《“数据要素 ×”三年行动计划（2024—2026 年）》强调打造高质量人工智能大模型训练数据集。2024 年 6 月，工业和信息化部等 4 部门联合发布的《国家人工智能产业综合标准化体系建设指南（2024 版）》提出规范数据采集、数据标注、数据治理、数据质量等标准。此外，国家还高度重视人工智能数据安全合规。2023 年 8 月，国家互联网信息办公室发布《生成式人工智能服务管理暂行办法》，要求训练数据在知识产权、个人信息、道德伦理等方面安全合规，提出采取有效措施增强训练数据的真实性、准确性、客观性、

多样性。

二是推进行业领域高质量数据集建设。2025 年 2 月，国家数据局组织 27 个部委召开了高质量数据集建设工作启动会，国家层面行业高质量数据集统筹协调机制正式建立。各部委依据自身职责，聚焦垂直行业领域持续推进高质量数据集建设。例如，国务院国资委加快推动央企布局建设交通物流、绿色低碳、金融服务等 10 个领域数据集，并已发布首批 30 项央企高质量数据集。中国气象局印发《中国气候数据集建设工作方案（2023—2025 年）》，提出要研发关键气候变量均一化数据集、风云卫星气候数据集、天气雷达气候数据集等。

三是推动数据标注产业升级带动高质量数据集建设。数据标注作为高质量数据集建设的核心环节，决定了数据集的生产能力。国家持续推进数据标注市场规模做大做强和产业升级。2024 年 5 月，国家数据局布局成都等 7 个城市建设数据标注基地，根据国家数据局发布的数据，截至今年 3 月底，7 个城市已形成医疗、工业、教育等行业高质量数据集 335 个。2025 年 1 月，国家发展改革委、国家数据局等部门发布《关于促进数据标注产业高质量发展的实施意见》，从深化需求牵引、增强创新驱动、培育繁荣生态、优化支撑体系等方面提出了一系列产业支持政策。

2.地方政府多措并举推进高质量数据集建设落地

一是明确高质量数据集建设规划。上海、湖北、江苏、浙江等多个地区明确了建设高质量数据集的规模、周期及激励机制。2023 年 7 月，上海市发布《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）》，提出到 2025 年形成 1000 个高质量数据集。2025 年 5 月，无锡市发布《无锡市促进数据产业高质量发展实施方案（2025—2027 年）》，提出到 2027 年建成高质量数据集 300 个。

二是打造高质量数据集试点示范案例。地方政府以遴选试点示范和典型案例的形式，带动行业高质量数据集建设。例如，湖北省数据局评审并发布首批 10 个高质量数据集，涵盖科技创新、交通运输、文化旅游等多个领域。苏州市评审并发布首批 30 个工业制造、交通运输、金融服务等领域高质量数据集。

三是通过奖补手段推进高质量数据集建设。目前北京市、山东省，以及武汉市、南京市、杭州市、呼和浩特市等 11 个地区制定出台了人工智能数据集的奖补政策。从政策发布时间看，主要集中在 2025 年 2 月-4 月，凸显了高质量数据集补贴政策实施的迫切性。从奖补主体看，为解决当前数据集规模不足的主要矛盾，多数地区以奖补数据集建设单位为主，但也有武汉市、无锡市、深圳市同时奖励数据集使用单位等。从奖补金额看，各地区额奖补金额分为固定金额补贴和固定比例补贴（按照投入成本或购买成本比例补贴）两种。

（二）技术层面：人工智能技术演进重构数据工程范式

一是人工智能技术演进对数据集建设规模和质量提出了更高要求。以大模型为代表的人工智能技术展现出了类人智能的“涌现”能力，呈现规模可扩展、多任务适应及能力可塑三大特征，对数据集的规模、质量等提出更高要求。**在数据集规模方面**，大模型的规模可扩

展是指提高数据质量、扩大数据集规模或提升算力规模水平，都能够显著增强模型的复杂性和处理能力，推动当前预训练数据集规模提高到数十万亿 Token 的水平。**在数据集多样性方面**，大模型的多任务适应是指大模型支持多任务多模态能力持续增强，要求当前数据集建设过程中融合文本、图像、语音、视频等多模态数据，形成跨模态关联。大模型的能力可塑是指通用大模型在训练阶段通过结合增量预训练、有监督微调等方法，在推理阶段引入检索增强生成等技术，能够提升大模型在行业领域的应用能力，要求建设丰富多样的行业领域数据集。**在数据集质量方面**，要求保证数据真实性与时效性，并通过数据过滤、去重、平衡类别分布等手段提升数据纯净度，同时注重隐私保护与合规性，避免敏感信息泄露。

表 1 人工智能技术发展各阶段对数据集的需求

人工智能技术发展阶段	人工智能技术发展特点	人工智能技术对数据集的需求	典型数据集
1.浅层学习阶段（2012 年之前）	以传统机器学习算法为主，包括支持向量机、决策树等。模型结构相对简单，数据特征提取依赖人工设计，需领域专家手动构建有效数据特征。	1.规模较小，通常数万至数十万样本即可满足训练需求。 2.注重数据清洗，需人工剔除噪声样本、处理缺失值。 3.特征工程至关重要，数据需具备清晰可解释的结构化特征，便于人工设计与筛选。	图像领域：MNIST 手写数字数据集（约 7 万张标注图像，用于字符识别）、Caltech 101/256 物体图像数据集（数千张人工标注图像，用于图像分类）。 自然语言处理领域：Reuters-21578 文本分类数据集（约 2 万篇新闻文本，人工标注主题标签）等。
2.深度学习阶段	神经网络技术崛起，人工智能模型转	1.规模显著扩大，需数百万至数千万样本支撑深度模型训练。	图像领域：ImageNet 大规模视觉识别数据集（约 1400 万张标注图

人工智能技术发展阶段	人工智能技术发展特点	人工智能技术对数据集的需求	典型数据集
	向自动提取数据特征，如CNN、RNN等算法架构逐渐普及，数据特征学习能力显著提升，但仍依赖大量标注数据驱动。	2.标注精度要求更高，数据标注需细化到像素级（如图像分割）或字符级（如自然语言处理）。 3.数据多样性增强，需覆盖多场景、多模态数据，以提升模型泛化能力，如图片数据集需包含不同光照、角度、背景的照片。	像，含2万余类别，推动CNN技术突破）、MS COCO图像数据集（超33万张图像，标注物体边界框及语义掩码，支持目标检测与分割）； 自然语言处理领域：WikiText语言建模数据集（基于维基百科的数千万词级文本，用于语言模型训练）、GLUE基准数据集（多任务自然语言理解数据集，含数万至数十万条标注文本）； 语音领域：LibriSpeech语音识别数据集（约10万小时标注语音，结合文本转录，推动端到端语音模型发展）。
3.大模型阶段	Transformer架构为核心的人工智能大模型成为主流，模型参数规模达数十亿至数万亿美元，具备更强的上下文理解与生成能力	1.规模呈指数级增长，需数十万亿Token（如文本）或样本（如图像）构建训练语料。 2.数据模态高度融合，需整合文本、图像、语音、视频等多模态数据，形成跨模态关联。 3.数据质量要求精细化，既要保证数据真实性与时效性（如实时更新的网页内容），又需通过数据过滤、去重、平衡类别分布等手段提升数据纯净度，同时注重隐私保护与合规性，避免敏感信息泄露。	自然语言处理领域：Common Crawl超大规模网页数据集（累计超万亿Token）、The Pile混合文本数据集（含3000亿Token，整合学术论文、代码、图书等多源数据）； 多模态领域：LAION-5B数据集（含58亿图文对，支持图文跨模态模型训练）、YouTube-8M视频数据集（约800万视频样本，结合文本标签与视觉特征，用于视频理解大模型）； 代码领域：GitHub公共代码数据集（数十亿行代码片段）；

人工智能技术发展阶段	人工智能技术发展特点	人工智能技术对数据集的需求	典型数据集
			具身智能领域：Open X-Embodiment 数据集（谷歌建设的 22 种不同本体机器人的操作数据） 复杂推理思维链领域：NuminaMath-CoT 数据集（数学 AI 公司 Project Numina 发布，包含包含有 86 万个数学问题，每个问题的解答都采用了思维链的方式进行格式化。这些数据集的来源涵盖了中国高中数学练习题到美国及国际数学奥林匹克竞赛题目。）

二是高质量数据集建设工程范式持续创新。近期 DeepSeek 系列模型在数据训练中构建了数据工程创新体系，持续推进人工智能数据工程范式变革。首先通过自动化推理和数据生成技术，显著降低对传统人工标注的依赖，实现数据标注方式的智能化升级。其次采用数据蒸馏技术提炼低质数据中的有效信息，结合自动化筛选与人类专家反馈机制，形成“机器预处理+人工校准”的双层质检流程，使标注质量与效率得到双重保障。最后创新性地运用强化学习框架，聚焦推理能力提升，构建了包含 60 万条推理型样本与 20 万条非推理型样本的训练集，通过 3:1 的优质数据配比优化模型架构，同时大幅削减监督微调数据占比。

三是人工智能发展迫切需要多模态、具身智能、思维链、长视频

等四类数据集。其中，**多模态数据集**使模型更加全面和精准地理解和处理任务，更好地应对复杂应用场景需求。如华盛顿大学等发布MINT-1T，具体包含1万亿文本Token和30亿张图像数据。**具身智能数据集**可以增强机器人在多样化环境和任务中的适应性和决策智能，实现更高级别的自动化和智能化，主要包括真实世界数据和仿真合成数据两大类，涵盖三维场景信息、物体物理属性以及环境交互细节，具体包含2D、3D、文本、触觉、声音等多种模态数据。**推理思维链数据集**能够促进模型推理能力大幅提升。**长视频数据集**能够辅助AI用于智能监控和预警系统的构建、辅助医生进行疾病诊断和治疗方案的制定等场景。

（三）产业层面：数据成为人工智能行业应用的护城河

在算法趋同、算力普惠的背景下，高质量、高价值密度的数据集将构建起企业差异化竞争力，成为企业人工智能业务发展的护城河。

一是人工智能技术和行业数据集融合是“人工智能+”的核心。人工智能行业的蓬勃发展，正加速向各领域渗透，任何创新应用的落地，都离不开高质量数据集的强力支撑。从智能语音交互到复杂的图像识别，从智能交通系统到精准医疗诊断，海量且优质的数据是算法训练的基石。只有通过不断输入高质量数据集，人工智能模型才能精准理解各类复杂场景，提升预测与决策能力，从而实现从理论模型到实际应用的跨越，让技术真正服务于生产生活。

二是高质量数据集将成为企业构筑人工智能应用壁垒的关键。在细分行业领域，高质量数据集的积累正成为人工智能企业的核心竞争

力。率先在垂直行业建成高质量数据集的企业将有能力研发出性能优异的行业大模型，进而在行业大模型应用、研发迭代过程中采集到更多高质量数据资源，形成“数据飞轮”的马太效应，最终以高质量数据集构筑起企业人工智能应用的壁垒。以医疗影像和工业制造领域为例，医疗机构长期积累的病理影像、诊断记录等数据，工业企业持续采集的生产流程、设备运行数据，是企业的核心数据资产。这些数据不仅具有高度专业性，还因行业特性难以被轻易复制或获取。基于此训练的人工智能行业模型，在性能上形成显著的代际优势，将能更好地满足行业特定需求，并进一步丰富企业数据资源优势，进而构建起难以逾越的竞争壁垒。

二、高质量数据集建设的现状和难点

（一）高质量数据集的“三高”特征

高质量数据集覆盖制造、金融、医疗、交通、公共安全、自然资源、地理信息、人力资源、社会治理、科学研究等重点行业的公域数据和私域数据，具有高价值应用、高知识密度、高技术含量的“三高”特征。

一是高价值应用。高质量数据集的建设重点瞄准产业刚需的核心场景。以医疗行业为例，美国国立卫生研究院发布的胸部 X 光数据集 ChestX-ray14，汇集了 11 万张标注精准的医学影像，覆盖 14 种常见肺部疾病。该数据集直接支撑了 AI 辅助诊断系统的开发，使肺炎等疾病的 AI 识别准确率与人类医生相当，诊断速度则是人类的 160 倍。在工业领域，苏州阿丘科技构建了 500 多万张 PCB（印制电路板）

的图片组成的数据集，并研发出能够识别 PCB 上百种缺陷的模板，包括铜面异物、线路漏铜等缺陷，实际应用将缺陷过检率降低了 90%。这些案例表明，高价值数据集往往锚定产业痛点，通过场景化数据服务创造倍增效益。

二是高知识密度。高质量数据集能够用于构建跨领域的知识图谱。优质数据集的精髓在于其蕴含的复合型知识体系。例如，自动驾驶数据集 Waymo Open Dataset 包含 2000 多个道路实景视频，整合了激光雷达点云、交通标志语义解析、驾驶员行为模式等多模态数据，涉及计算机视觉、交通工程等学科知识。在生物医药领域，AlphaFold 使用了 20 多万个蛋白质结构构成的数据集，集成了 X 射线晶体学、冷冻电镜成像等多学科实验数据，构建出覆盖 98.5%人类蛋白质的三维结构图谱。这种跨学科的知识融合，使数据集成为连接理论突破与应用创新的桥梁。

