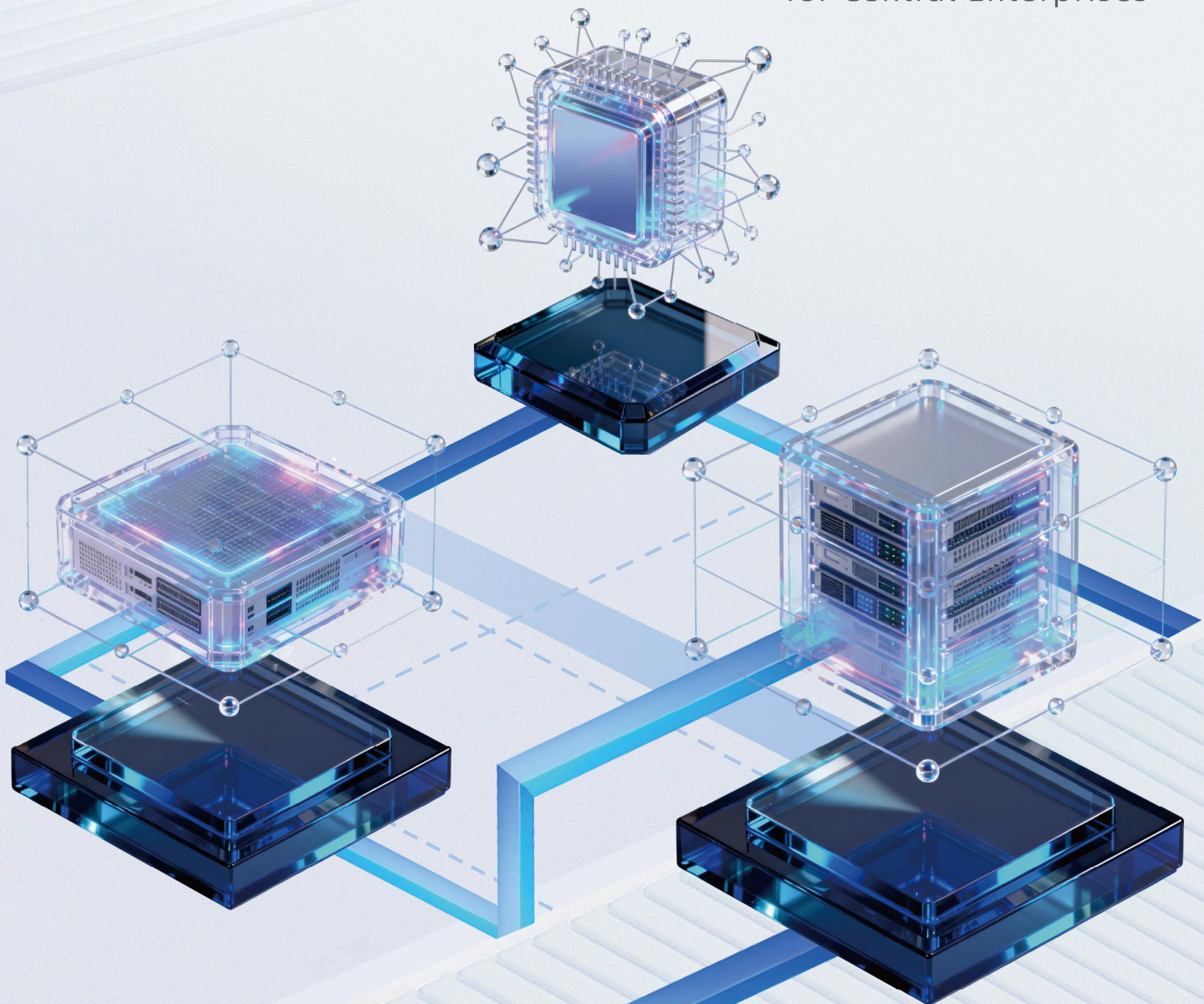
数据产业集团 | CAICT 中国信通院

中央企业高质量数据集建设 研究报告

Research Report on the Construction of High-Quality Datasets for Central Enterprises



—— 2025年12月 ——

版权声明

编制说明

本报告的撰写得到众多企业与专家的支持和帮助，牵头单位和参编单位如下。

牵头单位：中电数据产业集团有限公司、中国信通院云计算与大数据研究所

参编单位：中国石油天然气集团有限公司、国家石油天然气管网集团有限公司、中国南方电网有限责任公司、中国第一汽车集团有限公司、中国铝业集团有限公司、中节能大数据有限公司、中国交通信息科技集团有限公司、国机数字科技有限公司、中国移动通信集团有限公司、中国电信集团有限公司、中国联合网络通信集团有限公司、新兴际华集团有限公司

目 录

一、央企高质量数据集建设背景

在新一轮科技革命和产业变革深入推进的背景下，高质量数据集已成为支撑人工智能发展和行业智能化转型的关键基础。近年来，国务院国资委围绕实施央企“人工智能+”行动和产业焕新行动，将高质量数据集建设作为提升中央企业智能化能力和核心竞争力的重要抓手，通过专题部署、示范发布和平台建设等方式，持续推动数据资源向可用、可管、可共享的数据资产转化。与此同时，随着能源、制造、交通、通信等重点行业的智能化转型不断深化，对高质量、多模态、可持续迭代的数据集需求日益凸显，数据集建设正从单点建设应用，转向体系化建设和加速行业支撑。在产业需求与政策引导的双重驱动下，央企高质量数据集建设逐步进入系统性推进时期。

（一）发展趋势：行业智能化转型依赖高质量数据集

当前，人工智能正加速向各行业核心业务环节渗透，推动生产方式、管理模式和决策机制发生深刻变化。**行业智能化转型已不再停留在应用辅助分析，而是逐步向生产运行优化、风险预测预警和系统协同等方向拓展。**这一趋势对数据的规模、质量提出了更高要求，单纯依赖零散数据或业务系统数据已难以支撑复杂模型训练和规模化应用，高质量、可复用、可持续迭代的数据集正成为行业智能化发展的关键基础。

从数据需求来看，能源、交通、制造、通信等行业的智能化应用往往涉及设备端侧数据、业务数据与外部数据的融合，数据呈现出来

源多样、标准不一、时序跨度长等特点，需要建设高质量数据集实现统一组织与治理。一方面，企业普遍拥有大量设备、系统和长期积累的数据资源，具备开展智能化应用的基础；另一方面，由于业务专业性强、运行环境复杂，对模型的可靠性、稳定性及可解释性要求更为严苛，不仅要求数据规模实现“多”的突破，更强调数据质量达到“可用、好用”的标准。所以，需要通过系统化的数据集建设，将分散在不同系统、不同阶段、不同模态中的数据进行统一组织、规范处理和质量控制，形成能够真实反映业务运行状态和关键规律的数据集，从而满足人工智能模型对大规模、高质量训练数据集的需求。

从应用实践来看，高质量数据集正在逐步成为承载行业知识、支撑模型训练、提升人工智能应用能力的重要载体。通过围绕典型业务场景构建结构清晰、标签明确、质量可控的数据集，企业能够将隐含在长期运行过程中的经验、规则和模式转化为模型可学习、可泛化的输入要素，从而显著提升智能应用的落地效果和稳定性。高质量数据集已从支撑性资源转变为基础性能力，企业高质量数据集建设能力已经在很大程度上决定了央企智能化转型的深度和质量。

（二）政策驱动：全面支持高质量数据集建设

政策引导持续加强，行业高质量数据集建设提速。2023 年 12 月，国家数据局等 17 部门联合印发《“数据要素 ×”三年行动计划（2024—2026 年）》，提出在科研、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。2024 年，国家发展改革委、国家数据局

等四部门联合印发的《关于促进数据产业高质量发展的指导意见》首次明确提出“支持企业面向人工智能应用创新，开发高质量数据集”“大力支持重点行业高质量数据集建设”，该《指导意见》为后续各行业启动专项建设提供了政策依据。2025 年，围绕“人工智能+”行动和数字中国建设，国家层面加速系统化部署和行业政策出台。一方面，通过《数字中国建设 2025 年行动方案》《关于深入实施“人工智能+”行动的意见》等文件，将高质量数据集作为人工智能应用和产业发展的关键基础，作为重点任务统筹推进；另一方面，围绕交通运输、政务等重点领域，陆续出台行业高质量数据集建设方案和应用指引，明确数据集建设的业务场景和关键要素，推动高质量数据集加速向行业、场景建设。

表1 高质量数据集相关部分政策及内容要点

发布时间	政策名称	发布机构	相关内容
2025.10	《政务领域人工智能大模型部署应用指引》	中央网信办、国家发展和改革委员会	《指引》强调场景牵引、规范部署和运行管理。提出“政务部门应加强政务数据治理，持续提升数据质量，加快构建客观反映公共政策、制度规范、业务流程和治理实效的高质量政务数据集和知识库，支撑政务大模型的优化训练”“依托政务数据共享协调机制，统筹数据治理成果，推进高质量政务数据集的共建共享和生成数据的归集治理”。
2025.9	《交通运输行业高质量数据集建设方案》	中华人民共和国交通运输部	围绕“多模态覆盖、多场景贯通、多任务协同”的思路，提出建设行业通识数据集，并综合考虑场景应用需求必要性、技术可行性、经济可行性、数据基础条件等因素，面向基础设施、交通装备、运输服务、行业治理、绿色低碳、安全保障、科技创新等 7 个业务领域，提出公路基础设施状态监测、航道运行风险评估与智能监测、多运输方式物流运输成本优化等 39 个行业专识数据集建设场

发布时间	政策名称	发布机构	相关内容
			景。
2025.8	《关于深入实施“人工智能+”行动的意见》	国务院	提出加快实施 6 大重点行动，包括“人工智能+”科学技术、产业发展、消费提质、民生福祉、治理能力、全球合作；强化 8 项基础支撑能力，包括提升模型基础能力、加强数据供给创新、强化智能算力统筹、优化应用发展环境、促进开源生态繁荣、加强人才队伍建设、强化政策法规保障、提升安全能力水平。
2025.5	《数字中国建设 2025 年行动方案》	国家数据局	部署了体制机制创新、地方品牌铸造、“人工智能+”、基础设施提升、数据产业培育、数字人才培育、数字化发展环境优化、数字赋能提升共 8 个方面的重大行动，明确提出“积极开展人工智能高质量数据集建设”“建设行业高质量数据集”。
2025.1	《关于促进数据标注产业高质量发展的实施意见》	国家发展改革委、国家数据局等四部门	从深化需求牵引、增强驱动创新、培育繁荣生态、优化支撑体系、加强保障措施六个方面促进数据标注产业发展。明确提出“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”。
2024.12	《国家数据基础设施建设指引》	国家发改委、工业和信息化部	强调在建设数据高效供给体系方面，要在数据标注产业的生态构建、能力提升和场景应用等方面先行先试，链接公共数据、企业数据和个人数据，形成统一的数据资源开放目录，并研究制定高质量数据集建设的相关标准，确保数据标注的准确性和专业性。此外，要构建集成数据采集、存储、清洗、标注、管理、应用等功能的一体化数据基础通用工具平台，提升数据加工效率和保证数据质量。
2023.07	《生成式人工智能服务管理暂行办法》	国家网信办、国家发改委等七部门	使用具有合法来源的数据和基础模型，采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性。在生成式人工智能技术研发过程中进行数据标注的，提供者应当制定符合本办法要求的清晰、具体、可操作的标注规则；开展数据标注质量评估，抽样核验标注内容的准确性；对标注人员进行必要培训，提升遵法守法意识，监督指导标注人员规范开展标注工作。

二、央企高质量数据集建设主要问题和挑战

近年来，在政策驱动和需求推动下，央企高质量数据集建设加快，行业高质量数据集不断形成，在模型训练、业务应用中取得突出成效。但由于当前数据集建设以项目制为主，从整体能力来看，高质量数据集建设仍处于起步阶段，仍存在制度、标准、技术和生态等方面的问题和挑战。

内外部制度待细化落地，制约高质量数据集系统化建设。当前，央企在高质量数据集建设中普遍缺乏制度规范的约束，既包括跨主体的数据基础制度，也包括企业内部的配套制度和细则。一是数据所有权、使用权和收益分配等关键制度不清晰，缺乏明确的定价与利益分配机制，导致各央企对数据共享、流通交易普遍持审慎态度，数据资源获取和高质量数据集流通存在障碍。二是数据汇聚机制缺失，尤其是涉及行业基础性、共性数据时，因行业缺乏统一的汇聚共享机制导致部分数据集重复建设。另外，在央企内部，由于高质量数据集建设处于起步阶段，企业以项目化方式推进数据集建设，尚未形成覆盖全环节的制度体系，影响高质量数据集的持续建设。

标准体系不完善，数据获取与共享困难。多数行业尚未形成覆盖数据采集、格式、标注和质量评估的标准体系，行业内不同企业间的数据结构、元数据、标注方式存在差异，造成数据的互通、标注结果和质量结果互认存在困难。例如，工业领域，不同设备的协议、数据格式和接口不统一；交通领域，时序数据的标注需结合统一的时空现状和物理规则；生态环境领域，部分地方的生态环境开发的管理系统

多达数十套，各系统的数据标准和技术接口均不统一。

技术支撑能力不足，影响建设效率和效果。技术架构方面，尽管部分央企已搭建数据平台，具备数据标注、质量评估等工具，但数据采集、处理、标注、评估、应用全过程不打通，容易造成开发、标注等过程难以追溯、验证和迭代。数据处理方面，缺乏面向非结构化、多模态数据的智能数据处理工具，导致多模态数据的结构和特征提取能力不足，非结构化数据的联合建模可用性较差。数据标注方面，部分数据标注工具对 BIM 模型、影像等非结构化数据的标注效率较低，影响构建效率。质量评估方面，质量评估常作为事后环节开展，而非贯穿于采集、处理、标注等各阶段，导致系统无法实现对采集偏差、标注错误、样本分布失衡等问题的即时发现与修正。

协同生态未建立，阻碍数据集流通应用。目前不少央企虽已开始构建企业或行业可信数据空间，但尚未形成成熟落地的运行模式。空间内的主体准入、应用成效、收益分配等运营机制和规则均未完全明确，导致开发的数据集难被其他需求方使用，使用反馈、再开发和迭代更是难以实现。另外，数据提供方、服务方与使用方之间缺乏以“数据即服务”为核心的商业模式，数据产品供给形式单一、价值实现路径模糊，进一步阻碍了高质量数据集的服务化推广和产业化落地。

三、央企高质量数据集实践分析

高质量数据集建设是一项基础性、长期性、系统性的工程，建设过程不仅涉及数据采集、处理、标注、评估等多环节，也涉及多公司、多部门的协同。高质量数据集建设运营主要内容和实施路径如下图所示，本章节从数据集建设、数据集运营和基础保障体系三部分对央企的具体实践情况展开深入分析。