三是高技术含量。数据集质量的提升已进入技术驱动新阶段，正在重塑数据集生产的全流程。**合成数据技术**正引发变革，英伟达运用生成对抗网络创建的自动驾驶合成数据集，可模拟暴雨、暴雪等极端天气场景，缓解了恶劣驾驶环境数据不足的问题。微软 Phi-4 小模型（140 亿参数），在数学等测试中表现极佳，该模型合成数据占比 40%、互联网数据 30%、代码数据 20%、其他渠道数据 10%。马斯克称 xAI 公司的大模型 Grok3 大量使用合成数据。在**人机协同标注**领域，百度开发的智能标注平台内嵌 100 多种智能算法，助力标注效率提升 50%。

（二）高质量数据集分类维度

1.基于数据应用维度分类

从数据应用维度分类，高质量数据集能够分为通识类高质量数据集、行业通用类高质量数据集、行业专用类高质量数据集。**一是通识类高质量数据集**。通识类高质量数据集是指由政府机构、科研机构、开源社区或大型互联网企业等公开数据构建的数据集，具有广泛性和通用性，覆盖多个领域，如自然语言处理、计算机视觉、语音识别等，能够为企业提供丰富的训练资源和基准测试环境，有助于行业大模型快速验证算法、提升模型的基础能力。**二是行业通用类高质量数据集**。行业通用类高质数据集，是指针对某一特定行业或领域知识的具有事实性数据集，具有高度的专业性和针对性。这类数据集通常包含某一特定行业特有的知识、术语、场景和业务流程等信息，对于训练出适用于行业应用的大模型至关重要，能够覆盖行业领域专业知识，提高模型在行业通识领域的泛化能力。**三是行业专用类高质量数据集**。行业专用数据集，是指根据行业企业自身业务场景和需求收集的数据集。这类数据集通常包含行业企业内部业务流程、用户行为、产品信息等关键信息，具有针对性和定制化的特点，能够为行业企业提供高度个性化的训练数据资源，构建专属大模型。通过行业企业场景化数据集的训练，可以定制化地优化大模型算法和参数设置，深度挖掘内部数据价值，实现模型的定制化优化与业务高度适配，使其更好地服务于业务需求和发展战略，带来更加精准和有效的业务洞察和决策支持。

2.基于模型训练阶段分类

从模型训练阶段分类，高质量数据集包括预训练数据集、微调数据集、价值观对齐数据集、行业领域数据集等。**一是**模型预训练阶段，预训练数据集规模庞大且内容广泛，涵盖文本、图像、音频、视频等多种数据类型，旨在让模型学习通用的语言规则、视觉特征或语义理解等基础能力。**二是**模型监督微调阶段，微调数据集针对特定任务或领域，通过标注准确的输入输出对，引导模型在预训练基础上，进一步优化对特定任务的处理能力。**三是**基于人类反馈的强化学习阶段，价值观对齐数据集通过收集人类对模型输出的评价、偏好等反馈信息，帮助模型学习符合人类价值观和伦理规范的输出方式，避免产生有害或不当内容。**四是**模型行业领域适配阶段，行业领域数据集聚焦特定行业的专业知识和业务场景，使模型能够在专业领域发挥作用，满足行业特定需求。

表 2 不同训练阶段数据集的质量特征

	预训练阶段	**监督微调阶段**	**基于人类反馈的强化学习阶段**	**行业领域适配阶段**
训练目标	**预训练数据集**：通过无监督学习捕捉语言的统计特性，学习词汇和句子结构等。	**监督微调数据集**：通过有监督学习，微调模型使其适应特定领域或行业任务。	**价值观对齐数据集**：使模型能够根据人类的价值观偏好和评价来调整输出。	**行业数据集**：通过检索相关信息来增强生成效果，使生成结果更准确。
多样性	覆盖多领域（如文本、图像、语音）和多语言数据，提升模型的泛化能力。	覆盖模型应用潜在应用场景，例如医疗问答任务训练中，应包含疾病诊断、药物说明、患者咨询等多类问	数据集应涵盖多样化的人类反馈类型，包括不同场景、不同用户群体的反馈，以确保模型能够全面地学	覆盖行业特有知识、术语及业务流程。例如，金融数据集需涵盖财报分析、风险评估等场景，并确保术语

	预训练阶段	监督微调阶段	基于人类反馈的强化学习阶段	行业领域适配阶段
		题。	习到人类对各种输出的价值观评价和偏好。	准确性。
标注需求	无须标注。	标注准确性要求高。	标注准确性要求高。	标注准确性要求高。
内容平衡性	避免类别或领域偏差，防止模型输出偏向高频内容。	无相关要求。	避免模型过度偏向某些特定人群的价值观。	无相关要求。
内容时效性	内容时效性要求适中。	内容时效性要求适中。	内容时效性要求适中。	内容时效性要求高。
安全与合规性	内容符合社会主义核心价值观，不涉及个人隐私等，无版权争议。	内容符合社会主义核心价值观，不涉及个人隐私等，无版权争议。	内容符合社会主义核心价值观，不涉及个人隐私等，无版权争议。	内容符合社会主义核心价值观，不涉及个人隐私等，无版权争议。

3.基于数据模态维度分类

从数据模态维度分类，高质量数据集包括文本、图片、音频、视频和其他。**一是文本数据集**。文本数据集以离散的字符序列为表征形式，是自然语言处理任务的基石。在情感分析、机器翻译等应用场景中，高质量文本数据需具备标注一致性、语义多样性和领域覆盖度。通过预训练语言模型对大规模文本语料进行特征提取，可有效提升模型对语言结构和语义理解的泛化能力。文本数据集的歧义性和文化差异，使得数据清洗与标注成为构建高质量数据集的关键环节。**二是图片数据集**。图片数据集以像素矩阵存储视觉信息，广泛应用于计算机视觉领域。在目标检测、图像生成等任务中，高质量图片数据需满足高分辨率、多场景、精确标注的要求。同时，借助数据增强技术（如

图像旋转、裁剪、噪声添加）可扩充图片数据集规模。**三是音频数据集**。音频数据集以时域波形信号承载声学信息，是语音识别、声纹识别等任务的核心资源。其数据质量受采样率、信噪比、标注粒度等因素影响。**四是视频数据集**。视频数据集融合了时空维度的图像与音频信息，适用于动作识别、视频目标跟踪、视频摘要等任务。由于视频数据的高维度特性，常采用光流（Optical Flow）、时空特征金字塔等技术提取动态特征。基于 3D 卷积神经网络（3D CNN）和时空 Transformer 的模型架构，正推动人工智能大模型在自动驾驶、安防监控等领域的深度应用。**五是其他类型数据集**。除了上述常见的数据模态，还存在多种具有独特价值和模态的数据集类型。例如，地理空间数据集包含地理位置、地形地貌、交通网络等空间信息，广泛应用于城市规划、物流配送、环境监测等领域。金融市场中的股票价格走势、电力系统的负荷变化数据等时间序列数据集则按照时间顺序记录数据，常用于预测分析、异常检测等任务。多模态融合数据集整合了文本、图像、音频等多种数据模态的信息，为跨模态学习任务提供支撑。

（三）高质量数据集建设主体

一是高质量数据集开发和治理主体。该类主体是数据集的建设方。其中，算法服务方和技术服务方负责供给数据集采集、清洗、标注、合成、质检等相关基础算法和技术工具。平台服务方提供数据集开发供需对接、任务分发等服务。交易服务方提供数据集流通交易、资产入表、价值评估等服务。人力服务提供数据标注人才员招聘管理等服

务。**二是数据资源提供和应用主体**。该类主体是数据集原始数据资源的提供方，也是数据集建成后的开发利用方。主要包括公共数据、行业数据、互联网数据资源提供和应用主体。**三是能力支持与生态发展主体**。该类主体为数据集的建设提供人才、标准、安全、生态培育等服务，支撑提升数据集建设、使用、运营能力。人才培养方涵盖了高校、人才培训组织等。生态培育方包括有关联盟和协会，推动相关企业集聚、资源互通；数据安全方提供数据分类分级、数据安全监测预警等技术支持；标准制定方推动数据集格式、分类分级、技术工具、质量评估等标准起草和应用。

图 1 人工智能高质量数据集建设相关主体

（四）高质量数据集建设难点

当前，高质量数据集建设正处于探索阶段，主要面临目标定位模糊、实施路径碎片化与技术底座薄弱三重挑战。

一是目标定位模糊。人工智能高质量数据集建设常陷入“为数据而数据”的误区，智能场景需求与数据集建设目标脱节，企业未能将数据工程目标与核心业务指标深度绑定，导致数据价值难以转化为模

型性能提升。例如，制造业质检数据若缺乏缺陷分级标准，则无法支撑分级预警系统的落地。这种目标模糊性使得数据采集与清洗缺乏针对性，难以形成“数据采集—模型训练—业务反馈—数据迭代”的闭环优化机制。

二是实施路径碎片化。从数据采集到模型训练的全链路缺乏系统性规划和设计，无法形成体系化数据集构建和维护机制，造成多源异构数据标准难统一、跨部门跨层级难协作，致使清洗、标注等数据处理成本激增。例如，工业智能应用场景复杂多样，同样又涉及设备日志、检测数据和人工巡检记录等数据分散在不同系统中，数据格式差异大，导致采集难度高、清洗与标注需反复对齐。

三是技术底座薄弱。现有数据处理技术难以应对复杂人工智能场景需求，多模态数据处理能力不足，制约模型迭代与应用规模化。同时，缺乏适配行业特性的工具链，自动化程度低，人力依赖严重，工程落地效率受阻，行业特性适配工具链的缺失加剧了这一矛盾。例如，工业时序数据分析需专用特征提取工具，医疗影像标注依赖专业CAD辅助软件，但通用化工具难以满足需求，迫使企业依赖人工经验，工程落地周期延长。

三、搭建人工智能数据工程能力核心要素

面向人工智能的数据工程核心旨在提升模型数据集管理与运营效率、提升数据集质量和数量、充分挖掘数据资源价值、保障模型数据安全可信，涵盖管理体系、开发维护、质量控制、资源运营、合规可信等五大核心要素。

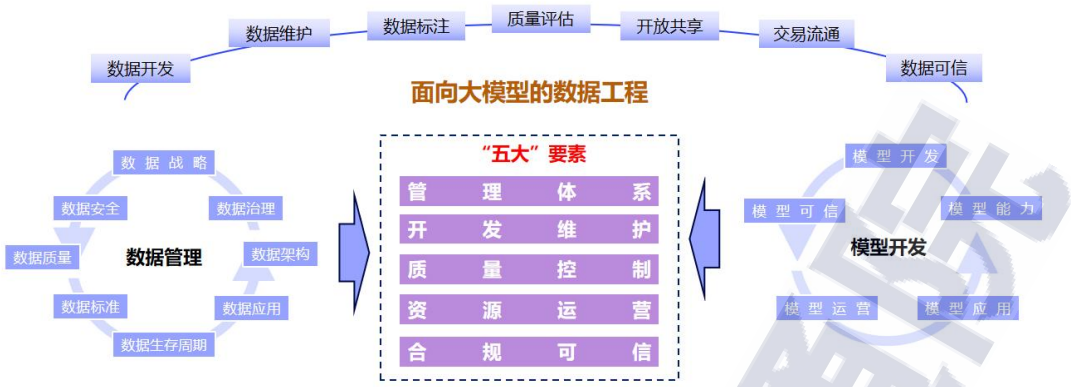


图 2 人工智能数据要素五大工程要素

（一）管理体系

全方位解决人工智能数据工程项目管理效率、团队协同能力以及技术应用标准化等问题，为大模型的成功部署与持续优化奠定坚实基础。

一是项目管理。专注于解决大模型数据工程全周期中的资源分配、进度控制、质量保证及风险应对等问题，旨在通过科学规划、精细执行与灵活调整，确保高质量数据集项目按时交付，成本可控。

二是组织建设。旨在设计并实施一个高效、协同的组织结构，确保从数据采集到模型应用的每一步都能够得到有效管理和支持，具体包括明确数据管理角色与责任、建立数据治理委员会、设置专门的数据科学与工程团队等。

三是人才管理。旨在建设一支跨学科、跨专业、跨领域的交叉复合型的大模型数据工程人才团队，具备数据科学和模型训练能力，掌握理解行业知识，了解预训练和微调数据构建、数据提示工程等关键数据构建和训练策略。

四是标准应用。指围绕大模型数据技术、平台、应用、管理、安

全等方面，制定数据生产标注服务能力、大模型数据集开发管理、数据集质量评估、合成数据生产与管理等数据服务标准和操作规范，并开展相关标准验证测试与持续优化。

（二）开发维护

人工智能数据集构建一般包括数据设计、数据采集汇聚、数据预处理、数据标注、数据质检等共性关键技术和环节，涵盖大模型数据集开发管理的预训练阶段、指令微调阶段以及反馈对齐阶段等定制方案。

共性人工智能数据工程技术工具，构建标准化底层能力。数据工程共性技术工具需覆盖数据全生命周期，解决标准化、自动化与质量控制问题。**数据设计规划阶段**，依据模型应用需求形成数据集设计方案和知识索引体系，梳理内外部数据资源，形成模型数据资源地图，为数据集开发和维护提供精准指引；**数据采集汇聚阶段**，需构建多源异构数据连接器，支持数据库、API、日志文件等格式的自动解析与融合，通过元数据管理实现数据血缘追踪。**预处理与标注阶段**，研发自动化工具链，例如基于规则引擎的异常值清洗模块、弱监督标注工具（如主动学习驱动的半自动标注）；**质量评估阶段**，建立“模型—数据”质量反馈评估能力，从完整性、一致性、语义合理性等维度自动检测缺陷，并联动修复工具实现闭环优化。

定制人工智能数据工程技术方案，面向应用的深度适配。预训练阶段需构建大规模高质量数据生成体系，通过定向采集领域知识，结合数据增强技术扩充样本多样性，同时需设计去偏处理算法，消除训

练数据中的性别、地域等潜在偏见。**指令微调阶段**需开发任务导向的数据构造工具，例如将用户问答数据转化为结构化思维链 CoT 数据集，结合知识图谱生成多轮对话上下文，并通过强化学习动态优化指令与答案的匹配度。**反馈对齐阶段**需搭建人类反馈闭环系统，设计多维度评价指标，利用偏好学习模型对齐人工评价与模型输出。此外，还需针对**行业特性定制方案**，确保数据工程与业务目标深度耦合。

（三）质量控制

大模型数据质量控制需构建覆盖“评估—优化—监控”全链路的闭环管理体系。数据质量直接决定大模型决策性能，需从评估准则、技术工具与流程管控三方面系统性突破。

评估准则层面，需建立多维度的量化标准。除传统完整性、准确性等基础数据质量指标外，需引入以模型应用为目标的人工智能数据集指令评估指标。参考 ISO 8000 数据质量系列国际标准，以及国家标准《信息技术 数据质量评价指标》，参照全国数标委《高质量数据集 质量评测规范》（征求意见稿）等标准文稿，中国信息通信研究院人工智能所建立了“可信 AI”人工智能数据集质量评估体系（ADAQ， Artificial intelligence Datasets—Quality evaluation），依据行业标准《面向人工智能的数据集质量通用评估方法 总体要求》（2021—1303T—YD），涵盖数据集完整性、规范性、准确性、及时性、一致性等 12 个一级指标和 36 个二级指标，如表 3 所示。此外，若数据集基于合成数据技术生成，应在质量评估中通过人工验证、自动化工具评估等方式评估合成数据集的保真度、隐私度、可解释度等，

以此提升数据集质量，并满足《数据安全法》《个人信息保护法》《生成式人工智能服务管理暂行办法》等法律、部门规章的合规要求。

表 3 人工智能数据集质量评估指标设计

序号	一级指标	二级指标	指标具体定义
1	完整性	规模完整性	数据集中的数据量是否满足人工智能模型应用需求的最低数量。
2		结构完整性	数据集中的样本所记录的信息是否完整，没有缺失值或缺失值在合理范围内。
3		领域完整性	数据集中数据可支持人工智能任务数量的覆盖程度。
4		任务完整性	数据集中数据可支持行业领域数量的覆盖程度。
5	规范性	形式规范性	数据集中数据符合一定的数据形式标准的程度，例如命名、创建、定义、更新和归档等需要遵循的标准，包括国际标准、国家标准、行业标准、行业实践经验或相关规定等。
6		隐私规范性	数据集中数据符合法律法规要求、行业标准、企业内部政策等隐私保护相关规范的程度。
7		安全规范性	数据集中数据符合人工智能模型训练的安全需求，确保模型的准确性和可靠性。
8	准确性	内容真实性	数据集中数据记录的信息是否与其所代表的实际对象、事件或事实一致和真实。
9		领域专业性	数据集中数据在特定领域的适用程度、专业深度和准确程度，且应当与特定行业的标准、术语、业务逻辑和专业知识紧密结合。
10	及时性	生成及时性	数据集中数据的产生时间应当尽可能接近当前时刻，尤其是在需要实时分析和响应的场景中。
11		采集及时性	数据集中数据采集的频率决定了数据集反映变化的灵敏度。高频采集可以提高数据的时效性。
12		处理及时性	数据集中数据从采集到数据可被分析使用的整个处理流程所需的时间也影响着数据的及时性，包括清洗、转换、加载等步骤。
13		发布与更新及时性	数据集更新的频率和更新后何时对外发布，直接影响数据使用者能否获取到最新的信息。

序号	一级指标	二级指标	指标具体定义
14	一致性	标签一致性	对于有标注的数据集，不同标注者之间是否对相同或相似实例的标注保持一致。
15		概念一致性	数据集中数据使用的术语、分类标准和定义在整个数据集内部是一致的。
16	稠密性	样本唯一性	数据集中样本是否存在大量重复记录，去除重复项可以提升信息稠密度。
17		内容信息量	数据集中单位样本所涵盖的信息量检测。
18	多样性	特征多样性	数据集包含特征种类的广泛和全面程度，是否覆盖了描述实体的不同属性和角度。
19		类型多样性	数据集包含类型种类的广泛和全面程度，是否覆盖了不同格式的数据，包括结构化、半结构化和非结构化数据。
20		来源多样性	数据集中数据源头或渠道的广泛和全面程度，以确保信息全面，减少偏见，提高应用场景的广泛性。
21		任务多样性	数据集支持人工智能任务的广泛和全面程度。
22		领域多样性	数据集覆盖行业领域的广泛和全面程度，以促进跨领域的知识迁移和应用。
23		文化多样性	数据集包含反映不同社会群体、文化背景、观点和利益相关者多元价值观的广泛和全面程度，避免出现偏见、歧视或违反社会主义核心价值观的数据内容。
24	均衡性	类别均衡性	数据集中各类别数据数量的均衡程度，确保数据集中各类标签的样本数量接近，防止模型偏向多数类。
25		来源均衡性	数据集中各来源数据数量的均衡程度，保证不同来源的数据量适当，提升模型的全面性和泛化能力。
26	相关性	领域相关性	数据集中数据与特定研究领域或应用领域的关联程度。
27		任务相关性	数据集中数据与所要完成的人工智能任务（如分类、回归、聚类等）的关联程度。
28		逻辑相关性	数据集中数据包含上下文无关的内容。
29	原创性	来源原创性	数据集中数据数据采集源头的原创性。即数据不是从现有的公开数据集中简单复制或改编而来，而是通过专门的实验设计、传感器网络、定制调查、独特观测等方式首次获得的。
30		内容原创性	数据集中数据的原创性。即使数据来源于已知的渠道，但其内容可能是前所未有的

序号	一级指标	二级指标	指标具体定义
			组合、特定情境下的记录或是独特现象的描述。
31	可溯性	来源可溯性	数据集中数据是否可以追踪其最初来源，包括数据是如何被收集的、由谁收集、在何时何地收集等信息。这对于验证数据的真实性和合法性至关重要，尤其是在科学研究和法律合规性方面。
32		链路可溯性	数据集中数据是否可以追踪其从原始形态到最终处理形态的整个转换过程，包括数据经过的清洗、转换、聚合、分析等每一步操作的记录。了解数据的演变路径有助于在发现错误或需要回溯分析时快速定位问题所在，同时也便于复现研究过程。
33		版本控制可溯性	数据集是否可以追踪其更新和迭代过程中每个版本的版本记录，包括变化历史、修改原因和责任人等。这对于团队协作、错误恢复及审计需求尤为关键，能够保证数据使用者能够获取并使用正确的数据版本，同时维护数据的历史完整性。
34	可访问性	获取难易性	数据集是否有清晰的使用许可协议，用户是否容易理解这些许可条款，以及获取数据是否需要特殊的权限或申请流程。
35		使用难易性	数据集是否配备详尽的文档说明，包括数据字典、元数据描述（数据来源、收集方法、更新频率等）、使用指南和示例代码等。
36		技术支持性	数据集是否提供官方或社区技术支持渠道，用户在使用数据过程中遇到问题时能否得到及时有效的帮助。

技术工具层面，需融合自动化与智能化手段。ADAQ 体系自建人工智能数据集质量评估工具平台，按照“规则检测+人工抽样+模型效果”的“三道关卡”融合思路进行开发，实现人工智能数据集质量评估指标的具体量化与项目执行，主要包括项目管理、参数配置、测评过程管理、可视化分析与总结、知识图谱管理、大屏展示等核心功能模块。

全流程监控层面，需贯穿数据生命周期。从采集阶段的元数据追

踪，到预处理环节的异常值实时清洗，再到模型训练中的质量反馈闭环。中国信通院将 ADAQ 数据集质量评估体系与“方升”大模型基准测试体系形成协同，通过对比模型输出与训练数据集，反向定位低质数据区间并提出优化机制。

（四）资源运营

人工智能数据集资源运营需构建“资源管理、开放共享、流通交易”三位一体机制，破解训练数据资源“存不好、管不住、用不活”的难题。

资源管理层面，需建立覆盖数据全生命周期的管理框架。以“资源目录”为索引，构建高质量数据集分类分级体系，按数据应用维度、训练阶段、模态情况细化分类标准。采用模型专家和业务专家联合的数据治理机制，按照模型需求梳理专业数据加工和标注策略，例如医疗数据需专家设计隐私脱敏策略，工业时序数据由专家规划预处理流程，确保数据集与行业需求精准适配。

开放共享层面，需考虑数据集和模型应用场景双重要素。内容规范，除数据集本身，需完整呈现采集来源、环境参数、结构规模、质量指标、隐私策略等，确保数据全要素透明。开放管理，明确开放时限、应用范围限制及版权协议，平衡开放力度与风险管控。场景化协议，针对 AI 训练、推理评测等场景，制定开放许可协议，规范数据供需方权责与使用方式。

流通交易层面，符合现有交易流通机制，鼓励模型数据生态合作。数据集首先明晰权属，确立数据所有权、使用权、交易权，构建登记

追溯体系，统一交易标准与合同范本，保障交易合规透明。鼓励数据方和模型方合作共建，形成资源融合推动产品和应用创新，建立共享联盟与合作框架，形成协同共进的流通生态，激活数据要素在模型训练、应用验证等环节的乘数效应。

（五）合规可信

高质量数据集合规可信是大模型可信的基石，需从数据合规与数据可信双向发力，既确保数据应用合法合规、版权清晰，又保障数据质量可靠、效果可溯。

数据合规以安全性、法律遵循和版权规范为核心，覆盖多重维度。一方面，数据需严格符合《中华人民共和国网络安全法》、《个人信息保护法》、《生成式人工智能服务管理暂行办法》等相关法律法规，以及网络数据爬虫有关合规要求，确保数据在采集、存储、使用、传输等各环节均合法合规，不涉及违法内容，建立数据安全管理机制、强化物理与逻辑防护。**另外一方面**，明确数据采集、生成、加工过程中的版权归属，避免权属纠纷，规范数据使用与分发的版权授权，确保数据来源合法，使用范围符合授权约定。

数据可信围绕来源、治理、结果、效果构建质量闭环。数据来源可信强调真实性、准确性、合法性，验证数据采集过程的客观性，确保获取渠道正规、授权完备。数据治理过程可信要求方案与流程透明可解释，治理规则（如清洗、标注标准）清晰留痕，操作过程可追溯。数据集结果可信以多维指标衡量质量，要求治理后数据分布合理，降低偏见样本率、毒化样本率，提升边缘案例覆盖度与标注准确性，

避免因数据偏倚导致模型决策偏差。数据集效果可信通过模型训练效果验证价值，对比治理前后模型的准确率、泛化能力等表现，以效果反推数据治理的有效性。

四、高质量数据集建设路径设计

为推动高质量数据集建设，相关主体可以采用“三步走”的战略，从体系规划到工程建设再到质量监测，形成一批行业高质量数据集。

图 3 高质量数据集建设路径

（一）体系规划阶段——构建高质量数据集认知框架

构建人工智能高质量数据集体系，规划阶段需以资源知识索引、地图构建、标准体系搭建为核心，驱动数据集系统化规划和建设。

一是构建知识索引，围绕智能化需求构建知识索引框架。结合行业特性，提炼核心知识节点，搭建层次化知识架构，如医疗行业以疾病诊断、药物研发为核心模块，金融行业围绕风险控制、客户营销构建体系。将既有数据资源与知识索引匹配，可实现数据的知识化归类。同时，知识索引为模型应用提供结构化路径，研发人员可依托索引快

速调用关联数据，加速模型训练迭代，让数据资源转化为驱动业务创新的智能生产力，最终实现从数据到模型的价值跃升。

二是锚定智能场景，构建企业人工智能数据集资源地图。深入拆解模型业务场景，如制造业的产线智能运维、服务业的客户需求预测等，明确场景对应的核心数据类型与质量要求，形成数据集设计方案。继而全面梳理内部数据资源，盘点数据目录清单，清晰标注数据类型（用户行为、设备运行数据等）、存储状态（实时流数据、历史归档数据）及权属关系。通过绘制资源地图，实现数据集的可视化呈现，形成“数据地图导航”能力。这不仅能快速定位场景所需数据，规避重复采集或资源浪费，更让数据检索从“模糊搜寻”转向“按图索骥”，为后续数据集开发提供精准指引，推动数据与智能场景的高效适配。