注：关于具体建设内容与实施路径欢迎读者与编写组联系探讨，联系人：白玉真（18810275013）。

（一） 数据集建设

企业以场景为牵引，梳理行业数据资源，以形成一批产业亟需、规模庞大、模态丰富、质量过硬、具备行业特色的数据集为目标。通过提升建设运营能力，构建一系列技术工具和运营管理措施，全面支撑开发运营全流程和数据采集、处理、标注、评估全环节。在数据集建设方面，主要包括需求管理、数据采集、数据处理、数据标注、质量管理和数据交付六部分能力。

1. 需求管理

当前，高质量数据集建设普遍以核心业务场景为牵引，通过规划业务目标、任务类型、样本形态和预期效果，进而明确所需的数据源、模态和规模。**从实践情况来看，数据需求正由以部门单点提出为主，逐步转向结合企业智能化应用规划进行集中判断和统筹安排。**部分央企围绕生产运行、安全保障、经营管理等重点方向，提前梳理可落地的应用场景，并将其中具备共性价值的应用进一步拆解为数据集建设任务，在启动阶段就明确服务对象和使用方式。在实际推进中，明确的数据需求有助于减少后续采集和标注环节中对数据源的调整，可有效提高数据集与模型训练、业务验证之间的匹配度。

2. 数据采集

围绕目标场景，通过梳理业务流程和样本覆盖度等质量需求，明确哪些数据需要采集和补充，进而获取分散在不同系统、不同环节、不同载体中的数据。**从实践情况来看，央企数据采集来源较为广泛。**既包括业务系统内的各类模态的数据资源，也包括采买的数据产品、公开网站或开源社区的数据集、合成数据等。例如，某交通物流行业央企，汇聚内部图纸、监控平台、政府网站等多来源的数据进行数据集建设；某能源行业央企加快行业数据资源整合，持续推进统建、自建等数据的入湖，推动数据在内外部的开放共享。总体来看，央企数据采集逐步形成“面向场景取数”“内部资源汇聚”“多源多模态”等特点，但在跨系统协同、数据连续性和采集成本控制等方面仍面临

一定挑战。

3. 数据处理

数据处理是连接原始数据与可用样本的关键环节，其成熟度直接影响数据集的可用性和稳定性。**从实践情况来看，在工具平台方面，部分央企开始优化技术架构，打通建设全流程。**针对结构化、时序、图像等不同类型的数据，研发覆盖全流程的工具链，开发格式转换、特征提取、数据融合、增强合成等工具算子。在处理流程方面，通过算子化、流水线化的方式，将常用处理规则固化为可复用的处理组件，并设置处理流程，通过清洗、规范化、切分和重组等操作，将原始数据转化为适合训练的样本形态，结合可视化、任务编排等能力增强任务配置的灵活性和多项目间的一致性。整体来看，央企在数据处理环节正逐步优化平台架构、完善工具体系，但在多模态协同处理、复杂场景适配和处理结果可解释性等方面仍需持续完善。

4. 数据标注

数据应用场景专业，规则复杂，需要通过数据标注将业务规则和专业经验转化为模型可学习知识，标注的质量直接影响数据集质量。**从实践情况来看，数据标注逐步形成业务规则、技术工具与人工校验相结合的模式。**围绕具体任务，企业通常先对样本单元、标签层级和判定口径进行统一约定，再通过规则化标注、模板化标注等方式完成基础标注工作，在此基础上引入自动标注或智能辅助工具，提高大规

模样本处理效率。例如，某通信行业央企，通过智能预标注、多模态标注和人机协同模式，大幅提升标注效率与准确性。专家标注则更加侧重于复杂、有风险、重要的任务，如边界样本判定、标注所用模型缺少样本数据等。例如，某工业行业央企，针对某冶炼数据集采用专家主导结合人工智能辅助校验方式进行标注，准确率和效率较之前有极大提升。整体来看，央企数据标注在通用场景以自动化标注结合人工审核为主，复杂场景以专家标注结合智能化判断为主，后续需要持续沉淀专业知识，增强跨项目复用能力。

5．质量管理

质量管理是指围绕数据集全生命周期，对数据采集、处理、标注、集成和应用等各环节建立统一的质量目标、评价标准和管控机制，确保数据集在规模化建设和持续迭代过程中保证可用性的管理活动。**从实践情况来看，质量管理不再只是建设完成后的检查环节，而是逐步前移并融入数据采集、处理和标注等各个阶段。**一是在评估维度方面，央企对数据质量的关注已从基础的完整性、准确性、一致性等扩展到更贴近任务和场景的维度，如全面性、均衡性、应用效果等。二是在评估方式上，为支撑数据集规模化建设，一些央企开始通过规则校验、自动检测和抽样复核等方式，对关键质量问题进行提前识别和集中管控，减少问题在后续环节的放大。同时，质量结果逐步被用于反向校正采集策略、处理规则和标注口径，使质量管理工作在实践中形成闭环。整体来看，央企在质量管理方面正从“事后把关”走向“过程约

束”，但在多模态数据的统一评估、质量指标与应用效果之间的关联分析等方面仍有提升空间。

6. 数据交付

数据交付是确保数据集能够被不同使用主体稳定、可信地使用的关键环节。**从实践情况来看，当前数据交付仍以企业内部使用为主。**围绕模型训练、分析验证和业务应用等场景，对数据集进行结构整理、格式规范和必要说明，使其能够被相关系统和团队直接调用。在此基础上，随着可信数据空间、数据元件、数联网等数据基础设施建设加速推进，**部分央企已开始探索面向外部协作的交付方式**，在科研合作、行业共建等场景中，通过限定数据范围、版本和用途，实现数据集的有序共享，或者上架数据产品进行数据交易。例如，某能源行业央业在交易所挂牌设备故障诊断数据集，该数据集通过对光纤振动异常信号的智能分析，可显著提高风险事件类型预警效率及准确率。但整体来看，大多央企的企业、行业可信数据空间仍在建设过程，数据交付与应用还是以内部为主，随着未来制度完善、技术成熟会逐渐延伸到外部流通使用。

（二） 数据集运营

1. 应用服务

在数据集运营阶段，应用服务承担着将数据集持续转化为业务价值的重要作用，其核心在于数据集是否能够被稳定应用于模型训练、

分析验证和业务流程中，以及对应用价值、实现可行性和扩展性的分析。**从实践情况来看，高质量数据集的应用服务正逐步从零散调用走向相对体系化的服务体系。**一些企业开始围绕模型研发、算法验证和业务试点场景，为数据集提供统一的访问接口、调用规范和使用说明，使不同团队能够在既定规则下多次使用同一数据集资源，从而降低数据集使用门槛，一定程度上也可以避免数据集的重复建设。随着央企内应用场景不断扩展，数据集服务将呈现出“按场景供给”的特征，即围绕特定业务提供对应版本的数据集，精准支撑模型迭代和应用优化。整体来看，央企在数据集应用服务方面仍处于逐步完善阶段，通过将数据集从一次性交付成果转变为可持续提供服务的资源，数据集的应用价值将不断放大。

2. 运营监控

在数据集进入持续使用阶段后，运营监控主要用于保障数据集长期可用。**从实践情况来看，运营监控指标体系较为缺失，当前以资源监控、任务状态监控为重点，还未扩展到数据集使用效果、质量变化和潜在风险的综合感知。**为实现数据集的可持续使用，一方面，需要对数据集的使用情况进行持续跟踪，包括被哪些模型或业务场景调用、使用频率如何、是否出现异常中断等，从而判断数据集在实际应用中的稳定性和活跃度。另一方面，质量相关指标也应逐步纳入监控范围，通过对样本分布、标签一致性、异常比例等变化的持续观察，及时发现数据老化或偏移问题。在此基础上，企业可进一步将运营监控结果

与模型效果、业务反馈进行关联分析，为数据集是否需要更新、扩展或重构提供依据。整体来看，央企在运营监控方面仍处于探索阶段，但正以实现“是否存在风险、能否持续创造价值”的综合判断为目标，构建全面的运营监控指标体系，增强运营监控能力。

3. 生态运营

生态运营是指围绕数据集的共建、共用和持续迭代，组织和协调多方主体形成稳定的协作关系，并通过明确运行规则和运行机制，推动数据资源和数据集在更大范围内流通、复用和价值释放。围绕数据集建设和应用构建稳定、可扩展的运营生态逐渐成为央企的共同选择。**从实践情况来看，央企正依托自身在行业中的组织和资源优势，将数据集运营从内部团队拓展至子公司、产业链上下游、科研机构、专业服务方等主体。**各方围绕典型场景开展共建、共用和协同优化，通过明确参与边界、使用规则和协作方式，实现数据集有序流通和价值放大。例如，某能源行业央企通过构建行业生态，按照约定规则与其他能源企业进行数据互通，有效提升模型准确性，赋能业务。同时，部分央企开始借助行业社区、联盟平台或既有协作机制，沉淀工具、规则、样本和开发经验，如在焕新社区发布大模型、数据集等。整体来看，生态构建能力正在成为决定数据集能否持续迭代、规模扩展并形成行业影响力的重要因素，央企的数据集生态运营仍以内部协同为主，但向外延展的趋势已经显现。

（三） 基础保障体系

1. 组织管理

组织管理是保障各项工作能够持续推进的重要基础。**从实践情况来看，高质量数据集建设已不再局限于单个团队，而是逐步向集团统筹、分级协同的组织形态演进。**一是组织架构方面，部分央企在集团层面明确数据集建设与运营的统筹责任，通过设立专门的协调机制或牵头部门，对建设方向、重点任务和资源配置进行统一安排，避免各子企业重复建设。由此，将数据集建设工作嵌入业务条线或专业单位，由熟悉业务场景和数据特性的团队承担，形成“**集团统筹、企业建设**”的协同格局。例如，某交通行业央企构建集团和二级单位的协同组织，成立数据治理和人工智能工作组，形成专业的统筹、业务、数字化、人工智能团队。二是人才培养方面，央企对人员能力结构的要求也在发生变化，既需要具备行业知识和场景理解的业务专家，也需要掌握数据工程工具、标注方法和质量控制手段的专业人员。一些央企通过人才引进、内部培养、项目轮岗和外部合作等方式，逐步积累稳定的数据集建设与运营队伍，使关键经验能够在组织内部沉淀和传承。总体来看，央企在组织管理上正逐步摆脱以项目为中心的松散模式，转向以长期能力建设为导向的组织安排。但在跨单位协作效率、职责边界清晰度以及与模型研发、业务应用团队的协同机制等方面，仍需要在实践中不断优化。

2. 制度规范

制度规范是支撑数据集在建设运营过程中，长期、稳定、合规运行的基础性要素，其核心在于通过明确规则、流程和责任边界，将数据集建设从项目行为转化为可持续的组织行为。**从实践情况来看，部分央企开始搭建制度框架，逐步探索建立建设职责、流程衔接、质量要求、使用边界等制度和机制，增强数据集建设安全合规和可复用性。**在建设阶段，通过明确数据采集、处理、标注和质量评估等环节的步骤和规则，提高单位和团队的执行效率；在运营阶段，通过对数据集使用范围、版本管理和迭代方式的制度化约束，保障数据集在长期运行中的稳定性和一致性。例如，某交通行业央企设计了完整的管理体系，明确数据采集、处理、标注、评估和发布等阶段的工作步骤、方法和技术要求等。同时，一些央企开始将数据集相关制度与既有的数据治理、信息安全和合规管理体系衔接，避免制度割裂带来的执行冲突。总体来看，央企已开始梳理和设计数据集相关的制度规范体系，正加速构建关键性的制度、机制和标准。

3. 资源管理

算力资源和存储资源是支撑数据集研发、管理和使用的关键基础要素。**从实践情况来看，数据集建设对资源的需求呈现出明显的阶段性和结构性特征。**一方面，数据处理、标注、质量评估等环节对算力提出持续需求，既包括集中式计算任务，也包括面向多项目并行的弹性计算需求，并且随着标注和评估环节的智能化能力引入，会进一步

拉高算力需求。另一方面，多模态数据集规模不断扩大，对存储资源的容量、性能和管理能力提出更高要求，尤其是在图像、视频、时序数据等场景下，长期保存、版本管理和高频访问并存。**当前，部分央企在算力和存储资源配置上仍以业务系统为主，数据集建设与运营过程存在资源分散、调度不灵活、建设与使用脱节等问题。**随着高质量数据集逐步从项目成果走向长期运营对象，围绕算力和存储资源形成更加清晰、可统筹的管理方式，已成为央企保障数据集建设和运营稳定开展的重要基础条件。

4. 安全合规

安全合规是贯穿数据集建设全生命周期的基础保障，不仅涉及各平台运行环境的系统安全、通信安全、存储安全等，也包含个人信息保护、商业秘密和国家秘密等数据的采集、使用、监管合规性保证。**从实践情况来看，央企的数据安全工作通常围绕数据分类分级、敏感信息识别与处置、访问控制与权限管理、使用留痕与审计追溯、风险评估与持续整改等关键要素展开，以确保数据在采集、汇聚、加工处理、存储管理、共享使用等环节均可控。**对央企的高质量数据集而言，安全合规要求往往面向多重约束，既要满足通用的数据安全与个人信息保护规则，也要符合行业监管要求和央企内部保密管理体系，并与业务场景中的人员、系统和流程保持一致。总体来看，安全合规能力直接影响数据集可用范围与可持续运营空间，决定了哪些数据能够纳入建设范围、以何种形态组织加工、以何种条件参与协同应用，是数

据集长期稳定运行的重要基础。

综上所述，央企高质量数据集建设应以国家战略和行业共性需求为牵引，依托其在关键领域、核心场景和长期业务运行中形成的稳定数据资源基础，将数据集建设嵌入生产运行和管理流程之中。首先，通过**“场景牵引、工程化推进、体系化治理、协同化扩展”**，充分发挥央企在资源整合、标准牵头和生态组织方面的优势，通过集团统筹、分批建设的方式推进数据集规模化构建。然后，夯实技术支撑能力，构建覆盖采集、处理、标注、质量管控和交付的全流程工具链，提升数据集建设的工程化和自动化水平。其次，加速完善数据集相关的内部制度机制，构建安全合规体系，支撑数据集高效、合规建设与应用。最后，深化标准研制与落地实施，持续推动数据集在央企内部持续使用和迭代，并逐步向行业协作和生态共建拓展。