三是搭建标准体系，指导面向人工智能全周期高质量数据集建设。围绕标注服务、合成数据等生产活动，制定数据标注规范、合成数据质量标准，结合技术服务、流程管理要求，保障数据生产规范化。明确质量评估指标，规范自动化/人工检测流程，建立质量管理组织规范，确保质检环节有章可循，提升数据可靠性。在大模型应用端，制定数据与模型对接标准，规范资源运营、风险管理流程。针对工具平台，统一数据清洗、标注、存储等技术工具开发标准，保障工具兼容性与易用性。通过覆盖生产、质检、应用、技术的标准体系，可实现数据工程全链条标准化管理，减少模型开发方、数据运营方、数据管理方协作误差，提升数据质量控制与合规可信水平，为高质量数据集体系构建坚实支撑。

（二）工程建设阶段——打造高质量数据集生产体系

在构建人工智能高质量数据集体系的工程建设阶段，打造高质量数据集生产体系需从三大方向协同发力，以技术驱动、模式创新与生态协同筑牢数据根基：数据工厂模式奠定规模化生产基础，前沿技术突破质量与场景限制，生态协作整合外部资源，共同构建起覆盖效率、质量、创新的高质量数据集生产体系，为企业人工智能应用提供核心数据动能。

一是建立高效数据工厂模式，推动规模化生产。高效数据工厂模式以标准化、流水线作业为核心，重构数据生产流程。参照工业制造逻辑，将数据生产拆解为采集、清洗、标注、质检等模块化环节，每个环节制定精细化操作规范。例如，在图像数据生产中，明确采集设备参数、清洗过滤规则、标注分类标准及质检抽检比例，形成“输入—处理—输出”的标准化流水线。同时，引入自动化工具提升生产效能，如开发自动化数据清洗脚本，批量处理重复、错误数据；搭建智能标注平台，通过预设标签模板、自动推荐标注结果，减少人工操作成本。这种模式不仅实现数据生产的规模化扩张，更通过流程管控降低人为误差，确保数据质量稳定性，为大模型训练提供持续、可靠的数据源。

二是探索前沿技术路线，突破传统数据局限。前沿技术的应用是提升数据集质量与创新性的关键。其一，合成数据技术模拟真实场景生成稀缺数据，解决数据标注成本高、敏感数据获取难等问题。例如，在自动驾驶领域，合成复杂交通场景数据，补充现实中难以采集的极

端天气、罕见事故场景数据，扩大训练数据覆盖度。其二，智能化标注技术借助小样本学习、主动学习等 AI 算法，辅助人工标注。模型自动预标注数据，标注员仅需审核修正，大幅提升标注效率。其三，深度推理思维链数据聚焦模型逻辑可解释性，采集记录数据推理过程的思维链信息。例如，在知识问答模型训练中，不仅保留问题与答案，还记录推理依据、逻辑步骤，帮助模型学习更严谨的决策过程，提升复杂任务处理能力。

三是构建生态协作机制，整合外部资源力量。生态协作机制通过引入外部资源，扩大数据生产“朋友圈”。建立开放合作体系，引入专业第三方数据标注服务商，借助其规模化团队与丰富经验，分担数据生产压力。同时，构建科学的供应商评价和管理体系，从标注质量、交付时效、数据安全等维度设定评估指标。定期对服务商进行考核，筛选优质合作伙伴，淘汰不达标团队。此外，推动生态内协同创新，与高校、科研机构合作探索数据生产新技术，与上下游企业共建数据共享联盟，共同优化数据生产标准。通过生态协作，企业突破自身资源局限，在保障数据生产规模的同时，借助外部专业力量提升数据生产的专业性与创新性，形成多方共赢的数据生产生态。

（三）质量监测阶段——构建高质量数据集全流程管控机制

质量监测阶段的全流程管控机制，通过量化评估奠定技术基础，动态机制实现过程把控，持续优化闭环推动迭代升级，三者协同构建严密的数据质量保障体系，为人工智能高质量数据集体系筑牢质量防

线。

一是量化评估模型和工具，构建科学评估基准。量化评估模型与工具是数据质量管控的技术基础，企业需针对数据生产各环节，设计精细化量化指标体系。例如，在数据标注环节，以标注准确率、一致性、覆盖率为核心指标，通过计算标注结果与真实标签的匹配度量化质量；在数据清洗环节，以错误数据过滤率、完整性提升率衡量工具效能。同时，引入自动化评估工具，开发集成式质量检测平台，内置AI算法自动扫描数据缺失值、异常值，生成可视化报告。如利用AI检测模型快速校验文本语义一致性、图像标注准确性，替代人工抽检，提升评估效率与精准度，为质量管控提供科学依据。

二是动态评估机制，贯穿事前事中事后全流程。动态评估机制以时序化管理实现质量全程把控。事前规划，生产前明确质量标准与流程，制定数据采集规范、标注规则，设定质量准入门槛，从源头控制数据生产。事中监控，生产过程中部署实时监测系统，追踪标注进度、清洗效果，对超标的错误率即时预警，触发人工复核，避免问题累积。事后复盘，生产完成后系统性复盘，对比质量目标与实际结果，分析偏差原因。如针对模型训练中的数据偏差，回溯采集、标注环节，定位漏洞并形成改进方案，实现质量问题早预防、早发现、早解决。

三是持续优化闭环，建立数据与模型的反馈联动。持续优化闭环以模型效果为导向，形成质量提升循环。企业需搭建反馈评价体系，将模型训练、应用结果反向传导至数据生产环节。例如，若模型准确率下降，溯源定位数据标注错误或缺失，启动修正流程。同时，建立

常态化反馈机制，分析模型性能指标（准确率、召回率等）与数据质量的关联，总结影响规律，优化生产流程（调整标注规则、升级清洗算法），形成“模型反馈—质量诊断—流程优化”的闭环。通过持续迭代，数据质量与模型性能相互促进，推动数据集体系向更高质量演进，确保人工智能应用基于优质数据释放最大价值。

五、高质量数据集“炼化”流程和技术

高质量数据集建设需经历数据设计和采集、治理、标注、质检、运营等流程类似石油“炼化”过程中的勘采、粗炼、精炼、质检、运营等流程，推动原始数据资源迈向智能应用。

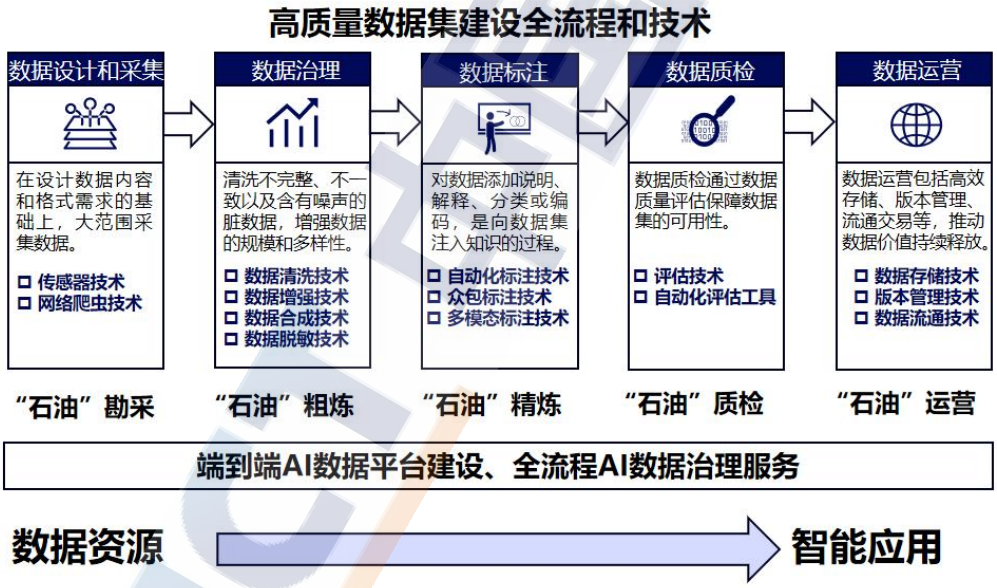


图 4 高质量数据集建设全流程和技术

（一）数据设计和采集

数据设计是指明确数据采集的内容、场景、格式、规模等要求，明确数据采集的手段。数据采集包括公开获取、数据采购、生态商业合作等方式，是汇聚高质量数据集建设原始数据资源的过程，技术层

面主要包括传感器技术、网络爬虫技术等。**一是传感器技术**。通过各类传感器，包括摄像头、麦克风、温度传感器等，能够采集图像、音频、环境等数据。例如在自动驾驶领域，激光雷达、摄像头等传感器可实时采集道路环境数据，为自动驾驶高质量数据集建设做好准备。**二是网络爬虫技术**。网络爬虫技术能从网页自动抓取数据，进行网页解析和数据提取，高效获取互联网上大量文本、图片等信息。

（二）数据治理

数据治理涵盖数据清洗、数据增强、数据合成、数据脱敏、数据质量评估等方面。**一是数据清洗技术是基础**。采集获取的原始数据往往存在噪声、异常值和冗余信息，需要进行系统性的数据清洗。例如，对于缺失值处理，常采用删除缺失值记录、均值填充、基于相似样本填充等方法，保障数据完整性；通过统计学方法或机器学习算法（如孤立森林算法）检测异常值，并进行相应处理，防止异常数据影响模型训练；利用哈希算法、字符串匹配算法等检测和去除重复数据，降低数据冗余。**二是数据增强技术用于扩充数据多样性**。为提升数据集的规模和多样性，需要采用适当的数据增强技术。例如，在图像数据集增强方面，通过旋转、翻转、缩放、裁剪、亮度调整、噪声添加等操作，增加图像数据的种类，提升模型泛化能力；在文本数据集增强时，采用同义词替换、随机插入或删除单词、句子重组等方式，扩大文本数据规模，提高文本模型性能等。**三是数据合成技术扩大数据规模**。为了提高数据的多样性、保护隐私、降低成本并增强模型的泛化能力，需要采用适当的数据合成技术。例如，基于统计模型的数据合

成技术通过模拟数据的概率分布来生成数值与时序数据，包括高斯混合模型和逆变化采样法等，具有较高的可解释性，适用于对数据统计特性要求较高的场景，如数据分析、数据补全和序列预测等。基于深度学习的数据合成技术包括生成对抗网络、变分自动编码器、Transformer 模型、扩散模型等，在图像、文本内容生成上表现出色。基于物理仿真的数据合成技术通过模拟现实中的物理现象和力学特征，生成三维空间数据和高仿真场景。例如，在自动驾驶领域，利用物理仿真与图形渲染等方法能够生成工业等领域逼真的场景数据。与此同时，当前合成数据技术使用中存在一定的风险隐患，例如可能合成存在带有歧视和偏见的数据、合成逻辑不合理的数据、对于长尾事件（如罕见疾病）合成与真实分布相差较大的数据等，需要进一步提升技术成熟度，并在合成数据应用中强化质量评估和管理。**四是数据脱敏技术确保数据安全**。为了确保数据集合法合规，需要对涉及个人隐私、商业秘密等数据进行脱敏处理。例如，替换脱敏用虚构或无意义的值替换敏感数据，如身份证号码、银行卡号等；加密脱敏运用加密算法对敏感数据加密，只有通过特定密钥才能解密还原，保障数据在非授权访问时的安全性等。

（三）数据标注

数据标注是指对数据添加说明、解释、分类或编码的过程，以便数据可以被人工智能算法所理解和使用，是向数据集注入人类知识的过程，是提升数据集质量的关键步骤。当前，数据标注技术创新主要聚焦在自动化标注、众包标注、多模态标注等多个关键技术方向。**一**

是自动化标注技术。自动化标注技术利用机器学习和深度学习算法，自动对数据进行标注。这种技术可以显著提高标注效率，减少人工参与，降低成本。例如，商汤科技通过大模型技术对自动驾驶的路测回流数据进行自动标注和重建。然而，自动化标注技术在某些复杂任务上可能无法达到手动标注的准确性，因此仍需要与人工标注相结合。**二是众包标注技术**。众包标注技术通过引入激励机制、质量控制和任务分配策略，将数据标注任务分发给大量网络用户，从而提高标注效率。这种技术可以充分利用互联网上的闲置人力资源，但需要注意保证标注质量和一致性。**三是多模态标注技术**。随着多模态数据集在人工智能模型训练中的广泛应用，跨模态数据标注技术变得越来越重要。例如，利用注意力机制等技术，关注多模态数据中的关键信息，提高标注的准确性和效率。

（四）数据质检

数据质检通过数据质量评估保障数据集的可用性。为评估数据集质量，需要研发质量评估技术，并针对评估指标研发算子及对应自动化评估工具，支撑数据质检高效开展。**在评估技术方面**，学术界和产业界使用了研究基于分类器和困惑度的数据集质量评估方法等，重点评估预训练数据集质量。例如，基于分类器的数据集质量评估技术的原理是，计算评估数据集与参考数据集（如维基百科）的相似度，相似度越高则数据集质量越高。GPT-3 使用该方法从 45T 的 Common Crawl 数据集中，挑选出 570G 高质量数据集。基于困惑度的数据集质量评估技术的原则是，利用语言模型（如大模型、n-gram 模型）计

算评估数据集中每个句子出现的概率，概率越大，则困惑度越小，数据集质量越高。**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聚焦如表 3 所示 12 个一级指标和 32 个二级指标，采用“自动化为主+人工校核辅助”的检测方式，共开发 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。例如，在规范性指标方面，开发了敏感字段检查工具；在稠密性指标方面，开发了相似样本查重工具；在均衡性指标方面，开发了公开网页来源均衡性检查工具等。

（五）数据运营

数据运营涉及数据存储、版本管理、流通交易等多个环节。**一是数据存储技术支撑数据集的有效存取**。例如，向量数据库专门针对向量数据的存储与检索进行优化，能高效处理高维向量数据，将向量数据与相关元数据关联存储，支持快速的向量相似性搜索，为大规模机器学习、深度学习模型训练和推理过程中的数据调用提供有力支持，极大提升数据获取与利用的效率。**二是版本管理技术工具助力数据集管理**。借助专业的版本控制系统，通过数据库+对象存储相结合的方式确保任何时间点的数据状态可以被追踪、恢复，并支持团队协作。在实际应用中，借助版本管理工具可以便捷地回溯到任意数据集历史版本，为数据分析、模型训练等后续工作提供支撑。**三是数据流通技术保障数据集安全可信流通**。例如，通过区块链技术保证数据集交易安全、透明、可追溯。通过搭建数据集交易平台，提供信息匹配、交易撮合、安全传输等服务，促进数据集在合规前提下流通。通过建设

数据开放平台提供高质量数据集查询、下载和 API 接口调用等，方便用户获取使用数据集。通过联邦学习技术，在数据不出本地的前提下，实现跨机构、跨地域的数据协同建模与分析，各方仅交换加密的模型参数或中间计算结果，避免原始数据的直接传输与暴露，实现“数据不动模型动”“数据可用不可见”，推动人工智能数据在更大范围内的安全共享与深度应用。

六、总结展望和建议

随着人工智能大模型应用从初步探索迈向更为复杂、智能的高阶阶段，对高质量数据集的规模、多样性、时效性以及处理速度的要求将会快速增长。展望未来，数据集工程、技术创新、质量评估、版权合规以及基础制度建设是推进人工智能高质量数据集建设的关键。

（一）建立 AI 数据工程体系

数据工程能力决定了数据集建设和处理的效率。数据工程能力弱将严重制约人工智能产业发展，导致模型训练滞后、应用效果不佳等问题，影响人工智能赋能千行百业发展。建议加快建立包括数据工程服务平台、标准规范、团队建设、项目管理在内的 AI 数据工程体系。**一是打造功能完备的数据工程服务平台**。整合各类先进的数据采集、清洗、标注等工具，以一站式服务模式，有效降低机构数据集建设成本，提升数据集建设和运营效率。**二是完善数据集工程能力相关的标准规范**。明确数据处理各个环节的技术指标和操作流程，推动机构数据工程朝着规范化、标准化方向发展。**三是加强数据工程团队建设**。积极引进具备数据科学、软件工程、行业专业知识等多领域交叉背景

的复合型人才，为团队注入创新活力。**四是建立科学高效的数据工程项目管理体系**。提高数据工程全流程进行精细化管控能力，从项目规划、资源合理分配到进度实时监控、质量严格把控，确保数据工程项目按时、高质量交付。

（二）推动 AI 数据技术创新

传统的数据采集、处理和存储等技术已难以满足高质量数据集建设的需求，迫切需要合成数据、自动化标注技术创新提升数据集的“量”和“质”。**一是攻克数据多模态融合技术，**实现图像、文本、音频等多种数据类型的高效融合与协同处理。研发自动化标注技术，利用深度学习算法对数据进行自动分类和标记，大幅提升标注效率与准确性。**二是探索数据压缩与高效存储技术，**采用新型编码方式减少数据存储空间，同时不影响数据的完整性与可用性。**三是聚焦向量数据集技术，**通过优化向量数据库架构，提升向量数据的存储与查询效率，为人工智能模型的快速检索与匹配提供支持。**四是攻关重点场景定制化数据集生产技术。**在数据较难采集的低空经济等领域，使用合成数据技术。在标准化程度较高的自动驾驶等场景下，采用众包等方式发动大量标注员快速完成标注；在专业化程度较高的医疗等领域召集专业医学专家组成标注团队，确保数据标注准确、高效完成。

（三）搭建全流程 AI 数据质量管理体系

通过科学、严格的质量评估，及时发现并解决数据集中存在的问题，能够为人工智能模型提供坚实的数据基础，提升模型的稳定性和泛化能力。**一是完善机构数据集质量评估和管理体系**。在国家和行业

标准的基础上，细化质量评估指标，定制测试方案，在评估指标和方法设计中，考虑涵盖测试数据集准备、测试条件初始化、前置检测、后置检测以及数据集质量评估得分计算；在测设方案设计中，要搭架测试公共服务平台，通过规则检测+人工抽样+模型效果等过程，客观、全面、公立的开展数据集质量评估。**二是推动数据集“以评促建”**。在客观、准确的数据集质量评估基础上，对于评估过程中发现的质量问题，及时采取整改措施，并对整改后的数据集进行再次评估，确保质量达标。通过评测真正实现数据集开发管理能力和质量水平提升。

（四）加快 AI 数据开发利用机制突破

加快完善数据权属、定价、收益分配、安全治理等数据基础制度，是规范高质量数据集市场秩序、保障各方合法权益的迫切需要，是实现人工智能数据相关产业可持续发展的关键所在。在市场机制和法律法规要求下，积极探索版权合规机制、数据集定价机制、数据集收益分配机制，形成行业典型示范。**一是强化数据集建设运营版权合规性**。机构应强化版权意识，确保采集数据合法授权，不擅自超出授权范围使用数据，对于机构自主生成的数据集及时进行版权登记，维护企业的合法权益和良好声誉。**二是积极探索数据集定价机制**。综合考虑数据集的规模、质量、稀缺性、应用场景、潜在价值等多方面因素开展定价探索。**三是积极探索数据集分润机制**。通过协商谈判等形式探索数据集收益分配机制，保障数据采集者、标注者等基础劳动者的合理权益，同时兼顾开发者、使用者等其他主体的利益，促进数据生态的平衡发展。

附表 1 国家高质量数据集相关政策

序号	政策发布部委	政策发布时间	政策名称	有关政策内容
1	国家数据局等 17 个部门	2023 年 12 月	《“数据要素×”三年行动计划（2024—2026 年）》	支持交通运输龙头企业推进高质量数据集建设和复用，加强人工智能工具应用，助力企业提升运输效率。 深入挖掘各类科学数据和科技文献，通过细粒度知识抽取和多来源知识融合，构建科学知识资源底座，建设高质量语料库和基础科学数据集，支持开展人工智能大模型开发和训练。 在科研、文化、交通运输等领域，推动科研机构、龙头企业等开展行业共性数据资源库建设，打造高质量人工智能大模型训练数据集。
2	工业和信息化部等 4 个部门	2024 年 6 月	《国家人工智能产业综合标准化体系建设指南（2024 版）》	规范人工智能研发、测试、应用等过程中涉及数据服务的要求，包括数据采集、数据标注、数据治理、数据质量等标准。
3	国家发展改革委等 4 个部门	2024 年 12 月	《关于促进数据标注产业高质量发展的实施意见》	加强交通、医疗、金融、科学、制造、农业等重点行业领域数据标注，建设行业高质量数据集，支撑人工智能在行业领域的应用赋能。
4	中国气象局	2023 年 5 月	《中国气候数据集建设工作	研发关键气候变量均一化数据集、风云卫星气候数据集、天气雷

序号	政策发布部委	政策发布时间	政策名称	有关政策内容
			方案（2023—2025 年）》	达气候数据集，开展实况分析产品历史回算，研发全球和中国区域大气再分析产品、中国极端气候事件专题数据产品、多场景应用的人工智能基准数据集。
5	国家数据局	2024 年 11 月	《可信数据空间发展行动计划（2024—2028 年）》	面向新药研制、新材料研发，推动基础科学数据集、高质量语料库汇聚，促进人工智能驱动的科研范式创新应用。
6	国家发展改革委等 6 个部门	2024 年 12 月	关于促进数据产业高质量发展的指导意见	支持企业面向人工智能应用创新，开发高质量数据集，大力发展“数据即服务”“知识即服务”“模型即服务”等新业态。
7	中央网信办	2023 年 7 月	生成式人工智能服务管理暂行办法	推动生成式人工智能基础设施和公共训练数据资源平台建设。促进算力资源协同共享，提升算力资源利用效能。推动公共数据分类分级有序开放，扩展高质量的公共训练数据资源。鼓励采用安全可信的芯片、软件、工具、算力和数据资源。

附表 2 地方高质量数据集相关政策

序号	地区	政策发布时间	政策名称	有关政策内容
1	山东省	2025 年 3 月	《山东省“数据要素×”创新应用省级财政支持政策及资金管理实施细则（暂行）》	原则上每个主体每年奖补额最高不超过 40 万元，每个设区市每年累计奖补额度最高不超过 200 万元
2	北京市朝阳区	2025 年 4 月	数据要素产业奖补项目方案	支持企事业单位围绕智慧城市、金融、医疗健康、教育、消费等领域建设开放高质量数据集、语料库等，支持通用人工智能大模型和面向行业应用的大模型训练。根据数据规模、数据质量、更新频率和应用效果等，给予最高 200 万元资金支持
3	湖北省武汉市	2025 年 3 月	《〈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）（征求意见稿）》。	每年发布一批高质量数据集建设任务专项，对完成专项建设并通过评审的企事业单位，按建设投入成本的 30%给予不超过 200 万元的奖励
				对利用数据产品支持人工智能发展且通过评审的企事业单位，按照数据产品购买金额的 30%给予最高 200 万元补助
4	江苏省南京市玄武区	2025 年 3 月	《玄武区关于促进数据产业发展的支持政策》	鼓励构建高质量自然语言、多模态等行业数据集和语料库，支持训练、验证、测试、语料等数据集通过江苏数据交易所挂牌，对于成功交易的，按数据规模给予每款最高不超过 5 万元开发奖励，同一主体年度奖励金额最高不超过 20 万元。

序号	地区	政策发布时间	政策名称	有关政策内容
5	江苏省无锡市	2025年4月	关于建设“人工智能+”标杆城市的政策意见	发放“数据券”，对企业购买非关联方数据集开展大模型研发应用的，给予不超过合同金额30%的补助、最高200万元。 支持企业建设人工智能行业高质量数据集，对被超过3家非关联企业用于大模型研发应用且示范效应明显的，给予补助最高200万元。
6	内蒙古自治区呼和浩特市	2025年3月	《关于促进绿色算力及人工智能产业高质量发展的若干意见》	围绕教育、医疗、文旅、交通、气象、农牧、林草、城市治理、能源等领域打造一批高质量数据集，每年安排1000万元资金，按训练使用量、数据质量等对符合条件的数据集给予奖励
7	浙江省杭州市	2024年7月	《杭州市人民政府办公厅关于高标准建设“中国数谷”促进数据要素流通的实施意见》	支持对具有商业价值的数据集合进行数据知识产权权益登记。 完善社会数据集中采购制度，支持行政事业单位、国有企业等加大对中小企业数据产品和服务采购力度，引导社会力量积极提供数据资源、数据产品和服务。 支持训练、验证、测试、语料等数据集通过杭州数据开放平台向社会开放，每年评选不超过5个数据集，按照不超过实际投入的30%给予奖励。同一单位年度奖励金额最高不超过100万元，同一数据集不得重复申报
8	河南省	2025年2月	《关于开展2025年“行走河南·读懂中国”文化和旅游数字	支持相关单位建设高质量文化旅游数据集，主要采取以奖代补、财政补助等方式，对符合条件的单位择优予以支持，单个项目不超过

序号	地区	政策发布时间	政策名称	有关政策内容
			化奖补资金申报工作的通知》	项目总投资的 30%。
9	广东省深圳市光明区	2025 年 3 月	《关于推动人工智能和软件信息产业高质量发展的若干措施》	对高质量数据集服务商，上年度数据服务业务达到一定规模的，按不超过服务合同金额的 30%给予最高 500 万元奖励，连续三年 对人工智能企业通过数据交易平台购买非关联方语料数据的，按其年度实际购买合同金额的 30%，每年给予最高 100 万元的一次性资助，连续三年
10	福建省厦门市	2025 年 4 月	《厦门市促进数据产业高质量发展若干措施》	鼓励企业汇集行业内上下游企业数据资源，开展数据清洗、标准化、标注等数据治理工作，形成多模态的行业高质量数据集，通过厦门公共数据融合开发平台向社会开放，提供跨企业、跨行业的数据服务。每年评选一批成效显著的项目，按不超过项目实际投入的 30%的比例给予一次性补助，每个项目最高补助不超过 500 万元。
11	广西省南京市	2025 年 4 月	《南宁市支持中国—东盟人工智能创新合作中心高质量发展第一批政策措施》	企业、高校、科研机构建设包括语料库在内的高质量数据集，最高给予 100 万元奖励。
12	上海市	2023 年 7 月	《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）》	建设高质量数据集，开展数据质量评估评价，构建面向大模型的高质量语料库，形成标准操作流程和技术规范。