四、央企高质量数据集建设案例

为系统呈现中央企业高质量数据集的建设经验，本章节聚焦智慧能源、工业制造、绿色低碳、交通物流、医疗卫生、现代农业、移动通信和应急管理这八个重点行业，选取具有代表性的数据集案例。通过对各案例的建设背景、建设方案和应用成效三个维度进行分析，进一步总结央企高质量数据集建设的特点与经验。

表 2 高质量数据集建设案例一览表（排名不分先后）

序号	所属行业	单位名称	数据集名称
1	智慧能源	中国石油天然气集团有限公司	中国石油油气地震勘探大模型高质量数据集
2		国家石油天然气管网集团有限公司	油气管道保护高质量数据集
3		中国南方电网有限责任公司	配电网智能规划多模态数据集
4	工业制造	中国第一汽车集团有限公司	乘用车产品智能数据集
5		中国铝业集团有限公司	铝合金材料金相组织图片数据集
6	绿色低碳	中节能大数据有限公司	企业远程非现场执法数据集
7	交通物流	中国交通信息科技集团有限公司	交通基础设施多模态三维构件数据集
8	医疗卫生	中国联合网络通信集团有限公司	肺结核影像精标注高质量数据集
9	现代农业	国机数字科技有限公司	农业机械及作业高质量数据集
10	移动通信	中国移动通信集团有限公司	中国移动人时空三元组高质量数据集
11		中国电信集团有限公司	通信行业网络大模型高质量数据集
12		中国联合网络通信集团有限公司	信息通信领域高质量数据集
13	应急管理	新兴际华集团有限公司	大型石油储罐火灾救援数据集

（一）智慧能源

在全球能源转型与“双碳”目标驱动下，智慧能源行业正以数据

资产化为核心目标，构建支撑能源系统智能化升级的高质量数据体系。当前，行业围绕能源生产、传输、存储、消费全链条，已整合石油气、管道、电力等多细分领域数据，通过标准化采集、实时化处理与场景化应用，形成覆盖源网荷储互动、多能协同互补、用户需求响应等场景的数据集体系。

下面以中国石油天然气集团有限公司的中国石油油气地震勘探大模型高质量数据集、国家石油天然气管网集团有限公司的油气管道保护高质量数据集、中国南方电网有限责任公司的配电网智能规划多模态数据集为例进行介绍。

专栏 1：中国石油油气地震勘探大模型高质量数据集

案例背景：

地震勘探是油气勘探最常用、最有效的技术之一。地震勘探是给地球做“CT”，是油气勘探的侦察兵，是后期开发的参谋长，是打赢油气增储上产攻坚战的关键。

近年来，国内油气勘探开发取得显著成效，但也面临油气资源品质持续下降、勘探开发难度增大等挑战。传统油气地震勘探处理解释方法已难以适应复杂勘探需要和海量数据的快速分析，亟须加强新一代人工智能特别是大模型技术的应用，提高地震处理解释效率与精度。

建设方案：

本案例遵循了《石油地震勘探资料采集技术规范》（GB/T 33583-2025）《石油地质与地球物理图形数据格式规范》（SY/T 6932-2024）等标准，实现基于勘探地震数据的高效处理解释及预测，为勘探领域提供高效的、可工业化推广的智能化解决方案。

1. 数据需求及采集

聚焦物探、测井和地质三大专业，采集国内主要含油气盆地高分辨率地震勘探数据、测井数据，相关地质资料及研究报告，明确数据需求包括时空序列、文本及图像等数据类型，构建多模态勘探数据集，满足大模型训练需求，建立地震初至、断层、地震相、地震属性等样本标签库，重点解决断层识别、河道预测等典型应用场景的数据适配性问题。

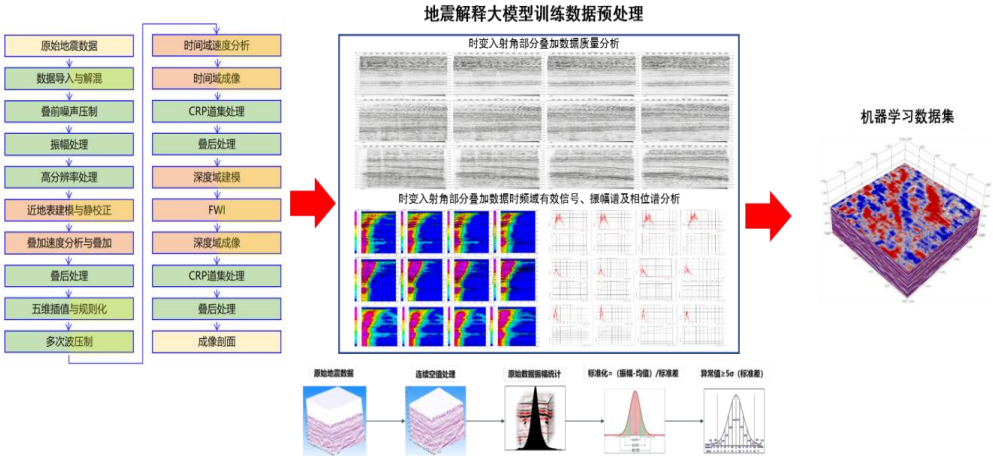
数据采用统一数据接口规范实现多源数据标准化接入。针对地震数据易受噪声干扰的特点，引入多通道同步采集技术，剔除了信噪比过低的无效数据，确保原始数据质量基本达标。

2. 数据处理与管理

构建数据治理与质控体系，按照统一标准开展数据治理、数据清洗和数据处理等工作。搭建分布式存储云，构建统一数据共享服务平台，实现多源

异构数据统一管理，支持动态扩展与安全防护；为模型训练、模型微调与数据共享提供数据服务，支持油气勘探大模型全生命周期应用。

针对地震解释大模型对数据质量、数据格式、数据对齐、数据增广、数据高效存储等方面的需求，开展了地震、测井、地质等核心数据的数据治理、数据处理与数据集构建技术研究，构建面向大模型的地震资料处理流程及方法，有效支撑地震解释大模型数据集构建，实现多尺度、多维度数据的高质量协同训练与应用。



图：油气勘探地震数据处理流程

3. 数据标注与评估

地震数据标注的工作量巨大，同时专业数据的标注需要很强的专家经验。研发了数据智能标注工具平台，基于预置的领域模型，针对地震、测井等数据进行智能标注，并通过专家审核，保证训练样本的质量，大幅提升了工作效率。

构建了典型盆地标准样本库：分别针对塔里木、鄂尔多斯、四川等重点含油气盆地，围绕主流地质类型和油藏类型，建立了断层、地震相、地震属性等样本标签库；建立地质分层、岩性判识、储层划分等样本标签库。

从准确性、一致性、完整性等维度开展，通过人工核对、模型交叉验证及标签覆盖热力图排查未标注区域等方式进行数据评估，确保数据质量可靠。

4. 模型验证

支撑了地震解释大模型的构建，开发了基于大模型的走滑断裂识别、缝洞体预测、生物礁滩相识别、河道识别、火山岩识别等应用场景，对比传统方法，大模型预测结果与实钻数据吻合度有显著提升。

应用成效：

本案例数据集及支撑构建的大模型，已在长庆、塔里木等 20 余个国内外油气田部署应用，取得显著成效。

在某地区碳酸盐岩缝洞体预测场景中，深层、超深层地震勘探面临有效信号衰减严重、信噪比低等挑战，传统方法刻画的缝洞体不清晰、识别准确度难以突破 80%；应用大模型技术预测的缝洞特征更加明显，断裂清晰，缝洞体沿主要断裂分布，地质规律性更强。经过钻井验证，符合率达到 90%以上，一定程度降低了钻井投资风险。

在某地区致密砂岩岩性预测场景中，因储层薄、非均质性强，储层预测准确率一直不高。传统方法解释的结果河道特征不清晰，信噪比低，利用地

震大模型进行重新解释，准确刻画河道砂体的展布特征，在没有进行井数据标定条件下，大模型预测准确度比常规方法高出 10%以上。

高质量数据集面向大模型应用需求，通过海量数据驱动，具有数据质量高，多模态融合，应用范围广等特点，减少了对数据标注的依赖，改变了人工智能碎片化应用范式，显著降低了人工智能应用门槛，加速了人工智能在油气行业的落地应用。传统地震解释方法围绕庞大数据体和各类复杂地质资料，从属性分析、地震反演到目标评价，研究工序多、周期长，利用数据集与大模型技术通过少量标注和微调直接产生预测结果，地震解释效率有较大提升。

专栏 2：油气管道保护高质量数据集

案例背景：

油气管道作为能源运输大动脉，其安全稳定运行直接关系到国家能源安全和经济发展。为有效管控风险，企业在管道沿线部署大量传感设备，常态化采集分析管道本体、附属设施、周边环境多源海量数据，随着管网规模的不断扩大，亟需构建行业高质量数据集，提升垂直领域 AI 算法模型性能，强化风险管控技术能力。本案例通过构建“采集—加工—应用”数据运营机制，面向油气管道保护核心业务场景，开发了 5 个高质量数据集，有效将多源异构海量数据资源转化为赋能油气管道保护的数据资产，实现油气管网安全风险管控从“经验驱动”向“数据驱动”转变。

建设方案：

聚焦油气管道风险管控，通过建立“采集—加工—应用”数据运营机制，开发了油气管道保护高质量数据集。

1. 场景选取与需求明确

选取无人机巡检、高后果区视频监控、分布式光纤振动监测、地质灾害监测、阴极保护数据分析 5 个典型应用场景，相关数据集用于对 11 个算法

模型进行训练以提升精度，支撑管道沿线威胁事件智能感知、管道本体缺陷精准判识等智能化场景，赋能油气管道保护 5 项核心业务。

2. 数据采集

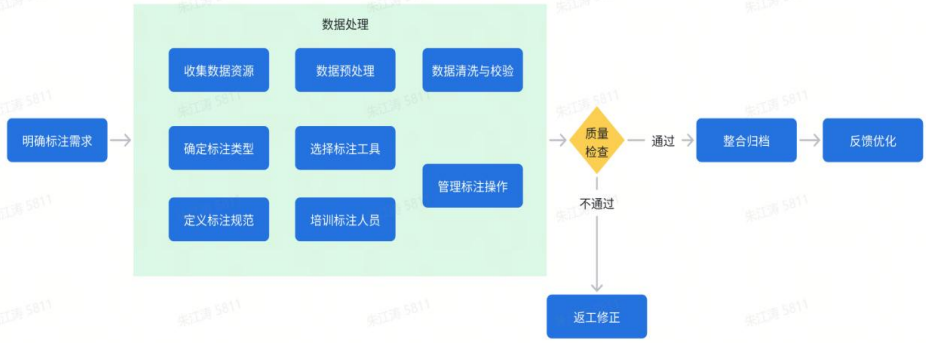
构建标准化、自动化、可审计的多源数据采集体系。卫星数据通过国家遥感中心 API 定期获取；无人机巡检采集的影像数据回传归档，形成历史图像库；地面传感类时序数据通过时序数据中台统一汇聚。项目制定统一接入规范：视频分辨率不低于 1920×1080、帧率≥25fps；时序数据采用 JSON 格式，并强制携带 UTC 时间戳与设备唯一标识；所有数据同步生成包含采集位置、光照条件、设备状态等信息的元数据包。经试点优化，无人机有效图像占比提升至 95%以上，光纤振动信号信噪比显著改善，切实保障源头数据“采得全、采得准、采得稳”。

3. 数据处理

针对多源异构数据特点，建立系统化预处理流水线，将“生数据”转化为“熟数据”。统一数据格式：视频转 MP4、图像转 JPEG/PNG、时序数据归一化为 Parquet 存储。实施多级清洗，剔除重复帧、过滤异常传感器读数、修复时间戳错位。引入数据增强与特征工程：对工程设备图像进行旋转、亮度扰动等扩增，提升复杂场景鲁棒性；从原始信号中提取地表形变特征、施工目标语义特征、腐蚀风险时序模式等高阶特征。处理后数据按“威胁类型—设备类别—地理分区”三级标签自动归档，并建立版本快照机制，确保过程可复现、可追溯。

4. 数据标注及数据质检

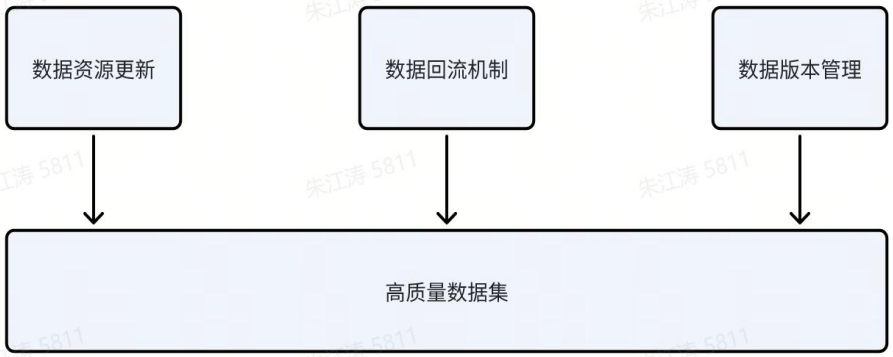
构建“规范先行、工具赋能、闭环优化”的协同标注体系。由管道安全专家牵头，针对漏油、挖掘机作业、人员入侵等 10 类场景制定精细化标注规范，采取矩形框、点线标注工具，明确边界框精度（IoU ≥0.85）与遮挡处理规则。开发智能标注平台，集成预训练模型实现 AI 初筛，人工校验修正，标注效率提升 40%以上。实行“双人标注+专家抽检”机制，关键场景标注一致率要求⩾95%。建立动态反馈闭环：将模型误检、漏检样本自动回流，定向补充难例，并定期优化标签体系，实现数据集“越用越准”。



5. 模型验证

以模型性能反向验证数据集质量。基于 5 个高质量数据集训练 11 个专用 AI 模型，在独立测试集上达成第三方施工识别准确率 92.5%、管道缺陷定位 IoU 0.88、地质灾害预警提前量≥72 小时等指标。若模型性能未达标，则触发上游环节复盘，针对性补充数据后重新验证。最终交付结构化数据集（含原始数据、标注文件、元数据及质量报告），并通过集团数据平台支持

API 调用与批量下载。模型验证与数据更新、规则更新形成闭环，及时修正偏差、完善标注规则，提升标注一致性与场景适配性，保障数据集的稳定产出。



应用成效:

油气管道保护高质量数据集的建设和应用，取得了显著的经济效益和社会效益，为油气管道安全稳定运行提供了有力保障。治理后的高质量数据集能够有效驱动管道保护相关业务管理提效与变革，提升油气管道本质安全水平。