附件 行业高质量数据集建设代表性实践

（一）教育领域：高等教育学科高质量数据集建设实践

（1）数据集建设背景

当下随着人工智能技术发展，高等教育领域的格局正在经历深刻的重塑。教育部、国家语委、中央网信办共同印发的《关于加强数字中文建设推进语言文字信息化发展的意见》提出，到2027年初步建成国家关键语料库。在关键学科、重点行业、战略区域、民生期待和社会急需领域，分批建设规范、安全、优质的国家关键语料库。高等教育学科高质量数据集建设中面临挑战具体表现在以下四个层面。

一是预训练数据技术壁垒突出。教育领域的预训练数据需处理大量非结构化内容（如公式、图表），这部分内容解析、标注自动化水平和准确度还不够，人工标注投入巨大。

二是微调数据积累处于初级阶段。教育大模型的微调需场景化、多模态数据支持，但当前数据采集尚未形成系统化能力。教育场景的强交互性（如师生对话、课堂行为分析）数据仍稀缺。教育部等九部门提出构建“多模态语料库”重点布局思政、心理健康等垂直领域大模型，推动数据采集与教学全流程融合。

三是高质量评测数据集不足。特别是在高等教育和职业教育的细分领域，存在明显的缺口，学科深度评测数据仍依赖专家知识动态更新，且标准化程度较低。

四是数据生态体系与商业模式待完善。头部企业通过关联公司封闭处理大量核心数据，导致教育数据流通率小，头部企业的标注工具链与其云计算架构深度耦合，外部不容易使用或者付费使用。

（2）数据集建设实践

高教社在过去 71 年的发展过程中，积累了大量内容经过专业审核和意识形态属性审核的高等教育、职业教育学科资源，利用高教社在教育领域的行业优势，汇聚、标注加工，正在建设高质量语料库资源，为国内大模型的训练和能力提升提供数据支撑。

提升数据标注质量。按照高等教育教学语料的特点和适用场景，分层分级的地针对视频语料设计了四级标注分类。第一级标注是补足元数据，包括来源、学科、标题、内容简介、分辨率、时长等信息。第二级标注是补齐精准的字幕文件，确保与音视频在时间轴上对齐。第三级标注是对音视频内容进行符号逻辑的拆分，并对不同片段添加描述信息。第四级标注是对整个音视频资源的知识点进行总结描述，并与教育部指定的知识图谱进行关联。

强化平台功能建设。围绕两个工具平台，一个应用中台体系进行建设。两个工具平台分别为：高教社语料库及 Lovong 创新应用平台，汇聚并萃取社内外海量高质量教育领域语料，提供数据采集、清洗、筛选、标注等功能；高教社语料存储与管理平台，该平台用于存储和管理已加工过的语料信息，方便进行快速检索、查找与导出。一个中台为 LoVong“龙凤”人工智能开放应用平台，平台具备简洁易用的操

作页面和标准化 API 接口，为高校用户及社内各业务部门提供技术和平台支撑，赋能高教社业务创新和重点项目发展。

提高质量管理能力。建立一套高效、严格的数据集质量评估管理方案。

一是构建多维度质量评估标准体系。结合不同行业特点和人工智能应用需求，制定完善细致、全面、科学的质量评估标准，涵盖数据的准确性、完整性、一致性、时效性、安全性、隐私保护等多个维度。加强对质量评估标准的推广和宣贯工作。①关于分级标准制定。建立基础性指标，覆盖准确性（如 OCR 校对后人工复核）、完整性（教材知识图谱全节点覆盖）、时效性（教材版本更新）。同时，增设教育专属指标，如教学适用性、学科权威性等。②关于安全规范落地。对一些教育数据进行脱敏，通过技术监控—加密—销毁全流程监控防护，实施全生命周期管控。

二是建立“生产—评估—优化”全流程闭环。①关于数据标注规范化流程。一方面，按数据类型进行分类分级生产管理：对文本类、音频类、视频类数据实施不同生产流程，对理工类和文史类教材数据执行不同标注标准。例如，针对理工类教材数据的复杂公式，进行 latex 转写和人工精准标注。对高等数学、线性代数、概率论与数理统计、数学分析等课程教材进行了精标，全部公式转为大模型可以识别的 latex 语句，确保数据准确性和权威性。另一方面，多模态数据实现线上平台自动解析，人工只需复核争议内容，大幅提升效率。②关于动态质量评测机制。一方面，开展企业自评，组建专项质检团队，按

数据类型设定差异化抽检比例。另一方面，组织第三方评估，委托高校实验室或第三方 AI 数据集质量评估机构，定期对题库、科研文献数据集进行公正、客观的评估，以数据集质量评测引导数据集建设。③关于持续优化机制。建立“问题溯源—流程改进—复检验证”闭环，对于评估过程中发现的质量问题，及时采取整改措施，并对整改后的数据集进行再次评估。此外，将典型错误案例转化为数据标注人员的培训素材，帮助标注人员不断提升能力水平，形成可持续发展的质量优化机制。

三是培育高质量数据生态的核心举措。①关于提升评测机制促进质量。充分利用质量评估结果，持续优化数据生产流程，不断提高数据集质量，为人工智能模型训练提供高质量的数据集支持。②关于产教融合促进人才培养。联合多所院校建立产学研实践基地，培养兼具学科知识和技术能力的数据专项人才，为打造高质量数据集提供生产力保障。

（3）数据集建设成效

在高等教育领域，无论是智能教学系统、个性化学习平台，还是学科研究中的人工智能应用，高质量的数据集可以确保模型的准确性与可靠性，高等教育出版社建设的高质量数据集在以下方面提供支持:

一是用于高教社龙凤教育文化大模型训练。结合大规模教材语料进行增量预训练，提升模型在教育领域 4 项特色能力（答案精准可信、启发性分步骤输出、正向激励情绪价值、推荐个性化教育资源）及相关领域的文本润色能力。训练数据来源于 50,000 多本教材和 5 万多

门优质课程，覆盖多学科知识体系。评测场景拓展至文科及文理科综合性质的采访、规划性材料、教材段落润色，由专业编辑团队验证能力。

二是应用于数学分析解题大模型。通过对数学分析单门课程的教材数据收集和 DeepSeek-R1 的蒸馏等，形成这门课程的高质量解题数据集，基于高质量解题数据集通过 LoRA 微调和全参 SFT 训练的方式训练大模型，提升了高等教育领域数学分析这门课程的解题能力。

三是应用于智能体训练。将整理后的数据集通过 RAG 的方式，应用于智能体训练。提升智能体回答的准确度，回复的内容更加贴近用户实际需求。

（二）科学领域：材料科学高质量数据集建设实践

（1）数据集建设背景

当前，材料科学研究正在经历传统“经验寻优”的第一范式、基于理论的第二范式和基于海量数据计算模拟的第三范式，构建大规模材料科学数据集至关重要。北京市计算中心有限公司作为国家材料基因工程专项的技术支撑单位，参与构建的多尺度材料数据库平台，整合了第一性原理计算、机器学习验证和实验表征等相关数据，目前已形成覆盖金属材料、半导体材料等八大类材料的高质量材料科学数据集，为科研院所和企业研发部门提供数据接口服务。

（2）数据集建设实践

一是在体系规划阶段，明确数据集采集的类型包括包含每种材料的晶胞参数、晶体结构、力学性质（弹性常数、杨氏模量、泊松比、

强度等），电子性质（能带结构、带隙、费米能级、载流子浓度等）等。明确数据集通过高通量计算模拟、实验测量数据、文献挖掘、数据集成与共享等方式采集。构建符合 FAIR 原则（Findable, Accessible, Interoperable, Reusable, 简称 FAIR）的数据本体与元数据模式，对材料数据进行结构化描述。参考中国材料基因工程标准化组织发布的材料数据标识（MID）规范等进行数据标注。

二是在工程建设阶段，构建端到端的自动化数据处理流水线：从数据抓取到数据验证的完整流程脚本化和调度化。包括定时抓取公开资源（如 Materials Project 更新、文献数据库接口）、自动格式转换（如将不同软件输出结构统一为 CIF/JSON）、自动预处理（单位转换、归一化等）以及自动校验（格式校验、逻辑检查）。借鉴 CI/CD 思想，使用流水线工具（如 GitHub Actions、Jenkins）或任务调度框架实现持续集成与部署。流程中自动记录数据版本和处理日志，关键步骤（如数据源变动、新增条目）触发通知或质检提醒，保证数据管线的透明性和可审计性。

三是质量评估与管理阶段，根据《信息技术 数据质量评价指标》（GB /T 36344-2018）提供的规范和框架与说明，建立数据规范性、完整性、准确性、一致性、时效性、可访问性等度量标准。构建自动与人工质量审查机制。自动审查通过预设规则（例如能量最小性检查、格式合法性检查）和统计方法筛查异常；机器学习方法也可用于异常检测（如孤立森林）。项目组定期评估数据集整体质量，必要时组织领域会议讨论数据质量标准，形成可执行的质控细则。

此外，对数据集进行版本管理，每次数据更新（新增、删除、修改）都生成新版本并记录变更说明。同时，建立数据可信度分级体系，根据数据来源、质量评审结果和贡献者信誉等维度给数据加权标记。所有可信度信息包含在元数据中，供机器学习模型训练时参考或作为加权依据。

（3）数据集建设成效

基于本数据集大量第一性原理计算数据，北京科技大学和中国航发北京航空材料研究院对镍基单晶高温合金开展了研究。通过大量计算指导研制“工业燃气轮机涡轮叶片用镍基单晶高温合金”的周期与国外先进水平相当。此外，与传统“试错法”相比，新型镍基高温合金研制成本降低一半以上。

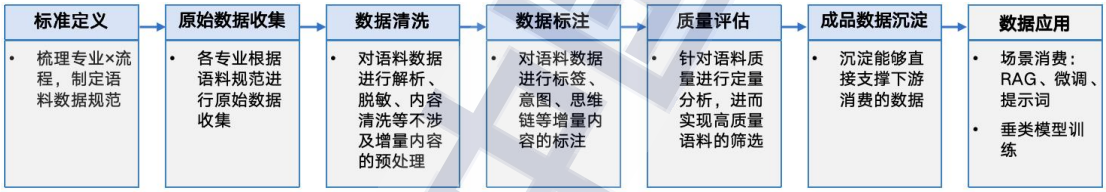
（三）通信领域：网络运维高质量数据集建设实践

（1）数据集建设背景

稳定、高效、智能的网络是运营商服务亿万用户、支撑千行百业数字化转型的生命线。传统的网络“规划、建设、维护、优化、运营”全生命周期管理，高度依赖人工经验与操作，不仅导致运维成本，更在效率、敏捷性及客户体验方面存在显著瓶颈。基于网络运维各阶段的痛点问题，中国移动通信集团浙江有限公司部署实施“点金行动”，以“流程、数据、AI”三元深度融合为驱动，围绕网络域开通、故障、优化、变更、投诉等重点场景打造30多个专业智能体，推动运维模式实现革命性升级。其中，核心是打造高质量网络运维数据集。

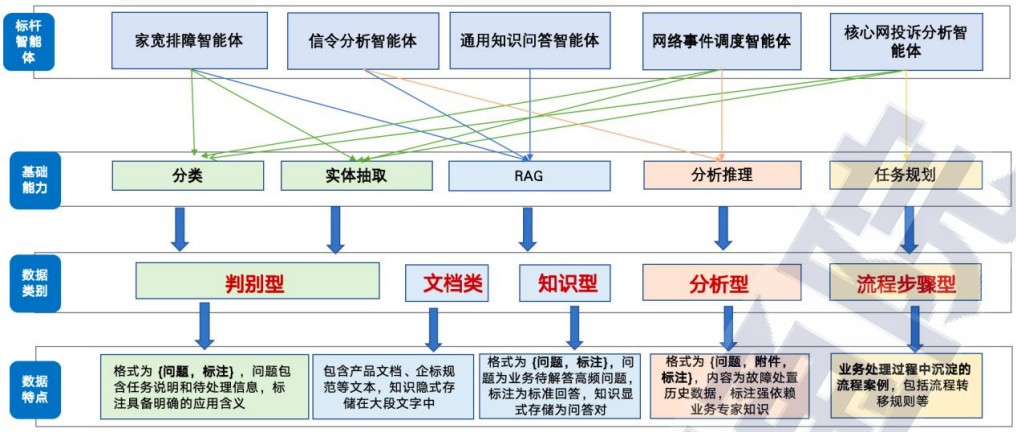
（2）数据集建设实践

在建设思路和方法论方面，数据集构建的方法论如下图所示，整体是一个“收集原始数据→加工处理数据→沉淀成品数据”的流程。首先，需要定义清楚数据的标准，包括数据收集的业务范围，即通信运营商哪些网络运维领域的哪些生产管理流程，以及数据收集的类型范围。获取原始语料后，对数据进行包括解析、脱敏等在内的清洗操作，这一步属于预处理，一般不涉及内容上的增加。接着，利用人工+自动的方式，对数据进行标签、意图、思维链等内容的标注。完成上述处理后，再对数据的质量进行定量分析评估，筛选出高质量的数据语料进行保留。最终，沉淀出成品数据集，能够直接给到下游（如需求场景的智能体建设）进行消费。



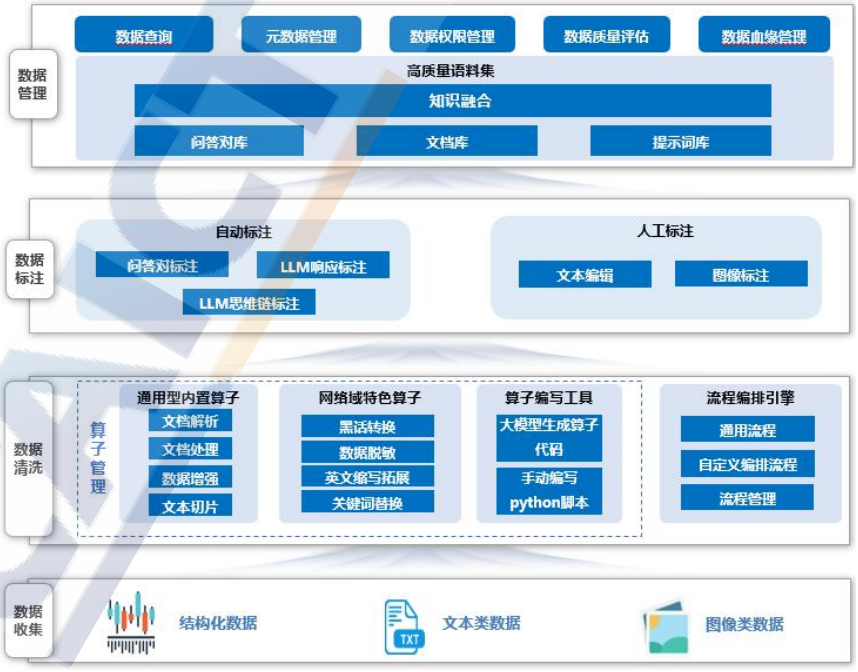
附图 1 网络运维高质量数据集建设全流程

从运营商网络域当前场景的实际需求出发，梳理出下图所示的数据类型，共两大类，一类是无标注、纯文本的文档类数据，一类是以“问+答”为基本形式的带标注数据，根据数据特点又进一步细分为判别型、知识型、分析型、流程步骤型。



附图 2 网络运维智能体数据需求情况

在数据使能平台建设方面，业界主流智能体开发平台在场景化应用开发中满足了基础数据使用需求，但在“AI+自智网络”业务实际落地中仍面临诸多挑战。为此，数据使能平台提供数据收集、数据清洗、数据标注、数据管理等数据一站式服务，打造问答对、文档、提示词三类网络域公有高质量语料集，为大模型训练测试、微调、知识检索提供数据服务。



附图 3 网络运维数据集使能平台建设

在数据质量评估体系方面，旨在以自动的方式对数据质量进行定量或定性的评分，助力筛选出高质量的数据进行沉淀。拟定出下图所示的评估体系，从结构完整性、信息时效性等表层的评估，到内容准确性、语言流畅性、标签一致性、合规性等内容性的评估，再到信息密度、信息价值等深度评估，全方位地产出语料的质量评分。各项指标具体的计算，采用大小模型结合的方式进行，大语言模型负责进行文本语义理解和定性的评估，基于 TF-IDF、信息熵等特定公式的小模型负责从某一特定维度进行定量评估。

附图 4 网络运维数据集质量评估指标

（3）数据集建设成效

在数据集建设方面，依托数据使能平台的建设，现已沉淀一批高质量网络数据集。例如：**核心网运维领域**，沉淀高质量文档类数据 721 份、微调语料 15465 条，供给研发侧进行知识库构建与模型微调；**家庭业务领域**，沉淀高质量文档类数据 209 份、微调语料 5600 余条，

供给研发侧进行知识库构建与模型微调；**监控运维领域**，沉淀高质量文档类数据 6 份、微调语料 2693 条，供给研发侧进行知识库构建与模型微调；**安全维护领域**，沉淀微调数据 80712 条，供给研发侧进行模型微调；**IP 承载网领域**，沉淀高质量文档类数据 45 份，供给研发侧进行知识库构建；**基础设施动力环境维护领域**，沉淀高质量文档类数据 911 份，供给研发侧进行知识库构建。目前，这些数据已在核心网运维、家庭和业务支撑、监控维护等专业的场景中实现落地应用，并打造了多项标杆案例，其应用成效也初步收获了域内的好评。

在数据集赋能智能体建设方面，一是建设网络运维问答智能体，旨在为网络条线打造“百科全书”式的全专业知识问答能力，当前已收集各类型知识文档超 2000 个，涵盖专业日月报、设备手册、企标规范、培训材料等多维度知识类型。当前该能力已经上线浙江移动玩过运维 APP，支撑网络条线日常知识查询，周调用量 200+。一些典型的用例如下图所示。**二是核心网信令分析智能体**，旨在应用于网络投诉处理中的投诉根因定位环节，通过模型原生能力来实现信令涵义解析、异常识别、流程识别和根因推理，解决人工分析效率低、门槛高等问题。目前信令分析智能体已接入投诉系统，用于支撑分析 4/5G 上网业务投诉根因定位，接入后自动化同时并发处理多个投诉工单信令分析任务，码流分析时长从 1 小时降至 10 分钟内，分析准确率 85%以上，投诉工单处理效率提升 15%以上。

（四）交通领域：交通运输政策法规和标准规范高质量数据集建设实践

（1）数据集建设背景

目前，交通运输政策法规与标准规范通过政府网站的政府信息公开和行业标准化管理平台发布，然而其智能化服务水平不高，存在着全文模糊检索信息过载、标准精准检索需专业背景、政策附件和标准全文需下载后阅读等问题。交通运输部科学研究院通过合规采集交通运输领域政策、法规和标准数据，建成交科智汇·交通政策法规与标准规范数据集，用于行业科技大模型以及细分子领域模型训练，有效提升通识类知识智能问答的性能表现。

（2）数据集建设实践

一是全面的整合政策法规与标准数据资源。通过公开数据库下载、购置等途径，收集交通运输领域（含公路、水路、铁路、民航、邮政）以及与交通运输行业相关的法律、行政法规、部颁规章、行政规范性文件、地方性法规、国家标准及行业标准等数据资源，形成待加工的行业文本知识。为探究政策法律变迁以及技术演进与优化，在采集中包含了失效法律和废止标准数据，并且通过人工标注梳理了替代历程，保障了数据间关联的准确性。

二是建设高质量交通政策法规与标准规范数据集。在对文本数据进行预处理后，通过文本加工大模型自动生成文本问题、评估问题质量、生成问题答案、将问答对组装，形成了 15 万对的交通政策法规与标准规范数据集，其中标准类 3.7 万对、政策类 10 万对、法规类

1.3 万对。针对在建设中出现的数据增强导致答案编造政策文件名称的问题，通过政策名称实体匹配策略，即提取问答对中的政策名称，将其与原始文本中的政策文件和文号进行匹配，再由人工核对、完善未找到对应政策的信息内容，较好地解决了数据集的命名实体幻觉错误。

（3）数据集建设成效

交通运输高质量数据集可应用于政策法规咨询，政策影响模拟分析，标准管理，工程设计审查，交通执法，物流运输合规管理，自动驾驶决策等智能化场景。目前，交科智汇·交通政策法规与标准规范数据集已应用于交科大模型智能检索/问答/咨询应用场景的微调训练，微调后交科大模型在政策法规与标准领域的对话准确率明显提升。未来，还将基于交通政策法规与标准规范数据集精选数据样本建立评估数据集，可用于评价交通运输细分领域或方向的政策法规与标准问答类模型/场景的泛化能力和理解指令要求。

（五）工业领域：基站机房运维高质量数据集建设实践

（1）数据集建设背景

传统通信基站机房资产管理模式面临三大核心痛点：首先是物理空间分散性带来的监管盲区，全国范围的机房设备分布呈现出地理跨度大、环境差异显著的特征；其次是人工巡检的时效滞后性，常规的周期性巡检难以捕捉设备状态的瞬时异常；再者是异构数据源的解析瓶颈，既有监控系统产生的视频流、红外热成像、夜间补光图像等多模态数据尚未形成统一的分析范式。

过往轻量级 CV 模型的解决方案，在复杂场景下暴露出泛化能力差、无法解析设备运行状态、难以捕捉资产异常行为时序特征等问题，为此需要基站机房运维的多模态大模型，本质是通过视觉理解技术重构资产管理范式。模型需同时实现四大核心能力：设备本体识别、运行状态监测、异常状态检测、风险预测预警。这种从“看得见”到“看得懂”的能力跃迁，对底层数据质量提出了前所未有的要求。

为此，中国铁塔积极推动建设服务机房运维的多模态大模型。该多模态大模型的训练依赖三大数据维度：空间维度需覆盖不同厂商设备的工业设计差异，时间维度需包含设备全生命周期状态变化，环境维度需囊括南北地域气候特征对成像质量的影响。具体表现为：在数据广度上，需采集配电箱在不同光照条件下的表面纹理特征；在数据深度上，需标注空调压缩机从正常运转到故障劣化的渐变过程；在数据关联度上，需建立开关电源状态与机房环境的耦合关系图谱。

（2）数据集建设实践

一是在体系规划阶段，数据维度设计采用"三层金字塔"架构：底层为设备本体特征库，包含空调压缩机锈蚀纹理、电池组接插件形态等静态属性；中间层为状态变化时序库，记录配电箱开关位移轨迹、机柜门开合、指示灯等动态特征；顶层为异常行为特征库，整合盗窃破坏的设备减少、误操作导致的指示灯不亮等风险场景。通过这种结构化设计，确保每个数据样本都携带设备 ID、时空戳、环境因子等元数据标签。

二是在工程建设阶段，数据生产线建设部署四条数据处理流水线。在多模态数据同步方面，对同一设备的可见光/红外视频帧建立时序对齐索引。在专家协同标注方面，开发三维标注工具支持旋转视角下的设备状态判定，并建立“双闭环”校验机制，前向闭环通过预标注—人工核验—模型迭代流程提升初始标注质量，反向闭环利用在线学习实时检测标注偏差。在对抗样本生成方面，模拟暴雨天气摄像头水渍、强光反射等干扰场景。在数据集版本化管理方面，基于时间、地点、环境等数据精准化数据版本。

三是质量监测阶段，建立质量评估体系，数据效能评估采用“五力模型”。具体包括，表征力（样本覆盖设备型号变体的完备性）、判别力（同类状态差异与异类状态相似度的量化指标）、泛化力（在未见过的新机房场景中的迁移表现）、鲁棒力（对抗光照变化、遮挡干扰的稳定性）、时效力（从数据更新到模型迭代的响应速度）。

（3）数据集建设成效

项目构建的多模态视觉数据集彻底改变了传统基站机房运维模式。基于深度学习的设备状态解析能力，能够识别空调压缩机状态、电池组运行状态等传统人工巡检耗时发现的潜在风险。特别是在配电箱状态监测场景中，算法通过分析开关手柄的毫米级位移，可准确判断是否存在虚接隐患，这种精细化识别能力直接推动了预防性维护体系的建立。

数据集支撑的异常行为检测模型展现出强大的场景适应性。针对基站机房常见的设备丢失场景，系统能够区分正常维护人员的工具操

作与盗窃行为的模式；对于设备操作规范性监测，可识别维护人员未关闭机柜门、工具遗落在设备区等风险行为。这种基于视觉语义理解的智能分析，使得安全隐患发现时效从原来的小时级提升至秒级响应。

（六）医疗领域：面瘫相关语音高质量数据集建设实践

（1）数据集建设背景

面瘫是一类常见的神经性运动障碍，直接影响面部肌肉控制功能，进而对个体的语言表达能力产生显著影响。由于构音受限、语速变化和发音不稳定，面瘫患者在语音交流过程中存在大量可观察的语言异常表现。当前，人工智能辅助诊断系统在医学语音领域正逐步推进，但面向此类特定人群的语音数据资源极为匮乏。

大多数公开语音数据集聚焦于普通话标准发音者，缺乏关于病理性发音状态、构音异常特征与生理评估信息的整合数据资源。这种缺口限制了语音识别模型、异常检测模型在医学语境下的应用，也阻碍了语言障碍相关研究的发展。