经济效益方面，数据集的应用有效降低了管道运营成本，提高了管网运行效率。通过分析无人机采集的视频影像数据进行巡检，每年节约人力成本 737 万元；数据集的开发应用有助于地质灾害隐患早期发现和预警，每年减少隐患治理成本约 320 万元；基于数据集对管道本体缺陷进行精准评估，年均节约维修费用约 300 万元。综合计算每年节约成本约 1357 万元。同时在提高管网运行效率、降低事故事件损失等方面还能够产生数以亿计的间接经济效益。

社会效益方面，数据集的建设和应用，有力支撑了地貌变化检测、威胁事件检测、多目标跟踪定位、阴极保护有效性评价等系列算法模型开发训练，提升了管道周边和本体威胁事件综合感知、识别和联动响应效率，保障了油气管道的安全稳定运行。在数据采集、治理、应用等方面的创新实践，有效推动了管道安全领域技术进步，为行业高质量发展提供了新质生产力。数据集+AI 构建的智能化业务场景释放了基层一线 5%的人力资源，实现管理效能提升。

专栏 3：配电网智能规划多模态数据集

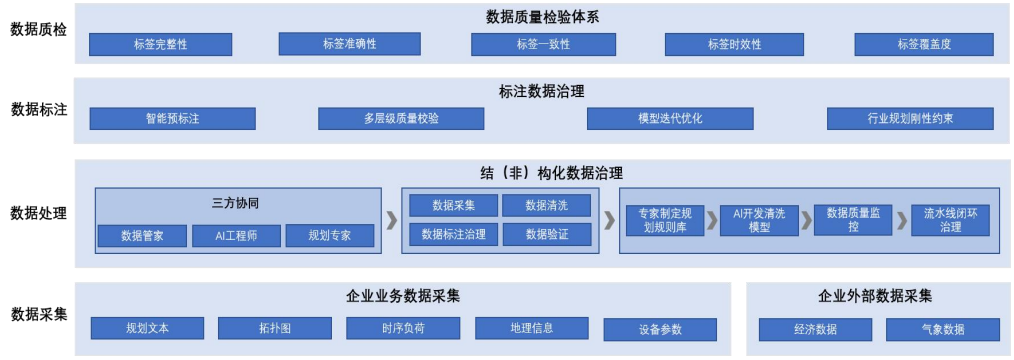
案例背景:

针对配电网规划依赖人工作业、数据壁垒严重、多模态数据协同困难等问题，本案例建立了“标准-技术-流程”三位一体高质量数据集治理体系，整合了规划文本、拓扑图，时序负荷等 7 类异构数据，构建了覆盖配电网规划业务全流程的多模态高质量数据集，有效支撑配电网规划大、小模型研发，显著提升南方五省区电网运行安全性、韧性与智能化水平。

建设方案:

本数据集的建设严格遵循数据驱动的核心理念，其构建过程是一个覆盖

从业务理解到质量验收的全生命周期闭环。其建设流程划分为“场景选取—需求明确—数据采集—数据处理—数据标注—数据质检”六个关键步骤，依次实现业务需求的精准转化、多源异构数据的系统治理，以及数据质量的规范管控，最终形成具备高可用性与可复用性的标准。



图：配电网智能规划多模态数据集建设流程图

1. 场景选取与需求明确

配电网智能规划多模态数据集的构建，首先紧密围绕新型电力系统建设与“双碳”目标下的实际业务需求展开。选取源荷预测、运行推演、问题诊断、规划方案生成及规划报告智能写作等具有代表性、前瞻性和高价值的规划业务核心场景作为数据建设的切入点。

2. 数据采集

在数据维度上，全面覆盖了规划文本、电网拓扑、时序数据、设备台账、地理信息、运行状态与外部环境数据等七类多模态信息，以确保数据生态的完整性与多样性。

3. 数据治理

围绕配电网规划业务特性，建立以“专家规则-智能处理-质量监控”为核心的数据治理体系。该体系致力于打造一个高效的配电网规划基础数据分析应用平台，通过引入行业专家经验形成规则库，结合自然语言处理、图像识别与时间序列分析等智能算法对多源异构数据进行深度融合与清洗，并建立全流程质量监控节点，从而确保所有数据达到高效、规范与安全可用的标准。

4. 数据标注

为确保用于模型训练与分析的样本数据具备高度的准确性与业务符合性，构建一套人机协同、持续优化的数据标注闭环流程。该流程以行业规范与专家知识为根本约束，深度融合智能预标注与人工专家智慧，形成了从“标注任务定义->智能模型预标注->专家审核与修正->标注结果反馈学习->模型优化迭代”的完整闭环，从而在提升标注效率的同时，极大地保障了标注结果的专业性与一致性。

5. 数据质检

最终，为确保产出数据能够真实、可靠地服务于配电网规划模型训练与深度业务分析，依托电力行业高质量数据集评估要求，建立一套量化的数据质量检验体系。该体系围绕标签完整性、标签准确性、标签一致性、标签时效性以及标签覆盖度五个关键指标对数据集进行严格把控，只有全面通过检验的数据才能被正式入库发布，从而为后续的智能应用提供坚实可靠的数据基石。

应用成效:

本数据集在配电网规划领域的深度应用，已转化为显著的业务赋能与综合效益。通过系统梳理规划全流程需求，成功整合了来自不同省份与业务系统的多源异构数据，构建起覆盖规划文本、拓扑图纸等七大类数据的标准化资源池，形成了配电网规划数据集建设的示范样板。

依托多模态高质量数据集，负荷预测、网架问题诊断及规划方案自动生成等关键模型研发取得突破性进展，成功将传统需耗时数周的规划方案生成时间压缩至十分钟以内，整体规划周期缩短 95%。经试点应用验证，年均节约人工成本约 3312 万元，人工成本降低 70%；同时，通过模型精准诊断并前瞻性消除潜在运行隐患 1435 项，有效避免了因大规模停电可能造成的巨额经济损失，预估达 14.35 亿元。

在创造直接经济价值的同时，本案例应用也产生了深远的社会效益。它显著增强了南方五省区电网的安全性与韧性，为超过 1.13 亿用户提供了更可靠的电力供应保障，并将极端灾害下的应急响应时间缩短至小时级别。在推动绿色转型方面，应用成果每年约助力减少碳排放 4.8 万吨。所形成的一套可复制、可推广的数据治理与智能化应用范式，已成功在南方电网多省公司及其他新能源企业中得到推广，赢得了行业与社会的广泛认可，为电力行业数字化转型树立了创新标杆。

（二）工业制造

工业制造包括汽车制造、有色金属、装备制造、船舶工业等多个细分领域。汽车制造领域围绕工业质检、智能驾驶等场景，产生雷达点云、车辆运行和车内音视频等数据，通过多源数据汇聚、预处理、标注、合成增强等方式形成高质量数据集。有色金属领域则主要面向金属生产冶炼，在工艺优化、生产监测和工业质检等方面构建高质量数据集，实现业务分析和效率的双重提升。下面以中国第一汽车集团有限公司的乘用车产品智能数据集、中国铝业集团有限公司的铝合金材料金相组织图片数据集为例进行介绍。

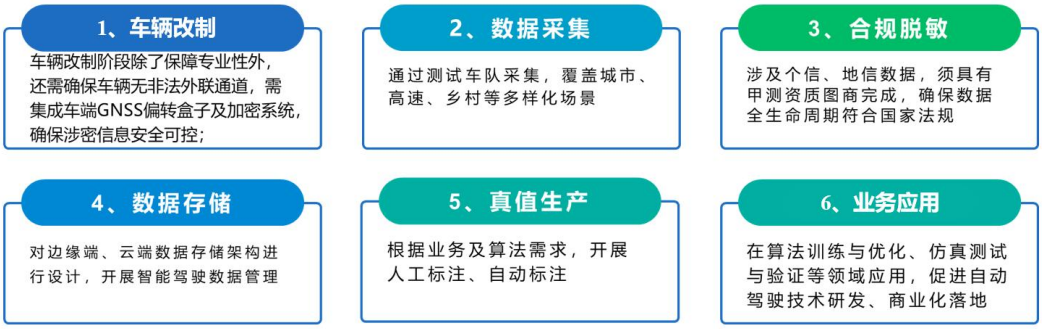
专栏 4：乘用车产品智能数据集

案例背景：

为解决自动驾驶技术研发中的关键瓶颈，包括环境感知、决策规划和控制执行的算法训练与验证问题。通过采集海量真实道路数据，覆盖复杂场景（如极端天气、突发障碍等），构建多模态数据集（摄像头、激光雷达、毫米波雷达、高精地图等），训练模型，降低长尾场景误判率，优化自动驾驶系统的性能，建立数据闭环，加速智能驾驶软件落地应用，并满足国家法律法规对智能驾驶数据合规要求。

建设方案：

智能驾驶技术的快速发展依赖于高质量的数据集，用于训练和验证感知、决策、控制等算法。本方案旨在建设一套完整的智能驾驶数据集，包含系统化的流程，涵盖车辆改制、数据采集、数据脱敏、数据存储、真值生产、业务应用六大核心环节。



1. 车辆改制

作为建设高质量数据集的起点，需将量产车或测试车改造为适合数据采集的智能车辆，确保传感器布局合理、数据同步精准、系统稳定可靠。在具体建设中结合各业务专业及算法需求，确定传感器选型与布局。此外，对车辆改制还需考虑搭载高性能车载计算单元，支持多传感器数据实时融合，采用 PTP（精密时间协议）或 GPS 授时，确保传感器数据时间对齐。为减少传感器标定误差，还需定期进行联合标定。

2. 数据采集

主要内容包括：设计数据采集方案，覆盖多样化驾驶场景（城市、高速、乡村、极端天气等），采集高质量、多模态的原始数据。在场景设计中，除考虑车道保持、跟车、红绿灯等常规场景外，还需重点关注行人突然闯入、车辆违规变道、恶劣天气等边缘场景用例，同时，还可利用仿真工具（如：CARLA、LGSVL）生成罕见场景数据，已实现对场景补充。

3. 合规脱敏

采用国家认可单位的空间位置处理算法和插件，对坐标数据提供各类合规偏转服务，支持对车端采集上传数据的解密功能，支持对空间位置处理插件的应用管理与日志审计，日志数据支持落盘导出带回的模式等。基于采集车数据采集类型进行分类分级，区分测绘数据与非测绘数据进行合规处理。实现 GNSS 数据脱敏、涉密数据脱敏、隐私数据脱敏。委托具有导航电子地图甲级测绘资质的合规服务商进行数据合规管理，确保数据的合规安全，定

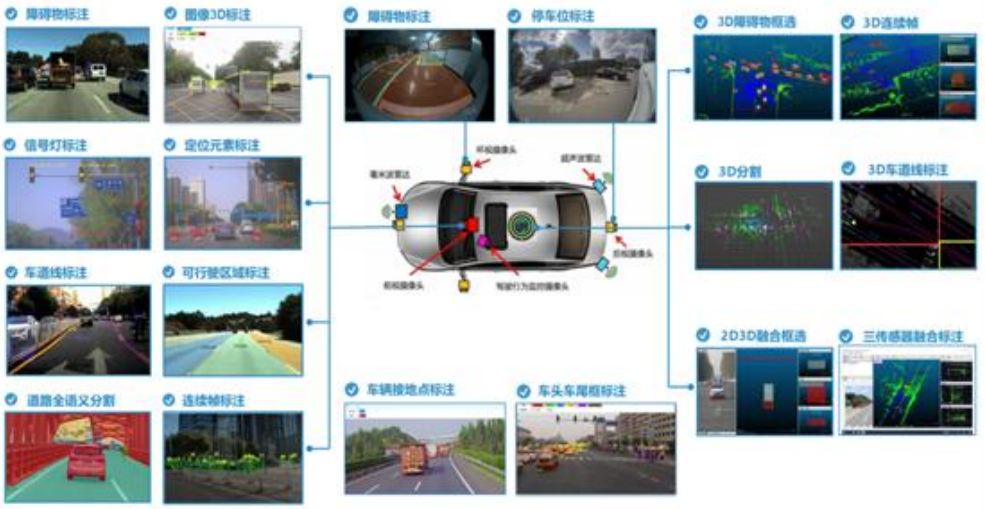
期进行第三方隐私审计。

4. 数据存储

本方案依托图商专属合规云实现数据存储，满足数据审核及监管要求。同时面向未来 3 年业务需求，制定存储资源规划。存储架构设计分别在边缘端和云端进行，其中车端考虑车载 SSD 临时存储（TB 级），支持断点续传。云端采用对象存储服务、分布式文件系统（HDFS）、 数据库（MongoDB）。为实现数据管理，对元数据进行场景类型、天气、路况标注。同时，实现数据集管理迭代及版本控制。

5. 真值生产

本方案真值生产由人工标注和自动标注两部分组成。满足红旗智能驾驶 HAD、VDC、L4 三大平台的开发。其中，人工标注实现 2D-3D 障碍物、3D 固态障碍物、3D 道路语义、2D 车道线/道路边沿、2D 道路语义、2D 可行驶区域真值生产。通过离线大模型预标注真值，再人工 100%审核修正，大幅降低标注成本，接近人工标注准确率。可实现 3D 障碍物检测/3D 语义分割，对激光点云数据中的所有物体进行语义分割，可识别车辆、行人、骑行者、路面、人行道、建筑物、栅栏、树木、杆状物、交通标志等多种目标物。



6. 业务应用

智能驾驶数据集是车企从研发到量产的核心竞争力，基于本方案构建的高质量数据集，在自动驾驶技术研发和商业化落地中发挥核心作用，主要业务应用包括：算法训练与优化、仿真测试与验证和法规认证与安全评估。

应用成效:

1. 算法性能提升

本方案的数据多样性使目标检测模型在复杂城市场景的准确率提升 20%-30%。多模态融合技术将恶劣天气下的误检率降低至 5%以下（对比单一传感器）。

2. 智驾产品商业化落地

本方案已应用至红旗 E001、E202、C801 量产车型的智驾产品。覆盖全国 30 多个城市的路网数据助力快速适配新城市。红旗量产智驾数据应用于事故责任判定、保险理赔等，助力实现问题场景快速复现、定位，提升问题解决效率。

3.智能驾驶研发

数据集应用于 L2-L3 视觉障碍物、车道线、标志牌等识别模型；L3 激光雷达障碍物感知模型；L4 激光雷达模型；端到端大模型。依托合规云开展数据存储、管理、处理直至算法训练，对数据进行分类分级、上云管理，旨在提升数据复用率，其中 70%标注数据可用于多车型开发。

专栏 5：铝合金材料金相组织图片数据集

案例背景：

微观组织是决定金属材料性能的关键因素，金相分析在铝合金材料研发与质量控制中具有核心地位。然而，传统金相分析方法存在效率低、精度差、数据管理分散等痛点：单张图片人工标注耗时约 3 分钟，分析误差可达 10%～15%，且数据格式不统一、标注规范缺失，制约了 AI 模型开发与行业智能化转型。针对这一现状，中铝材料应用研究院有限公司构建了高质量的铝合金材料金相组织图片数据集，覆盖 2xxx、3xxx、5xxx、6xxx、7xxx 系五大铝合金系，包含 11GB 专家标注金相图片，旨在突破数据瓶颈，支撑高精度 AI 模型训练，赋能材料研发降本增效。

建设方案：

数据集建设遵循标准化流程，涵盖场景选取与需求明确、数据采集、数据处理、数据标注和数据质检五大环节，确保数据高质量与可复用性。

1. 场景选取与需求明确

建设聚焦铝合金金相组织智能识别与定量分析场景，覆盖航空航天、轨道交通等六大领域，针对 5 大铝合金系的化合物、疏松等目标组织。业务需求包括实现自动化识别、精度保证（误差率低于 20%）及结构化输出；数据需求涵盖光学显微镜（OM）和扫描电子显微镜（SEM）图像，要求格式统一、规模达 500 张以上，并关联材料牌号、工艺状态等元数据；标准化需求是确保数据质量和模型性能的关键，涵盖图像采集、处理、特征提取和数据分析全流程。

2. 数据采集

数据来源包括历史检测数据与新增采样，严格遵循 GB/T 3246.1-2024 标准制备试样，使用蔡司 AxioObserver7 显微镜、FEI Apreo 等设备拍摄，确保分辨率不低于 1292×968 像素、照明均匀性误差≤5%。通过剔除比例尺错误、污染等不合格图像，最终有效 OM 图像 766 张、SEM 图像 1099 张，覆盖多种工艺状态，并通过均衡采样保证数据代表性。元数据管理包括采集时间、设备信息、样品详情等，实现全流程可追溯。

3. 数据处理

数据处理分为预处理与增强两阶段。预处理采用自适应滑动窗口切分算法将图像统一为 512×512 像素，避免边缘变形；去噪环节针对 SEM 图像使用非局部均值滤波消除荷电效应伪影，OM 图像采用中值滤波（kernel size=3）；对比度增强通过 CLAHE 算法提升组织边界清晰度，尤其优化 7 系铝合金浅色化合物区分；格式转换将.tif 统一转为.jpg，标注文件适配模型训练需求。

数据增强采用 Mosaic 技术、随机翻转、亮度调整及高斯模糊等方法，将 500 张专家样本扩增至 20000 余张训练数据，有效解决小样本瓶颈，提升模型鲁棒性。

4. 数据标注

标注工具选用 PixelAnnotationTool，支持像素级画笔标注与拓扑嵌套处理。标签规范明确定义：背景（黑色，ID=1）、化合物（绿色，ID=2，包括片状/条状/网状组织）、疏松（红色，ID=3，深色孔洞），并规范阴影去除、晶界标注等特殊情况。标注流程创新采用“人工初标-模型预标-专家矫正”半监督模式：先由专家标注 100 张样本训练预标注模型，再对剩余样本自动预标，最后由专家矫正至精度 95%以上，并通过交叉验证保证一致性。

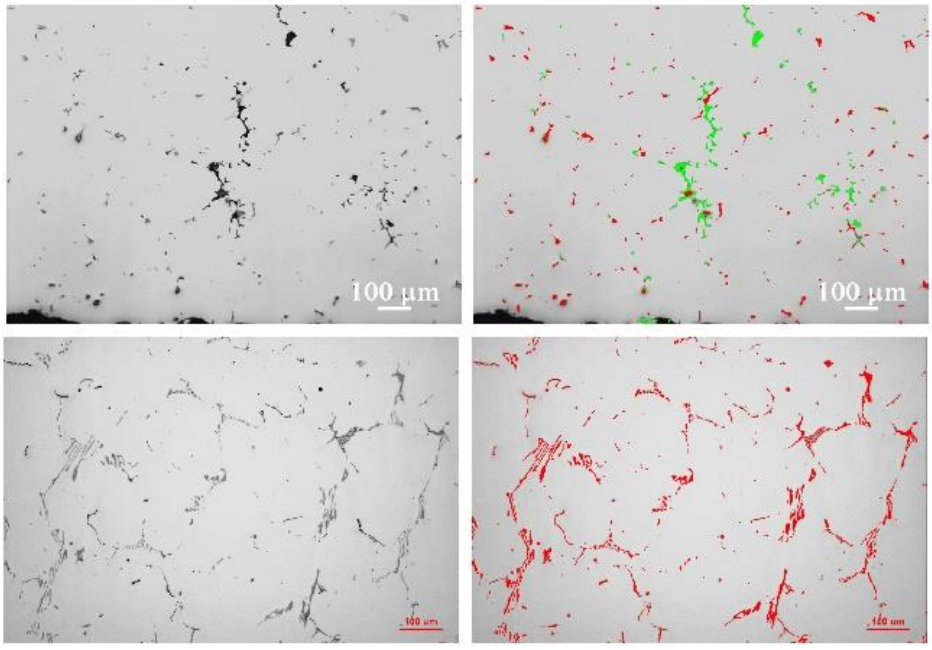
5. 数据质检

质检从准确性、一致性、完整性三维度展开：准确性采用人工像素级核对与工具辅助脚本验证标签匹配度；一致性分“内部一致性”与“外部一致性”两层验证；完整性使用标签覆盖热力图排查未标注区域。全量样本分层质检确保数据质量可靠。

应用成效:

效率提升：铸锭组织分析效率较人工提升 300 余倍，年处理 5 万张图像耗时从 312 人天降至 4 天；

精度提升：分析误差从传统方法的 10%左右降至 1%以内，使不同分析人员或不同批次得到的结果更具可比性；OM 图像整体推理准确率达 99%以上。对于 7xxx 系铝合金化合物与疏松色值相近易混淆的情形，也可以达到很好的效果；



图：OM 疏松与化合物推理效果示例

成本节约:单次分析人力成本从数千元降至 300 元左右,年节省成本 30.3 万元（按 5 万张图像估算）；

资源释放：检测设备占用率降低 30%，释放的设备可承担更多高价值检测任务。

（三）绿色低碳

数据要素是数字化时代最为关键的生产要素，也是当前促进经济

社会发展全面绿色转型的重要驱动力。“十四五”以来，“数字化”和“绿色低碳融合”已成为全球主要经济体的共同关注方向。绿色低碳数据是依托 5G、物联网、云计算、人工智能、区块链等大数据技术，对园区、企业和设施进行全面监测、深度挖掘、精细治理获取的数据，包含生态环境、能源碳排放领域的公共数据和企业数据。下面以中节能大数据有限公司的企业远程非现场执法数据集为例进行介绍。

专栏 6：企业远程非现场执法数据集

案例背景：

为推动环境粗放管理向精准管理转型，破解排污企业数量多、分布广、偷排时段隐蔽带来的执法难题。传统现场执法模式不仅需要投入大量人力、物力和时间，还会对企业形成频繁打扰的问题。中节能大数据有限公司构建了高质量的企业远程非现场执法数据集，通过整合分析环境监测数据、污染源信息、企业工况等多源信息，实现对潜在环境违法行为的智能识别、预警和线索推送，有效降低企业入户执法频率，提升执法效能。

建设方案：

企业远程非现场执法数据集建设过程遵循标准化流程，主要包括场景需求分析、数据采集、数据治理、数据质检等环节，从而确保数据高质量和可复用性。

1. 场景需求分析

针对排污企业远程执法难题，构建企业远程非现场执法高质量数据集，主要应用于企业生产、治理及排放数据异常发现场景，数据需求包括企业排污设施运行数据、污染物排放数据、污染治理数据，通过多元关联数据组合推理排放数据，再与实际监测到的排放数据进行比对，与非对称异常检测计算，以发现在线数据异常的问题。

2. 数据采集

数据来源包括企业端在线监测、大气分表计电、无组织系统、网格化、河流水质、企业自行监测、门禁、第三方检测机构报告及排污许可数据。治理数据取自环保设施运行监控设备，包括废水废气处理设备曝气量、除尘设备压差等运行参数；排放数据通过污染源自动监控系统获取，含废气 SO_2、NO_x、颗粒物及废水 COD、氨氮等污染物浓度、排放量数据。采用 5G + 边缘计算技术实现数据实时传输，确保采集时效性与完整性，所有来源均符合《生态环境监测数据管理办法》要求。

3. 数据治理

数据治理聚焦异常检测核心需求，分三步推进：一是标准化处理，统一不同企业、设备的数据格式与单位，建立生产、治理、排放指标映射词典，实现跨系统数据互通；二是数据融合，通过时空对齐技术将标准数据、设备

数据、生产数据等多维度数据与排放数据关联，构建联动数据链；三是数据标注，采用人工 + 智能辅助的方式构建标注体系。搭建标签库，涵盖设备启停、原料切换等工况标签，以及传感器故障、药剂投放不足、数据篡改等异常类型标签。智能辅助标注依托 “产污 – 治理 – 排放” 逻辑规则，自动识别明显异常数据并预标注，人工团队对预标注结果复核校准。标注信息同步录入数据溯源体系，保障标注结果可追溯、可核验。

4. 数据质检

数据质检围绕准确性、一致性、有效性三大核心指标展开。准确性通过设备校准记录核验实现，对各类监测设备数据设定误差阈值，超阈值数据标记后复核；一致性方面，校验产污、治理、排放数据逻辑关联，验证实际排放数据合理性，排查数据篡改风险；有效性方面，剔除设备故障、断电等情况下的无效数据，保留正常生产工况下的有效数据，同时设定数据完整性阈值（单企业单日有效数据占比≥90%）。建立多级质检机制，系统自动初检与人工复核结合，生成质检报告，确保数据集满足非对称异常检测的精度要求。

应用成效:

通过高质量数据集的构建，为生态环境执法智慧决策能力提供了坚实的数据底座，显著增强了企业对自身环境风险的发现与预警能力，同时大幅降低了企业为配合现场检查所承担的迎检负担。基于该数据集训练的企业远程非现场执法场景模型，已深度集成至生态环境智慧管控平台，并在多个关键环节中实现落地应用。应用场景包括烟气旁路排放行为分析、企业产污、治理及排放数据异常发现、疑似偷排行为识别、疑似干扰采样管路行为识别等，试运行半年累计发现线索数 3586 条，人员效率提升 35%，推动生态环境监管由“人海战术”向“智慧监管”转型。

（四）交通物流

交通物流行业数据资源呈现出“体量大、类型多、时效强、增长快”的特征，细分领域涵盖交通流量、路网结构、运力调度、运输状态、物流节点等。在政策推动下，交通物流行业高质量数据集的建设正在快速推进，下面以中国交通信息科技集团有限公司的交通基础设施多模态三维构件数据集为例进行介绍。

专栏 7：交通基础设施多模态三维构件数据集

案例背景:

为系统性破解传统交通基建全生命周期数字化转型中的信息壁垒与数据价值释放瓶颈，中国交通信息科技集团有限公司于 2020 年 7 月启动了《交通基础设施多模态三维构件数据集》的建设，旨在铸造一个高保真、标准化的三维数字资产“底座”，为交通基建的智能设计、车路协同感知、数字孪生运维等前沿应用提供高价值的核心数据要素。该高质量数据集涵盖交通基

础设施对象的点云、图像和文本信息，共计 11.8TB，已入选国资委首批央企高质量数据集建设优秀成果。

建设方案:

本数据集建设过程遵循了源数据层、数据处理层、数据标注层，数据建模层、数据应用层，以及贯穿全流程的数据质量控制层和模型闭环迭代层的“5+2”架构，旨在确保数据集的高质量、高标准化和高可用性，有效支撑算法优化和场景落地。

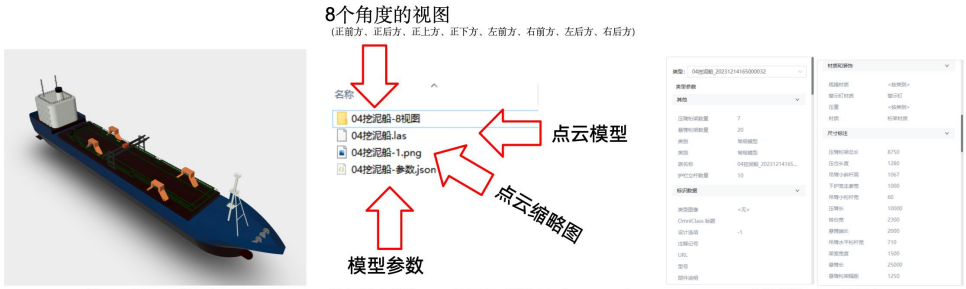
图：建设架构图

原始三维数据的获取遵循多元化与高标准原则。数据来源广泛整合了集团内部承建的重大工程项目（例如 CZ 铁路、平陆运河项目）所产生的 BIM 模型，并辅以外部专业构件资源库的采购以及合作伙伴的共享资源。在数据采集与整合阶段，所有数据均严格遵循中交集团 BIM 系列标准，包括编码标准、建模标准及审核标准，以实现源头统一。此过程借助自动化工具对模型的几何精度与参数合规性进行前置校验，确保了数据格式的一致性与模型属性的精准对齐。同时，依托中交集团在公路、水运等核心业务的深厚积累，构建了覆盖 2327 个细分构件类型的五级分类体系，为后续应用构建了清晰的标准体系。

为保证数据高质量，对原始 BIM 模型实施精细化预处理与数据清洗。核心任务是将多样化的原始模型统一转换为标准化的三维点云、多视角二维图像及参数化文本，同步完成坐标精确对齐与必要的参数脱敏。通过集成应用点云坐标归一化、去噪，图像分辨率标准化与压缩，文本语法校正与单位统一等技术手段，全面保障了数据的几何精度与信息完整性。

在数据标注环节，遵循“专家经验智慧+AI 辅助赋能”的混合样本生成

模式。由于 BIM 数据采集阶段的高标准，具备深厚行业背景的专业工程师已对构件类型、材料属性等核心参数完成了精准标注，确保 BIM 构件参数内容的专业性与准确性。然后，基于已提取的 BIM 参数，通过融合图像描述模型、大语言模型和多模态大模型等多种策略自动或半自动生成与三维模型匹配的自然语言描述，极大提升了标注效率与描述丰富度。这一策略确保了“点云-图像-文本”三者之间形成紧密对齐、语义一致的多模态样本。标注流程中还内嵌了自动化格式完整性校验程序与人工语义一致性核验机制，进一步提升了标注数据的质量。



图：数据样例

最后共产出 59308 个样本，每个样本包含一个对象高密度点云、8 个方向的截图和对象文本描述数据，数据总量约 11.8TB。与世界上先进三维数据集多聚焦日常物品、室内场景或三维游戏场景不同，本数据集专注于交通基建领域和交通运输行业，以通用、水运、机场、公路、市政、轨道交通进行分类，涵盖了从设计到施工的全流程数据，可广泛应用于逆向建模、智能设计、施工养护运营场景生成及具身智能仿真环境构建等，是交通基建行业内规模最大的同类多模态数据集。

数据集名称	真实数据	完整三维数据	对象数量（千个）	分类数量（个）	发布单位
ShapeNet		√	51k	55	斯坦福大学
ModelNet		√	12k	40	普林斯顿大学
3D-Future		√	16k	34	清华大学
ABO		√	8k	63	加州大学伯克利分校
Toys4K		√	4k	105	美国佐治亚理工学院
CO3D	√		19k	50	Meta AI
DTU	√	√	124	-	丹麦技术大学
ScanObjectNN	√		15k	15	香港科技大学
GSO	√	√	1k	17	Google
AKB-48	√	√	2k	48	上海交通大学
交通基础设施多模态三维构件数据集	√	√	**59.3k**	**2327**	**中交集团**

图：三维数据集对比

应用成效:

该数据集已成功应用于多个国家重点课题研究和国家重点项目，如科技部重点研发计划-2.1 交通基础设施数字化软件技术研发、工信部 2022 软件专项-201BIM 软件开发及产业化项目等；此外，CZ 铁路、平陆运河等重大工程也已采用该数据集，其应用显著提升了模型建设效率，提升率达 60%。

同时，该数据集在实际场景中落地成效显著。基于 BLIP-2 多模态对齐的三维构件分类算法，将构件审核效率从传统人工审核的数天缩短至数小时，人力成本降低近 70%，资源发布周期压缩 80%，日均处理构件量达 2000+。

而基于 PointLLM 的 BIM 构件智能三维分类系统，分类准确率提升至 90%以上，较传统人工标注效率提升超 3 倍，支撑日均处理 3000+构件，大幅加快了工程项目推进速度，提升了资源利用效率。相关算法已斩获多项成果，包括授权国家发明专利 4 项，授权国际 PCT 发明专利 1 项，软件著作权 9 项，发表 IEEE 学术论文 4 篇，CCF A 类期刊论文 1 篇，获得省部级特等奖、三等奖各 1 项。

在社会效益层面，该数据集同样具备重要价值。在逆向建模与智能设计领域，其可通过快速生成高精度三维点云模型，有效缩短设计周期、降低设计成本；进入施工阶段，结合生成式算法可优化施工流程、减少施工风险，预计能为项目节约 10%～15%的施工成本。此外，该数据集还蕴含巨大发展潜力，可支撑数字孪生、智能交通系统与载具协同、具身智能等前沿技术发展，助力行业实现全方位变革升级。

（五）医疗卫生

在医疗行业的疾病筛查、临床诊断与疾病预测等核心场景中，高质量数据集已成为推动智能化落地的关键基础，下面以中国联合网络通信集团有限公司的肺结核影像精标注高质量数据集为例进行介绍。

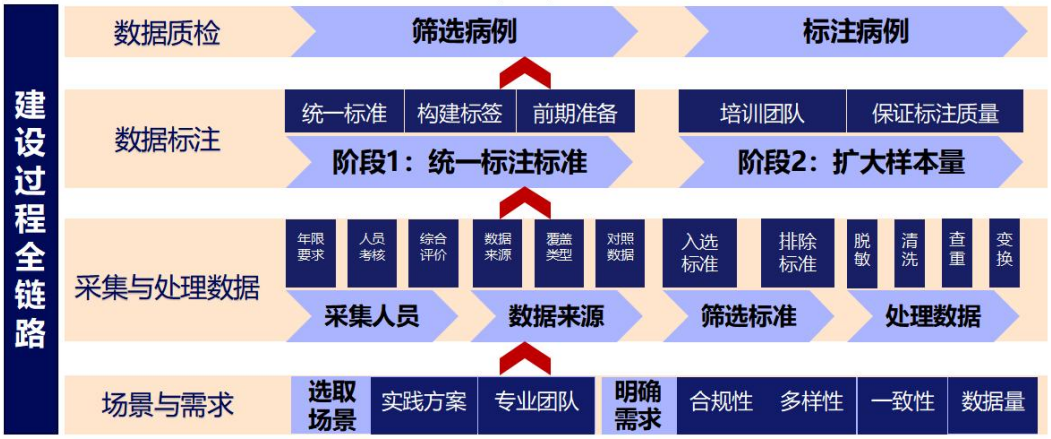
专栏 8：肺结核影像精标注高质量数据集

案例背景：

肺结核是一种严重危害人类健康的慢性传染病，当前肺结核的诊断主要依赖临床症状、实验室检查以及影像学检查等手段，其中胸部 CT 影像学发挥着关键作用。肺结核影像常与肺炎、肺癌等其他肺部疾病混淆，影像学诊断受到阅片医师水平和工作量的多重影响，而传统的人工诊断方式需要医生花费大量时间观察影像，对每个患者的诊断过程耗时较长。智能诊断模型能够快速对胸部 CT 影像进行分析，在短时间内给出初步诊断结果，大幅缩短诊断时间，而高质量数据集是训练诊断模型的核心要素。基于这一现状，北京医疗健康大模型有限公司联合北京胸科医院，构建了 2 万例确诊肺结核患者的 1.25mm 层厚胸部 CT 影像高质量数据集，为训练智能诊断模型提供了数据基础。

建设方案：

数据集建设遵循标准化流程，涵盖场景选取与需求明确、数据采集、数据处理、数据标注和数据质检五大环节，以确保数据高质量与可复用性。



1. 场景选取与需求明确

数据集建设依托国内结核病学科排名第一的首都医科大学附属北京胸科医院，该院具有连续 10 年的包含全类型肺结核的胸部 CT 影像资料，并拥有一支影像学和结核病专家队伍。胸部 CT 影像肺结核数据集使用的临床胸部 CT 图像，均是获得医院伦理委员会批准或者豁免的临床脱敏数据，数据集尽可能覆盖到更多通用的统计维度，以达到高质量数据集构建的安全、丰富的要求。

2. 数据采集

数据来源为 2014—2024 年期间经明确诊断并完成 1.25mm 薄层胸部 CT 扫描的肺结核患者影像资料。数据选择确保肺结核病例涵盖不同类型（如原发性肺结核、血行播散性肺结核、继发性肺结核等）、不同病情阶段（初治、复治、轻度、中度、重度）。

3. 数据处理

数据处理包括数据脱敏、数据清洗和数据变换三项工作。**数据脱敏**要求患者个人信息、医保信息、活动轨迹、经济社会状况、家庭情况、财务信息等个人隐私不得搜集和显示、图像头文件中涉及患者隐私的字段必须进行脱敏处理。**数据清洗**是根据图像质量要求，对图像进行检查和确认，去除不合格数据。图像的质量要求包括格式的有效性、单个文件的完整性、序列的完整性、图像内容的合理性等，不符合 DICOM 标准的图像、破损或无法读取的图像、不连续的图像序列、与肺部无关的序列都是应去除的对象。**数据变换**包括图像标准化、图像分割和数据增强，其中标准化是将所有图像统一调整为 1.25mm 层厚、分辨率（如 512×512 像素），并采用合适的窗宽窗位（如肺窗：窗宽 1500-2000HU，窗位-600--800HU；纵隔窗：窗宽 300-500HU，窗位 30-50HU），以保证图像特征的一致性；图像分割是利用分割算法将肺部组织从 CT 图像中分割出来，去除背景及其他无关组织，保留有效信息；数据增强将原始数据集扩大数倍，使模型在训练过程中能够学习到更丰富的图像特征。

4. 数据标注

第一阶段，组织影像学领域专家针对肺结核患者胸部 CT 影像的基本病变、基本征象和病理特点，形成统一的标注规范和自定义语义标签体系。组建影像学和结核病专家团队，基于肺结核的形态、分布、大小、病理等基本征象，构建 17 个自定义语义标签，使用医学专业 3D Slicer 软件在前期完成了 2000 余份肺结核胸部薄层 CT 的精标注，并开展小样本模型训练验证，模型诊断准确率达到 85%，验证了可行性和准确度。第二阶段，组建百人肺结

核影像标注医师队伍，通过培训、质控、验收等多个环节确保标注质量和一致性实现数据集大规模生产，最终完成 2 万份肺结核影像 CT 自定义语义标签精准分割标注。

5. 数据质检

质检从扫描层厚、窗宽窗位一致性、病灶标注准确性等维度进行系统核查。通过分级质控流程、标注医师能力校验机制及重复抽检制度，确保标注结果在准确性、一致性与完整性方面达到高标准要求。

应用成效:

该数据集整合了首都医科大学附属北京胸科医院连续 10 年的 2 万例患者胸部 CT 影像，覆盖全部结核类型及病程阶段，并依据 17 类统一征象标准进行智能化辅助标注，标注效率从 30 例/周提升为 1800 例/每周。

（六）现代农业

现代农业正在加速向数字化、智能化和绿色化转型，整体上涵盖种植业、养殖业、农产品加工、农业机械装备制造等多个细分领域。各领域数据不断融合，构成了涵盖生产、加工和装备的综合性农业数据体系，展现出规模大、类型多和周期性的特征。下面以国机数字科技有限公司的农业机械及作业高质量数据集为例进行介绍。

专栏 9: 农业机械及作业高质量数据集

案例背景:

农业机械是农业生产和农业应急救灾的主力军，是驱动农业现代化的重要生产力。农机作业过程包括耕整地、播种、植保、收获等关键环节，具有作业场景复杂、工况多变、时空跨度大等典型特征。随着智能农机、自动驾驶和精准作业技术快速发展，农机装备在运行过程中持续产生海量多源数据，包括作业轨迹、作业参数、机具状态、环境感知及运维信息等。高质量农业机械及作业数据集是支撑农机智能控制、作业质量评估、生产调度优化和安全监管的重要基础。然而，当前行业普遍存在数据采集标准不统一、标注体系缺失、数据质量参差不齐、难以支撑算法训练和规模化应用等问题。为此，围绕农业机械典型作业场景，系统构建农业机械及作业高质量数据集，对推动农机装备智能化、农业生产数字化和行业高质量发展具有重要意义。

建设方案:

农业机械及作业数据集建设遵循“统一标准、场景驱动、多源融合、质量优先”的原则，形成覆盖场景选取、数据采集、数据处理、数据标注和数据质检的全流程建设体系，依托农业机械行业可信数据空间，通过数据采集、数据处理、数据标注，实现数据接入到数据应用的闭环。

1. 场景选取与需求明确

数据集聚焦农业机械典型作业场景，重点覆盖耕整地、播种、植保和收获四大核心环节，涉及拖拉机、播种机、植保机械、联合收获机等主要机型。业务需求包括：支撑作业轨迹识别、作业质量自动评估和异常行为识别；支撑智能驾驶与作业参数自适应调控模型训练；实现作业过程可视化、可量化和可追溯；服务农业金融保险场景，提供农机全生命周期主要数据。

数据需求涵盖高精度北斗定位数据、作业质量数据（速度、深度、幅宽等）、机具状态及环境信息，并要求数据具备统一格式、完整元数据和明确质量等级。

2. 数据采集

数据采集依托智能农机作业示范区和规模化生产场景，通过车载终端、传感器系统和视频采集设备同步获取多源数据。定位数据采用亚米级/厘米级北斗高精度定位，采样频率不低于 10 Hz；作业参数数据来自农机控制系统，覆盖关键作业指标。同时，该高质量数据集与农业农村部农机化有关系统打通，汇集全国农机购置补贴、农机安全监理系统数据，基本覆盖农机关键环节的相关数据。

在对农机作业过程中的数据采集过程中，对异常数据、信号中断和明显失真的样本进行剔除，确保数据真实反映农机作业状态。同时建立完整元数据体系，记录作业时间、地点、机具型号及作业模式，实现全流程可追溯。

3. 数据处理

数据处理包括预处理、对齐融合和结构化整理三个层次。

在预处理阶段，对定位数据进行漂移修正和平滑处理，对视频数据进行去抖动、帧率统一和关键帧抽取；

在融合阶段，基于时间戳对多源数据进行精准对齐，实现轨迹、参数与视频信息的同步映射；

在结构化阶段，将非结构化视频和日志数据转化为标准化数据格式，形成可直接用于算法训练和分析的数据样本。

4. 数据标注

数据标注围绕“作业行为”和“作业质量”两类核心目标展开。

行为标注包括直线作业、转弯、重叠、漏作、异常停车等典型行为；质量标注包括作业深度合格度、行距一致性、覆盖均匀性等关键指标。标注流程采用“规则初标—模型辅助—专家复核”的方式，结合农机作业规范和农业生产经验，形成统一标签体系和判定标准，显著提升标注效率和一致性。

5. 数据质检

数据质检从准确性、一致性和完整性三个维度开展。通过抽样复核和自动化脚本检测，校验标注与原始数据的匹配程度；通过跨批次、跨场景对比，验证标签定义和判定标准的一致性；通过完整性检查，确保关键作业阶段和核心字段无缺失，全面保障数据集质量。

应用成效:

1. 经济与业务价值

一是支撑企业装备研发与性能优化。通过高质量数据集，有效降低集团内部农机企业的研发试验成本，基于作业数据的算法模型帮助其开发智能控制系统并在量产机型中落地，显著提升市场竞争力。

二是促进智能算法模型研发与验证。依托数据集，行业内 10 余家科研机构构建了作业识别模型、油耗预测模型、田间效率评估模型等，直接推动智能作业监控、智能调度、智能农机导航算法提升。

三是赋能政府监管与政策制定。多省农机管理部门使用该数据集开展作业补贴核验、区域农机投放量测算、农时分析等工作，决策依据更加精确，政策执行透明度显著提高。

2. 社会价值与行业影响

一是推动农机行业数字化转型。数据集成为行业共享标杆，促进了装备企业、合作社和政府部门之间的数据融合与协同。行业首次形成从采集到治理再到发布的全链条标准体系。

二是提升农业生产效率与资源利用率。基于数据集构建的智能饲喂、智能调度、精准作业系统。

三是构建行业生态与开放合作体系。数据集发布后，吸引行业头部 AI、气象、遥感、农机企业等 20 余家参与联合创新，形成跨行业协作生态，为我国智慧农业提供长期稳定的数据底座。

（七）移动通信

移动通信行业是数字经济的核心支撑，也是数据规模最大、更新速度最快的行业之一。随着 5G、物联网和即将到来的 6G 通信，移动通信行业数据不仅可服务于大众通信需求，还延伸至工业互联网、车联网、智慧城市等关键场景，高质量数据集在行业中既承载着通信基础设施运行的核心信息，也支撑跨行业数字化转型的应用创新。下面以中国移动通信集团有限公司的中国移动人时空三元组高质量数据集、中国电信集团有限公司的通信行业网络大模型高质量数据集、中国联合网络通信集团有限公司的信息通信领域高质量数据集为例进行介绍。

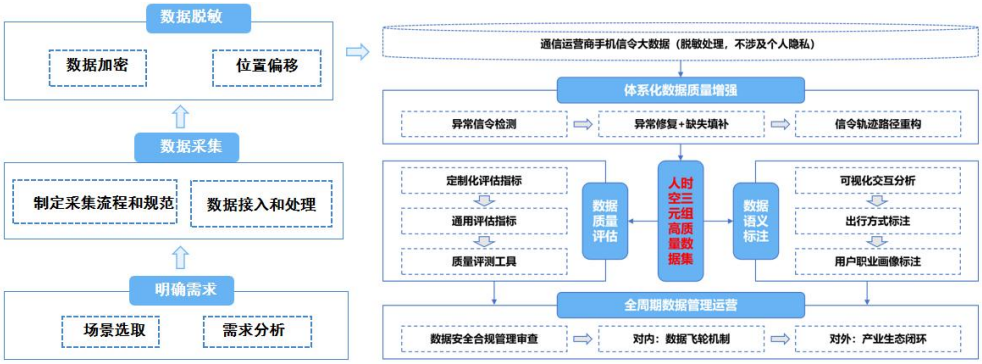
专栏 10：中国移动人时空三元组高质量数据集

案例背景：

中国移动十亿级手机用户和百万级通信基站产生的信令数据蕴含了大规模人群 7×24 小时连续的时空活动信息，具有“大样本、全过程、广覆盖”的优势，在交通运输、城市治理等领域具有广泛应用价值。但是，当前原始信令存在数据质量低、语义信息缺失、开发利用不足等问题，严重限制其价值释放。为解决该问题，本案例突破信令异常多元复杂、工具机制不完备等挑战，构建形成一套高质量的人时空三元组数据集，并将其应用于出行大模型训练、微调，以及交通、应急、文旅等行业相关下游任务推理。

建设方案：

本案例将围绕场景选取与需求明确、数据采集、数据处理、数据标注、数据质检五个环节，系统阐述如何将原始信令数据转化为能够准确反映个体出行轨迹与人地交互关系且具备丰富活动语义标签的高质量人时空三元组数据集。



图：人时空三元组高质量数据集整体构建流程

1. 场景选取与需求明确

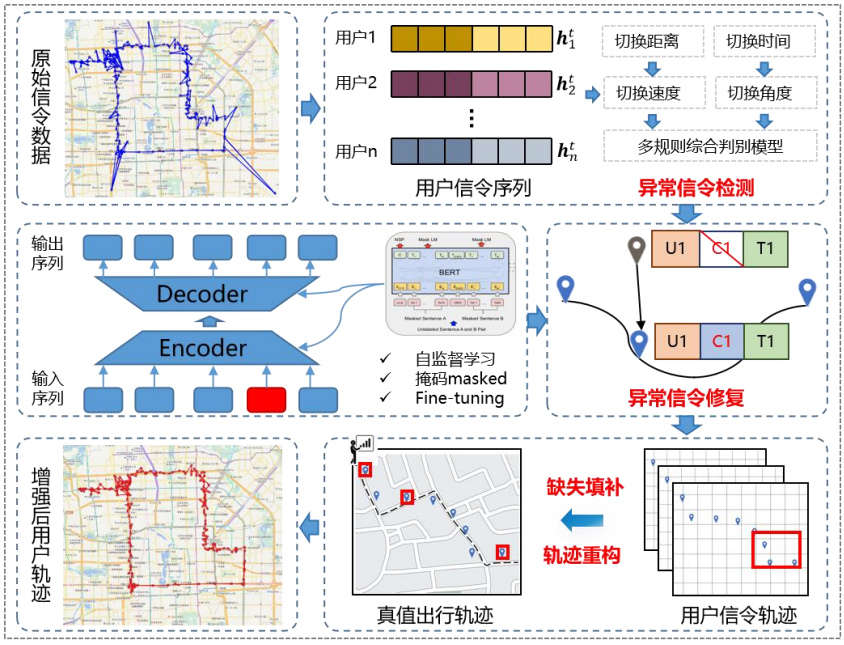
本案例聚焦于“辅助精准预判节假日交通态势、优化人流与拥堵动态管理”这一核心场景，构建以“人-时间-空间”为核心要素的高质量数据集，并将其应用于“九天川流出行大模型”，实现对城市人群出行链的精准识别与预测，提升节假日交通管理的智能化水平，为上层业务模型提供稳定可靠的数据基础。

2. 数据采集

数据由省级共享平台统一将 2G/4G/5G 等多制式信令数据汇聚至九大边缘节点，通过 Flume 进行采集，随后依次经过解析、清洗与集成等处理流程，最终输出至 HDFS 实现持久化存储，并同步写入 KAFKA 以供下游流式计算与实时应用。

3. 数据处理

数据质量增强是本案例数据处理的核心环节，旨在通过深度自监督学习与时空插补技术，系统性修复数据缺陷并重构完整轨迹，为下游任务奠定高质量、可信的数据基础。针对原始信令数据中的位置异常、缺失等问题，提出基于深度自监督学习的数据质量增强方法。该方法通过建模信令间的时空关联，实现对异常、缺失信令的修复与填补，使修复后信令平均切换距离从 800m 缩至 160m，500 米内占比达 96%，数据完整度从 90%提升至 100%；针对采样间隔不规律与长时间缺失导致的轨迹稀疏问题，提出基于时空插补的轨迹路径重构方法。通过学习用户时空行为规律及区域位置关联特征，对缺失轨迹进行时空插值，将稀疏轨迹重构为完整度较高的连续路径。



图：体系化数据质量增强方案

4. 数据标注

本案例打造了一款专门面向信令数据的标注工具。以 GIS 地图结合交互式问题引导的方式，在有效避免敏感信息泄露的基础上，实现了对数据的可视化展示以及对出行方式（如地铁、公交、驾车等）和与出行紧密相关的职业画像（如网约车司机、快递员、外卖员等）标签的高效标注，构建出具有丰富活动语义标签的出行标注数据集，用于支撑面向特定领域下游任务的模

型微调与效果评估。

5. 数据质检

本案例参考《自然资源领域应用手机信令数据技术指南》等行业标准，建立了一套信令数据质量评估体系。首先针对信令数据的独特性，从相邻信令之间的切换距离、角度、速度等维度设计定制化评估指标，同时引入数据完整性、唯一性、合规性、准确性等通用评估指标，全面评估信令数据的质量。

应用成效：

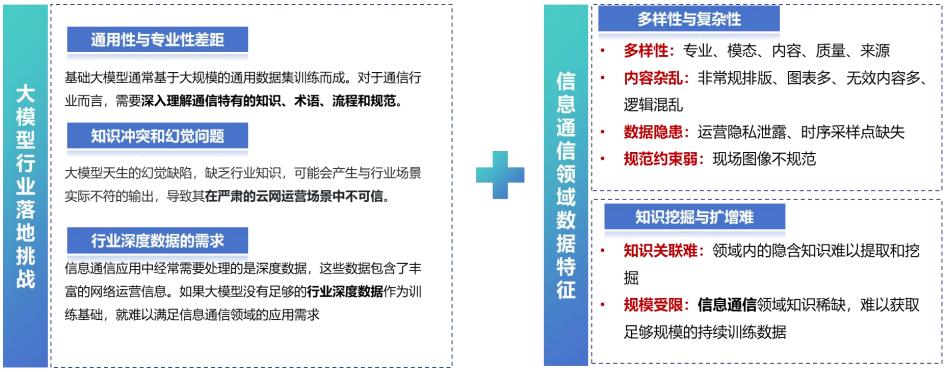
本案例入选“首批中央企业人工智能行业高质量数据集优秀建设成果”，并获评“国家数据局高质量数据集典型案例”。基于人时空三元组高质量数据集，中国移动打造了超 15 亿参数规模的“九天川流出行大模型”和 1 个智能体应用平台“川流智行”，构建覆盖交通治理、政务服务、选址推荐、文旅运营、应急决策等领域的 5 个 AI+标杆应用，广泛应用于交通、文旅、政务、电力、应急等 5 大行业，推广至央企、政府部门等十多家单位。

案例成果在 2025 世界移动通信大会（MWC）、第八届数字中国建设峰会、“和美乡途”央企文旅帮扶平台发布会、2025 中国国际大数据产业博览会、2025 中国移动全球合作伙伴大会等 5 项重要行业展会中亮相，显著提升了行业知名度与生态影响力；荣获“2024 数博会十大领先科技成果”“2024 年 AI For Good 全球峰会（ITU）杰出案例奖”“第五届中国科技产业化促进会科学技术奖二等奖”等 10 余项国内外重要奖项。

专栏 11：通信行业网络大模型高质量数据集

案例背景：

大模型落地信息通信行业面临通用性与专业性差距、知识冲突与幻觉、深度数据缺乏等挑战，中国电信网络大模型高质量数据集确立以云网知识体系为基础，以组织、流程、运营和安全为支柱，建设统一的知识管理平台的科学体系化建设方法，整合通信领域多源复杂数据，构建云网知识体系，打造 5 大类高质量数据集及数百个场景化知识库。通过数据集工具、关键技术研发与高效运营，形成“专业深度+实时进化”双能力，促进人才转型，支撑云网运营 AI 化及自智水平达 L4 级，显著提升运营效率与稳定性，成效惠及企业及行业客户，促进产业链协同。



建设方案：

网络大模型高质量数据集以云网知识体系和技术规范为基础，以组织、流程、运营和安全为支柱，建设统一的知识管理平台，实现为网络大模型持

续提供高质量的数据集以及知识增强服务，为中国电信人才转型提供培训支持。

一是确立科学体系化数据集构建方法。以“5+2”云网知识体系为基础，建设统一知识管理平台，重构组织流程与安全体系，形成领域全覆盖分类标准，实现跨域知识穿透协同，建成大规模、高质量、多模态领域数据集，提供分钟级建库工具、数据增强服务及动态管理能力，知识增强服务检索量达千万次。

二是突破数据治理技术瓶颈。融合多源智能解析、动态清洗、多级去重及数据合成技术，研发隐私保护算法与知识图谱流水线。自研 SIE 分层建库技术实现多模态数据分钟级建库入库，创新多模态混合检索架构及重排序技术，攻克通信数据规范难题，隐性知识发现效率提升 5 倍，召回准确率提升 35%；建成高质数据治理范式。

三是构建全生命周期运营模式。建立“生产—规划—采集—加工—应用—评估—反馈”闭环。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，保障动态更新，建成 5 大类数据集，支撑数百个知识库高效运行。

四是规模化赋能产业转型。提升云网运营效率，助力云网自智达 L4 级。培养千余名“三懂”复合人才（懂业务/网络/模型），支撑数十类云网 AI 助手与数字员工，覆盖全网。

图：网络大模型高质量数据集构建方法

网络大模型高质量数据集实现技术、模式、管理三方面创新：

一是高效数据标注与增强技术创新。融合多源解析、智能标注、动态清洗、多级去重、隐私保护、知识图谱及数据合成技术，首创 SIE（来源－索引－编码）分层建库技术实现分钟级建库，创新多路混合检索及重排序、检索交错生成与时序数据均衡技术，打通实时数据与大模型动态交互链路。

二是全生命周期管理模式创新。建立“规划－采集加工－应用－评估”闭环管理机制，依托“5+2”知识体系分类分级数据，通过平台精准治理，结合 AI+场景转化高安全性知识。

三是两级联动长效运营机制创新。构建两级多级闭环管理机制，明确语料审核、定期通报及数据报送职责；强化协同联动，创新知识入库时效管理，建立问题跟踪闭环，形成“组织－流程－协同－时效－反馈”运营体系。

应用成效：

网络大模型数据集从 2023 年开始建设，通过整合中国电信云网用户、订单、产品、业务、云网的知识与数据，构建数据集业内领先，先后入选国

务院国资委首批 30 项高质量数据集优秀建设成果、国家数据局首批 104 个具有推广示范价值的高质量数据集典型案例，位列信息通信领域首位。

启明知识管理平台应用覆盖集团+31 个省份，建成 5 大类数据集，包括各类训练、微调、推理文本数据，云网运行产生的结构化时序数据，中国电信机房巡检、资源核查等多模态数据集，覆盖云网多业务、专业、场景的知识图谱数据，支撑数十类云网 AI 助手与数字员工构建数百场景知识库，检索服务调用量达千万次。极大提升了云网运营效率，助力云网自智达 L4 级，培养千余名“懂业务、懂网络、懂模型”复合 AI 人才。

网络大模型高质量数据集构建方法与数据标注等能力成功输出给某工业制造央企，构建了该行业知识体系和知识分类标准，启明知识管理平台的数据标注加工能力为其完成几百万条数据标注，输出 150 亿 Token 数据集，约 10 万余条问答对。

专栏 12：信息通信领域高质量数据集

建设背景：

信息通信行业存在海量数据资源管理难、多源异构数据融合难、场景化应用不足等痛点。针对上述挑战，中国联通发挥集约化数据资源优势，统一归集与处理全域数据，打造信息通信高质量数据集，覆盖网络运营、客户服务、智能终端、电信反诈、经营管理、科研创新六大人工智能应用场景，总量达 53.5TB，质检合规率超 98%，在网络故障智能诊断、电信涉诈智能识别、提升客户服务与企业管理的智能化水平场景中广泛应用。

建设方案：

中国联通以核心生产经营业务数据为基础，深度挖掘结构化、非结构化数据资源，依托统一的数据标准体系和平台化工具，构建标准化、规范化、体系化的数据集生产体系。基于联通大数据平台打造覆盖“采集-清洗-标注-质检-应用-评估”的端到端多模态一站式数据生产流水线，自研 43 项智能数据处理算子、26 项智能标注能力，依托 AI 智能工具集与标准化处理，实现通信领域文本、音视频、日志等数据的自动化采集、跨模态对齐与高效流转，同时建立包含 70 余项量化指标的数据集质量评价体系，实现高质高效的数据供给，整体提升模型训练效率 50%。

1. 多源数据采集

网络数据集：汇聚网络领域多维度核心数据资产，包括运维解决方案、历史案例库、技术标准、操作规范、咨询报告等内部专业文档。

客户服务数据集：汇聚客服热线音频数据、客户意图标注数据、智能家庭工程师施工场景图像等多模态数据，扩展多方言场景语音交互数据。

智能终端数据集：汇聚中国联通终端日志等核心数据资产，涵盖设备运行日志、性能监测数据等多类型终端交互信息。

电信反诈数据集：汇聚中国联通涉诈短信文本、风险网页内容、黑白名单记录等核心业务数据，结合伪基站诈骗案例库、国际涉诈短信语料库、人脸特征视频及深度伪造检测视频等多模态数据。

经营管理数据集：汇聚中国联通办公公告与公文数据、中国联通多维度内部经营分析问答对、数据查询检索脚本文本数据，深度融合互联网新闻、行业报告、市场动态等多模态数据。

科研创新数据集：整合中国联通多维度技术研发数据，包括内部设计文档、编码规范、测试用例、代码生成标注数据等核心技术资产。

2. 智能化数据处理

多模态数据在清洗后通过语义增强、几何与色彩变换、切片匿名化与方言合成等技术提升文本、图像与音频样本的多样性与鲁棒性。数据蒸馏环节依托专家模型与扩散模型生成高密度与虚拟样本，构建非法请求等负样本库用于强化学习，如音频类数据集扩充覆盖 31 省的方言语料（1542 小时），识别准确率达 89%。同时，通过分布式向量引擎实现 TB 级数据向量化与音文、图文的跨模态语义对齐，为行业模型训练提供一致规范的语料基础。

3. 自动化数据标注能力

整合多模态标注工具，提供覆盖文本、图像、语音等 7 种模态数据的 26 项智能标注能力，并通过预处理和模型训练持续提升预标注质量，从而提高整体标注效率。网络运营场景中，大模型自动抽取指令类知识，用户基于预标结果微调，效率提升 30%；客服场景中，依托元景大模型进行意图识别和对话分类，效率提升 50%；在电信反诈场景利用 NLP 实现短信风险类别快速标注，减少人工依赖；经营分析场景通过实体关系模型自动提取关键信息，效率提升 60%；在科研创新场景提供代码错误识别、注释和总结等能力，效率提升 50%以上。

4. 多维度标准化数据质检体系

建立“标注-稽核-评估”三级联动质检机制，融合 AI 预标注算法与专家在线校验，形成六个基础维度和十个个性维度共 70 余个可量化质量指标（如字符重复率<3%、敏感词占比<0.1%），严控数据一致性、完整性风险，支持数据回流与数据集评价。

5. 全域标准化分级管理

构建数据分类分级、访问控制、动态脱敏三大核心能力，支持数据集的多维检索与订阅，为数据科学人员提供一站式的 AI 数据生产管理服务，实现敏感数据识别准确率>98%、脱敏覆盖率 100%。

应用成效：

中国联通通信行业数据集高效供给 30 个大模型训练与微调，赋能近百项细分场景应用，有效提升运营管理效能。

网络运维效率提升：实现基站故障提前预警、故障根因定位效率提升至 92%，网络知识问答大模型性能提升 35%;

智能客服体验优化：形成“语音转写-诉求分析-方案推荐”全链路服务闭环，日均处理 200 万次交互，满意率提升 6.3%，咨询意图识别准确率 95%，客服大模型问题响应率达 95.9%;

终端运维效率提升：支撑产品经理进行终端运营分析、智能定价与进销存预测，按需分析终端日志，通讯终端问题响应速率提升 20%;

诈骗识别能力增强：涉诈短信、对话等 5 类场景识别能力显著提升，诈骗骚扰电话提醒率提升 100%，用户受诈风险预警准确率超 85%;

经营决策效率提升：交互式智能问数功能累计使用 50 万次，交互分析自动生成报告累计 251 万次，制作时间由一周减到一天，面向广东、山东等 12 个省份推广，好评率评价为 87.8%;

研发效率提升：赋能元景代码助手，支持超 8.4 万用户构建轻应用，研发效率提升 95%。

（八）应急管理

应急管理行业是国家安全体系和社会治理体系的重要组成部分，主要涵盖自然灾害防治、事故灾难应对、公共安全管理和应急救援等多个细分领域。随着应急管理体系完善，行业对多源数据的需求持续增强，高质量数据集在其中扮演着基础性支撑角色，通过数据应用可有效提升应急管理的风险防范和事故处置能力。下面以新兴际华集团有限公司的大型石油储罐火灾救援数据集为例进行介绍。

专栏 13：大型石油储罐火灾救援数据集

建设背景：

大型石油储罐灭火救援领域的相关数据，长期存在碎片化、实战案例稀缺、多模态数据融合不足的问题，且数据未进行完整标注，这直接导致该场景下的模型训练效果和场景适配能力大打折扣。建设火灾救援多模态数据集的核心目的，是补齐数据短板，为火灾识别、灭火决策、装备问答等核心模型的研发提供精准全面的数据底座，同时推动应急救援领域从经验决策向数据驱动转型。

建设方案：

大型石油储罐火灾救援数据集遵循“数据采集 — 清洗标注 — 增强合成 — 模型训练 — 应用”的全流程构建逻辑，核心包含四类核心数据：一是火灾多模态数据，涵盖火场高清图像、动态视频等，覆盖储罐火灾不同燃烧阶段与场景；二是灭火救援技术数据，收录真实大型石油储罐火灾案例，包含火情研判、战术部署、装备使用等完整决策链路；三是应急装备数据，涵盖消防车、无人机、灭火机器人等应急装备的基础参数、操作视频、性能边界等全维度信息；四是消防规范数据，包含石油化工防火标准、灭火理论公式、救援规程等权威文档与专业知识。

采集阶段整合智能应急装备、储罐运行参数、火场实时影像、实战救援案例，以及公开的消防规范与行业研究成果等各类数据。

清洗环节针对性处理不同类型数据：文本数据通过自研工具完成实体识别、字段补全与格式统一，剔除异常值；图像与视频数据进行尺寸归一化、去噪处理，保障格式与分辨率统一。标注环节构建智能预标注 + 专家精修机制，图像按火焰区域 + 燃烧物类型 + 火情等级三级体系标注，文本遵循“Alpaca 指令 - 输入 - 输出”三段式结构。

增强合成环节基于生成对抗网络生成浓烟、夜间等稀缺火情场景数据，通过数字孪生系统模拟储罐火灾救援场景，填补实战数据空白；利用大模型生成标准化灭火案例文本，扩充数据集多样性。同时建立包含说明文档、数据质量、模型应用三类指标的评估体系，从准确性、完整性、时效性维度常态化校验数据，确保多模态数据语义对齐。

该数据集主要支撑三大应用场景：知识问答场景，依托应急装备数据和

消防规范数据训练应急装备大模型，实现装备参数、操作规范的精准问答；指挥决策场景，基于灭火救援技术数据训练灭火战术辅助决策大模型，生成科学的救援策略；火灾识别场景，利用火灾多模态数据训练火灾识别大模型，精准判定火灾类型及详情。

应用成效：

该数据集为大型石油储罐灭火装备智能化升级提供了核心数据支撑，显著提升了装备实战效能。依托该数据集训练的相关模型，有效优化了灭火救援流程，全面赋能灭火装备适配各类火情场景作业，提升了装备操作响应效率。

数据集可通过基础数据授权、数据查询服务、定制化数据报告等方式交付科研机构、应急装备企业；也可针对大型石油储罐灭火装备定制专属模型，作为边缘大模型部署在装备端，为相关企业的灭火装备赋能。此外，数据集的建设经验可复用于化工储罐、天然气储备库等高危险设施的应急装备研发，推动整个应急装备行业的智能化升级，助力减少火灾事故造成的财产损失，为能源行业安全生产提供坚实保障。

五、主要结论及未来展望

通过对多家中央企业的深入调研与案例分析发现，在国家“人工智能+”行动和数据要素等相关政策推动下，央企高质量数据集建设已从散点探索进入规模化推进阶段，在能源、制造、交通等关键领域的模型训练和业务应用中取得初步成效，并形成了一批行业示范成果。然而，通过调研和研究也揭示，当前数据集建设仍普遍处于“强项目、弱体系；重内部、轻生态；有数据、缺机制”的初级阶段。面向未来，央企高质量数据集的建设亟需从以单一项目为导向，转向构建可复用、体系化的基础能力，并推动高质量数据集由企业内部资产，逐步发展为支撑全行业协同创新的基础性资源。

（一）现状评估：成效显著，但深层次矛盾亟待破解

当前央企高质量数据集建设取得了三方面初步成效：**一是场景化落地验证了数据价值，**在故障诊断、智能巡检、工艺优化等场景中，数据集驱动的人工智能应用已产生显著的经济与社会效益。**二是行业共识基本形成，**企业已明确将数据集作为 AI 时代的核心战略资产。**三是技术工程化能力初步构建，**部分领先企业已搭建平台技术架构，形成支撑多模态数据处理的工具链，覆盖采集、标注、治理、评估等流程，并逐步完善质量评估流程和指标体系。

同时，制约高质量数据集建设从“可建”走向“优建”、从“自用”走向“共用”的深层次矛盾依然突出：

一是制度性瓶颈构成阻碍。数据权属、收益分配、安全责任等基础制度缺失，导致企业间数据共享“不愿、不敢、不能”，数据集价

值无法释放，重复建设与数据孤岛并存。

二是工程化能力存在短板。多数企业的数据处理流程仍依赖人工和定制化开发，面向多模态、大规模数据的智能处理、自动化标注与持续质量监控体系尚未形成，建设成本高、周期长、质量不稳。

三是项目制思维阻碍发展。数据集建设多依附于特定 AI 项目，项目结束则运营停滞。缺乏跨部门、跨业务的统筹规划与持续运营机制，导致数据集难以迭代、复用和积累成企业级核心资产。

四是生态位角色亟待厘清。央企在行业数据生态中的角色多维，“资源整合者”“标准制定者”“平台运营者”等生态位可能同时出现，在实际建设中容易导致各方协作不畅，不仅会影响各方在数据集共建共享的积极性，也会造成央企在资源统筹、规则制定和组织协调等方面的制度性优势难以有效发挥。

（二）核心发现：央企高质量数据集建设的关键要素

央企正从平台建设、生态运营、基础保障三大维度协同发力，全面推进高质量数据集体系化构建工作。在数据集建设方面，央企正加快优化已有平台技术架构，完善开发运营全流程中的必要能力，搭建支撑多模态数据处理的工具链，推动数据采集、处理、标注、质量评估等环节流程化、自动化和智能化，并逐步完善质量评估指标体系。在数据集运营方面，央企正探索完善应用服务和运营监控体系，加快打造行业生态，加强供需对接与标准规范共建，完善生态内流通规则，推动数据资源汇聚与高质量数据集的流通利用。在基础保障方面，央企以集团结合子公司的协同模式推动高质量数据集建设，并加快探索

数据集建设过程的质量、安全等各类管理制度和机制构建。

基于案例实践，报告总结出央企成功建设高质量数据集的关键要素：

一是“业务-数据-模型”闭环驱动。业务牵引效应显著，成功案例均始于明确的业务痛点（如管道安全风险、勘探效率低等），部分案例通过业务效果来反哺数据工程与模型训练，打造“业务产生数据、数据优化模型、模型赋能业务”的数据飞轮，有力提升数据质量。

二是“专家知识+智能工具”人机协同。行业专识知识的注入至关重要，在能源、工业等强专业领域，纯自动化标注无法满足业务需求，必须将领域专家（如工程师、医生、研究员）的知识，通过标注规则、质检标准、预训练模型等方式固化到工具链中，实现“专家定义规则，机器规模化执行”。

三是“集团统筹+一线创新”协同组织。集团重点承担顶层制度设计、共性能力建设和运行机制的统筹，围绕跨业务、跨场景的普遍的共用能力进行统一规划和建设。一线业务单元或专业公司作为创新主体，负责具体场景的数据集构建，并选择合适的运营策略和方法，确保数据集紧贴业务需求，有效支撑业务应用的智能化落地。

（三）未来建议：夯实央企作为行业高质量数据集建设的“压舱石”与“发动机”地位

展望未来，央企高质量数据集建设是关乎国家数字竞争力、产业安全与现代化产业体系构建的战略工程。为推动中央企业高质量数据集建设迈上新台阶，充分发挥央企产业支撑和托底保障的作用，央企

必须超越企业边界，明确自身定位，承担起更宏大的国家使命：**在供给侧，**成为国家关键领域高价值、高可信数据的核心供给者；**在标准侧，**成为行业数据规范与治理规则的主要制定者；**在生态侧，**成为融通产业链、创新链、价值链的数据生态组织者。

央企高质量数据集建设将呈现标准化、智能化与生态化的发展趋势。在制度突破与标准先行层面，数据基础制度、治理体系和安全监管将不断完善，为数据集建设、流通和使用提供更清晰的边界与运行规则；与此同时，行业数据规范、标签体系与质量评估体系将逐步统一，跨企业、跨场景数据可用性显著提升。建议央企在集团内部率先探索数据资产确权、内部核算与收益分享机制；由行业主管部门联合头部央企，在基础较好的领域，率先研制并发布行业级高质量数据集建设标准（含数据格式、标注规范、质量评估等），为跨企业数据互通奠定基础。

在技术攻坚与平台赋能层面，多模态数据处理、数据合成等技术将逐步成熟，促使数据生产方式更加智能高效。建议央企进一步加大数据工程技术投入，研发适配行业特点的智能标注、质量评估、动态监控工具。支持建设“集团级数据资产运营平台”，将分散的项目能力沉淀为可复用、可共享的平台服务，降低下属企业建设门槛。

在生态构建与价值释放层面，基于行业数据空间、区域数据平台和跨主体协同机制的新型生态将加速形成，将推动央企高质量数据集在更大范围内流通利用。建议有条件的央企牵头建设行业可信数据空间，在保障安全与权益的前提下，探索数据“可用不可见”的合作模

式，完善空间内相关制度规则。将自身高质量数据集作为“锚点”，吸引上下游企业、科研机构贡献数据、共同开发，将数据优势转化为产业生态领导力。

综上，通过标准、技术与生态持续投入，央企将具备更持续的数据资源供给能力、更强的国际竞争力和生态组织能力，成为推动我国数据要素市场繁荣的关键力量。通过系统性解决制度、技术与生态问题，央企不仅能筑牢自身智能化转型的基石，更将有力牵引我国重点行业整体数字化水平提升，为深化落实“人工智能+”专项行动、壮大国家数据要素市场、发展新质生产力提供坚实支撑。这条路任重道远，但方向清晰，行动刻不容缓。

联系方式:

一、中电数据产业集团有限公司

地址：深圳市南山区粤海街道科技园社区科发路 3 号

邮编：518052

邮箱：cedc-zhb@cecdt.com.cn

二、中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮编：100191

邮箱：yangjingshi@caict.ac.cn