为此，中国科学技术大学组织相关项目，旨在构建一套具备医学背景标注与语音表现特征并存的面瘫语音数据集，满足以下多重需求：①为语音识别与发音异常检测模型提供真实、结构化的训练样本；②为面瘫辅助诊断与治疗跟踪提供量化语音指标；③为语言学与康复医学交叉研究提供可复用的高质量数据基础。数据集将主要服务于以下几个方向：①医疗人工智能系统中的语音筛查与分级评估；②病理性语音识别与语音异常建模研究；③医学康复过程中的语言恢复状态追踪；④医工交叉的辅助诊断工具开发与模型验证。

（2）数据集建设实践

在体系规划阶段，一是明确数据建设定位。项目初期即将数据集定位为医学语音专项资源，目标明确服务于语言障碍分析与辅助诊断研究，覆盖采集、标注、处理、管理、运营等全流程。规划中充分考虑语音信号与病理状态的对应关系，设计了结合结构化语音样本与临床分级信息的复合数据结构。二是系统设计语料内容结构系统化。语料设计依据语言学覆盖原则与面瘫发音特征分析结果，分为三类：基础单字、构音短语及复杂句子，确保覆盖各类音节、声母韵母组合、语调模式与句法结构，适应模型训练与病理分析的双重需求。三是注重采集标准与伦理合规。从项目启动阶段即纳入数据伦理管理，采集设计严格遵循医学伦理审批流程。明确不采集视频、生物识别信息或个人身份内容，全部语音样本仅绑定匿名受试者编号。所有被试均签署知情同意书，数据使用范围限定在研究场景内部。

在工程建设阶段，一是自主研发采集系统。建立专用语音采集平台，支持 Windows 操作系统，客户端可自动加载任务文本、录音、命名、标记状态，并实时同步到后台系统。系统具备异常检测与重录功能，支持任务调度与轮次控制，方便采集管理人员按计划推进录音。二是执行高保真录音标准。采用 48kHz 采样率、16 位深度、单声道 PCM 格式进行采集，统一使用定向麦克风设备。采集环境为低噪声、无回声录音间，确保语音数据质量达到语音识别与声学分析研究标准。三是数据结构化管理设计。每条音频样本与任务编号、受试者 ID、录制时间、轮次等信息自动关联，采集完成后即时分类归档。元数据

同步记录采集设备参数与受试者状态标签，建立数据追溯机制，支持多维度数据检索与访问。

在质量评估管理阶段，一是自动检测与人工复核结合。治理阶段引入静音比例、响度分布、信噪比等指标进行音频质量检测，对采集样本进行问题筛查。发现缺损、杂音、异常时长等情况后，由采集人员复录。对可保留样本执行静音段裁剪、响度调整等修正处理。二是标注体系与临床评估融合。除文本时间戳标注外，项目特别引入面神经功能评估标注，记录每轮数据对应的 H-B 分级分数，并由专业人员补充标准化的症状观察信息。标注结果与音频数据共同组织为结构化数据单元，便于建模与统计分析使用。三是版本控制与访问权限管理。治理与标注完成后，数据按版本编号归档，记录处理时间、修改人、变更内容。系统设定多层访问权限，根据角色授权采集人员、标注人员与建模人员使用不同数据子集。全部使用操作均自动记录，形成审计日志，保障数据使用合规可控。

（3）数据集建设成效

项目采集了 100 例面瘫患者和 25 例正常受试者的普通话语音样本，是目前国内首个系统化融合面部神经功能评估信息与标准语音数据的资源库。每一位患者的录音数据均附带 H-B 分级评分，并结合医师观察记录了其典型症状表现，包括额部运动、眼睑闭合与外翻、口角运动、鼻唇沟变化等多个维度。这种“语音+分级+症状”一体化的结构设计，使数据具备较强的解释性和分析价值，能够支持临床语音表现与生理指标之间的建模研究。

项目数据目前已被用于开展语音识别模型训练、发音异常检测实验以及语音与病理状态关联建模等工作，初步结果显示，面向面瘫患者语音的识别精度与病理状态分类能力在使用该数据集后得到明显提升。此外，数据也作为语料基础，服务于高校课题研究与医学会议投稿，进一步验证其在实际科研场景中的适用性与价值。系统支持数据多版本管理和权限划分，具备依据伦理审批流程向合作单位分级开放的能力，为后续研究拓展与联合攻关打下基础。

（七）文化领域：方言高质量数据集建设实践

（1）数据集建设背景

语音数据是方言与少数民族语言保护的数字化载体。中国语言资源保护工程已建成世界最大的汉语方言资源库，收录 1700 多个调查点的语音数据，为方言语音特征分析、濒危语言抢救提供了实证资料。此外，高质量语音数据可通过技术手段（如语音合成）重现方言发音，助力地域文化传承。为持续拓展方言数据集，中国社会科学院语言研究所持续推进方言数据集建设，记录方言区普通话的发音特征，为语言学研究与方言保护提供了双重支持。

（2）数据集建设实践

在体系规划阶段，一是明确战略定位与应用场景，以解决方言口音与场景噪声对语音识别的挑战为核心目标，针对智能家电、汽车语音交互等产业场景，规划覆盖多方言、多环境的数据集架构。二是构建多维度数据框架。例如，发音人分层设计方面，按年龄、性别、教育背景等社会属性配置发音人结构（如每方言点 200 人，男女各 100

人），避免数据偏差。语料内容设计方面：整合独白（160 个话题）、日常问题（工作、数字等）、口语句子（500 个）、方言词汇及语音平衡句（2200 个），覆盖口语与朗读语体，确保语言特征的全面性。

在工程建设阶段，一是数据采集与质控。通过学校、社区招募发音人，确保地域口音纯正；同步录制近距（2-3cm）、中距（20-30cm）话筒信号，室内环境增加远距离拾音，捕捉不同声场特征；实时检查录音质量，发现模糊或噪声超标即刻重录，避免后期返工。二是多层级数据标注与处理。对正则文本、声韵母、重音、韵律单元等 7 层信息进行标注，实现从音素到语义的全维度刻画。先通过机器自动标注提升效率，再由专业人员人工校正（如方言发音偏差、副语言学现象），确保标注准确率。

在质量评估管理阶段，通过实时监听、波形可视化检查录音清晰度，记录背景噪音分贝值，剔除信噪比低于阈值的样本。采用“双标注+交叉验证”模式，随机抽取 10%标注数据进行一致性检验，误差超过标准的批次重新标注。依据《评价标准》检查语音特征完整性（如声韵母覆盖率、韵律标注准确率），通过专家评审与模型测试（如识别准确率评估）双重验证，确保数据符合项目要求。此外，还要基于用户反馈与模型应用效果，定期更新标注规则（如新增方言词汇标注），补充稀缺场景数据（如极端噪声环境），保持数据集的时效性与适用性。

（3）数据集建设成效

建成汉语地方普通话语音数据库 RASC863，覆盖 10 大方言区（如

上海、广州、重庆等 10 个方言点），收录 2000 人（男女各 100 人/方言点）在 4 类场景（室内安静、室外马路等）的语音数据，标注正则拼音、声韵母、韵律等 7 层语言学信息，为方言口音识别提供核心训练素材。同时，基于数据集建设经验制定了《语音数据集标准化建设指南》，成为国内“少数民族语言语音库”等项目的建设模板。支撑发表语音学论文 200 余篇，获国际奖项（如国际语音科学会议最佳论文），相关技术应用于华为、科大讯飞等企业的方言识别系统，推动智能设备在方言区的普及率提升 15%。

（八）商贸领域：商贸流通行业高质量数据集建设实践

（1）数据集建设背景

当前，人工智能技术进入与实体经济深度融合的 2.0 阶段，行业数据的质量、规模与多模态融合能力直接决定 AI 模型的产业落地价值。商贸流通行业具有供应链链路长、参与主体多元、数据模态复杂（如商品编码、物流轨迹、交易文本、视频监控、物联网传感数据等）的特点，亟需构建覆盖全场景、全链条的行业级数据集，以支撑线上购物、新零售业态、机器人导购、数字贸易、移动支付等场景的算法模型迭代。

商贸流通行业智能化转型面临三大核心制约，一是数据资源化程度低，行业数据标准化缺失导致跨系统、跨企业数据融合困难，仅 15%的企业实现供应链全链条数据贯通（据 Gartner 统计）。二是工程化能力薄弱：传统数据处理工具难以支撑多模态数据（文本、图像、时序、空间数据）的联合分析，模型训练数据准备周期长达 3-6 个月。

三是应用碎片化严重：90%的 AI 应用停留在单点场景（如线上购物），缺乏覆盖“商流-物流-资金流-信息流”的一体化数据驱动决策体系。为此，北京海天瑞声科技股份有限公司积极构建商贸流通高质量数据集。

（2）数据集建设实践

在数据采集方面，商贸流通行业数据广泛分布在政府、行业协会、社交媒体、行业企业、第三方机构等单位。在数据采集渠道商，以采集电商平台数据为例，数据获取途径包括：①API 接口。通过电商平台（如淘宝、京东、拼多多等）提供的 API 接口，开发数据抽取工具，获取用户行为数据、交易数据、商品评价、搜索记录等。②数据合作。与电商平台签订数据合作协议，获取特定数据或参与数据共享项目。③数据市场。通过电商平台的数据市场或数据交易平台，购买相关数据产品。

在数据标注方面，一方面强化标注内容多样性和针对性。在商品描述文本中，会标注商品的名称、品牌、规格、材质、价格区间等属性信息；对于用户评价，会标注情感倾向（正面、负面、中性）、评价主题（如质量、服务、物流速度等）。物流文本标注涉及订单编号、发货地、收货地、运输方式、配送时间等关键信息。合同协议类文本则着重标注合同双方、关键条款、有效期、违约责任等内容。另一方面，强化标注流程严谨性。首先对原始文本数据进行清洗和预处理，去除重复、错误或无效的信息。接着对标注人员进行全面培训，使其熟悉商贸流通行业知识和标注规则。标注人员按照统一标准对文本进行细致标注，之后由经验丰富的审核人员进行多轮审核，对标注质量

不达标的数据返回进行修正，直至达到高质量标准。

在数据质量控制方面，为了控制标注数据的质量，采用了如下的策略。①制定明确的标注标准和指南：包括标注的详细要求、示例以及标注过程中需要注意的事项，并确保所有标注人员了解标注标准。②自研了高质量的标注工具和软件：这些工具可以提供标注模板、自动化标注功能和校验功能，帮助标注人员提高效率和准确性。③进行标注质量控制：设置质量检查机制，对标注数据进行随机抽查或全量检查，以确保标注质量符合标准。我们采用了自动化检测工具或人工检查相结合的方式来发现和纠正标注错误。④标注数据管理系统：建立有效的数据管理系统来跟踪标注数据的版本和变化，确保标注过程的透明性和可追溯性。⑤实施标注人员培训和反馈机制：对标注人员进行培训，使其熟悉标注标准和工具，提高标注人员的技能水平。此外，定期收集标注人员和使用者的反馈，并根据反馈不断优化标注标准和流程。

（3）数据集建设成效

海天瑞声已建设总量约 12.96TB 的商贸流通行业高质量数据集。具体包括:①品牌营销数据集,包含数字人图像数据库 248 人 3600GB、直播带货广告营销语音数据库 1041GB; ②数字贸易数据集，包含多语种客服数据库 2135GB、商贸语料库 10,000,000 句; ③数字供应链数据集，包含发票数据库 5000 张 1GB、交通图片库 20,000 组 4GB、仓储场站高精标图片库 1,000,000 张 5.93TB; ④新型消费数据集。包含消费票据数据库 122,000 张 24.5GB、超市监控视频库 380GB、商

品图片库 50,000 张 10GB。相关数据集已服务众多商贸流通场景，如消费场景、电子商务场景、商超等场景，服务代表性企业有阿里、京东等单位。

编制说明

本研究报告自 2025 年 3 月启动编制，分为前期研究、框架设计、文稿起草、征求意见和修改完善五个阶段。本报告整体分析了高质量数据集总体发展现状，提出了人工智能数据工程的核心要素，以及高质量数据集建设路径、流程、技术，展示典型行业代表性案例，并进一步提出了发展建议，为政策制定者、行业从业者及企业投资者等提供全面的行业洞察、策略建议与决策依据。本报告由中国信息通信研究院人工智能研究所撰写，撰写过程中得到了人工智能关键技术和应用评测工业和信息化部重点实验室的大力支持。

参编单位：中国信通院河北科技创新研究院、中国社会科学院语言研究所（中国社会科学院语言学重点实验室）、交通运输部科学研究院、中国科学技术大学、北京交通大学网络空间安全学院、中国邮政集团有限公司、高等教育出版社有限公司、中国铁塔股份有限公司信息技术研究院、中国移动通信集团浙江公司、中国大唐集团技术经济研究院、贵州大数据产业集团有限公司、北京市计算中心有限公司、北京海天瑞声科技股份有限公司、亚信科技有限公司、高颂数科智能技术有限公司、北京希尔贝壳科技有限公司、浪潮卓数大数据产业发展有限公司、软通智慧数据科技有限公司、菲特检测技术有限公司、标贝科技有限公司、北京轩宇信息技术公司、中伯伦信息技术有限公司、北京识因智能科技有限公司、北京猎户星空科技有限公司、北京枫清科技有限公司、中关村数智人工智能产业联盟。

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn