大数据技术标准推进委员会
Big Data Technology and Standard Committee

高质量数据集实践指南（1.0）

CCSA TC601大数据技术标准推进委员会

2025年6月

版权声明

编制说明

本报告的撰写得到了数据智能领域、高质量数据集领域多家企业与专家的支持和帮助，主要参编单位与人员如下。

参编单位：大数据技术标准推进委员会、中国联通软件研究院、中国联合网络通信有限公司智能城市研究院、中国铁塔股份有限公司、中国移动通信集团有限公司、北京亦庄智能城市研究院集团有限公司、科大讯飞股份有限公司、中电数据产业集团有限公司、中国交通信息科技集团有限公司、中国航天标准化与产品保证研究院、中航信数智科技（北京）有限公司、蚂蚁区块链科技（上海）有限公司、联通数据智能有限公司、亚信科技（中国）有限公司、软通智慧科技有限公司、四川数通智汇数据科技有限公司、蓝象智联（杭州）科技有限公司、振华智造（西安）科技有限公司、上海市数字证书认证中心有限公司、东软集团股份有限公司、通用技术集团财务有限责任公司、重庆金山科技集团股份有限公司、中国医学科学院医学信息研究所、重庆祥富机电技术服务有限公司、普元信息技术股份有限公司、杭州数蜜科技有限公司、中国石油国际勘探开发有限公司

参编人员：白玉真、杨靖世、尹正、姜春宇、刘渊、王思佳、童锦瑞、袁博、康宸、王宇、武天富、李桐、孙亮、董正浩、杜鹏、韩丽、蔡伟霞、唐双林、路骁虎、石庆华、陈雷、杨鹏、王刚、方飞、时蒙福、李嘉宁、刘彬彬、王晶、莫洋、张蕊、刘晓玉、刘锴、叶可、孙晓峰、崔杨、张博、乔娇娇、蔡健生、王昊、陈亚乐、冯文、王立

冬、林镇阳、胡鑫、张冰、李由、王超、奚瑜、李晓燕、王会、杨晶、许强、崔朝辉、祝旭明、方桂全、吴吉芳、李杰、吴思竹、曾祥富、万强、王仕亿、薛良、刘楚、李晓雄、王春红

前　言

随着人工智能技术迈入以大模型为核心的新纪元，数据已成为驱动模型能力跃迁与产业智能化升级的战略资源。DeepSeek 的横空出世颠覆了 “高算力和高投入是发展人工智能唯一途径”的固有认知，引领从业者高度重视数据质量与规模，高质量数据集成为人工智能发展的关键要素。然而，当前产业界面临着高质量场景数据供给不足、建设路径模糊、标准规范缺失、技术工具需提升、数据价值难以释放等多重挑战。

为推动高质量数据集建设，明确建设和运营方法论，加速赋能场景应用，总结未来发展趋势，大数据技术标准推进委员会牵头，联合行业专家共同编制《高质量数据集实践指南（1.0）》。本指南适用于从事数据管理、人工智能研发、数据产品运营的企业管理者、数据工程师、算法科学家及相关从业人员，旨在为其提供一套可参考、可落地的方法论与操作指引，助力业界构建并用好高质量数据集。有以下亮点:

一是“理概念”。从数据集的概念、数据集的分类、高质量的内涵深入阐述高质量数据集的概念内涵。

二是“建体系”。基于理论研究与产业实践，总结高质量数据集建设模式，提供一套覆盖数据集研发、交付、运维、运营全生命周期的建设方法论，并搭建建设运营的成效评估体系。

三是“促应用”。梳理分析高质量数据集的应用情况，并辅助具体场景实践案例为方法论落地实施提供参考。

四是“看趋势”。从建设运营能力成熟度、行业场景应用、协同生态建设等方面展望高质量数据集的未来发展趋势。

高质量数据集是快速发展的新兴领域，新问题、新理论、新技术、新方法层出不穷，我们将持续深耕研究。由于时间仓促，水平所限，本报告仍有不足之处，欢迎联系白玉真（18810275013）交流探讨。

目 录

一、高质量数据集概念与问题

（一）高质量数据集的发展背景

1. 高质量数据是人工智能发展的关键要素

随着人工智能技术迈向大模型时代，行业发展正经历从“以模型为中心”向“以数据为中心”的范式转移。近年来以 GPT、DeepSeek 为代表的大模型技术突破实践表明，数据质量与规模已成为决定模型性能的核心要素。尤其是 DeepSeek 模型在复杂逻辑推理任务中取得突破性进展，源于其 R1 模型采用的数学推理数据集，不仅要求答案正确性，更对解题步骤的规范性、逻辑链的完整性提出严格标准，这种精细化的数据设计使得模型在抽象思维能力上实现质的提升。

大模型参数规模指数级增长与多模态能力的拓展，促使数据需求从量级积累转向质量提升。一方面，模型训练需要覆盖更广的知识范畴、更多元的数据场景，这对数据的多样性与代表性提出更高要求。另一方面，大模型从通用能力向垂直领域深度融合时，面临着数据瓶颈的严峻挑战。尤其是医疗、法律、工业等专业领域存在明显的“数据孤岛”现象，领域知识密度高但结构化程度低，且涉及隐私保护与数据安全等问题，高质量的数据集构建成本往往成为技术落地的主要障碍。

数据资源已成为全球人工智能产业竞争的核心战略要素。欧盟于 2022 年通过的《高价值数据集实施法案》已率先在环境、地理空间等关键领域明确了数据开放的标准与规范，推动公共数据的高效流通

与再利用。以 OpenAI 为代表的国际领先企业正通过强化微调等技术手段，依托小规模但高度精准、结构化的高质量数据集，实现大模型在垂直领域的专业化和实用化演进。这种“以质取胜”的数据策略显著提升了模型性能与落地能力。面对全球 AI 竞争的新格局，我国亟需加快构建标准化、合规化、可持续发展的高质量数据供给体系，为大模型技术研发和产业化提供坚实支撑。这不仅是提升国家人工智能核心竞争力的关键环节，也是实现数字经济高质量发展的重要路径。

2. 我国高质量数据集建设进入加速期

在人工智能产业发展浪潮中，高质量数据集建设已成为核心战略方向，从国家顶层设计到地方创新实践，各行各业都在积极探索。

国家层面，多举措陆续完善顶层规划。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）》，强化场景需求牵引，带动数据要素高质量供给、合规高效流通。2024 年，政府工作报告提出开展“人工智能+”行动，从顶层设计层面规划人工智能技术与大模型数据集建设。同年，《关于促进数据产业高质量发展的指导意见》首次明确提出“高质量数据集”，将其作为人工智能与实体经济融合的核心载体，并提出开发行业数据集的具体要求。随后一系列政策相继发布，《关于促进数据标注产业高质量发展的实施意见》《关于促进企业数据资源开发利用的意见》以及《国家数据基础设施建设指引》均提出建设行业“高质量数据集”，由此数据集高质量发展成为行业发展重要目标。2025 年 2 月，国家

数据局组织27个部委召开高质量数据集建设工作启动会，全力推动高质量数据集建设高效赋能行业发展。

地方层面，各地立足区域特色，积极探索高质量数据集建设创新路径，形成了各具特色、协同发展的良好局面。有的出台政策积极鼓励，包括江苏、苏州、贵州、成都、上海、宁波、广东、福建、杭州、河南、山东等地分别从数据集建设、数据质量评价、数据产品开发等多方面建立相互补充、各具特色的政策体系。有的发布打造具有领域特色的行业案例，比如苏州发布30个高质量数据集，覆盖工业制造、交通运输、金融服务等领域。北京国际大数据交易截止目前为大模型提供覆盖32个行业475个数据集，形成覆盖自然语言处理、多模态交互的行业专有高质量数据集体系。

研究层面，大模型企业和科研机构也积极贡献力量，丰富行业数据资源，为人工智能技术的持续创新提供动力。百度发布的百度百科数据集、百度搜索数据集等，凭借其庞大的数据规模和丰富的信息内容，成为研究人员和开发者进行模型训练和算法优化的宝贵资源。阿里巴巴发布的中文问答数据集，为智能问答系统的研发提供了高质量的训练数据。哈工大自然语言处理实验室发布的大规模中文分词、词性标注和命名实体识别数据集，为中文自然语言处理领域的研究提供了重要数据支撑。智源研究院发布的中英双语数据集IndustryCorpus1.0包含3.4TB开源行业预训练数据，覆盖18类行业，为人工智能领域的跨语言研究和应用提供参考。

（二）高质量数据集的概念内涵

《高质量数据集 建设指南》（征求意见稿）中定义**高质量数据集（high-quality dataset）是经过采集、加工等数据处理，可直接用于开发和训练人工智能模型，能有效提升模型性能的数据的集合。**基于高质量数据集是面向人工智能应用的前提条件，那么它与人工智能数据集是什么关系、有哪几类的数据集、怎么才算是高质量，这些都是在建设数据集之前需要探讨清楚的问题，下面将从数据集的概念、数据集的分类、高质量的内涵三方面理清高质量数据集的概念。

1. 数据集的概念

人工智能数据集是指用于训练和开发人工智能模型的数据集合。包含图像、文本、语音等大量标注的数据样本，用于训练人工智能系统识别和学习特征模式。通常一个数据集由四个主要部分构成: 特征、标签、元数据和样本。

特征是数据集的输入变量，它们描述了每个样本的具体属性。标签是数据集的输出变量，是需要预测的目标。元数据提供数据本身的信息，如数据收集的时间、地点、来源等。样本则是单独的一条数据记录，由一组特征向量和对应的标签组成。例如机器学习的经典数据集鸢尾花数据集（Iris Dataset）包含 150 条样本，均匀分为 3 类鸢尾花，每类 50 个样本，以花萼长度、花萼宽度、花瓣长度、花瓣宽度作为分类的核心特征。图像领域的 ImageNet 视觉识别数据集，包含超过 1400 万张高分辨率图像，涵盖 2 万多类别，每张图像标注了类

别标签，以及超 100 万张图像甚至还包含物体边界框的标注信息。

2. 数据集的分类

从数据模态来看，可以分为单模态数据和多模态数据。单模态包括文本、图像、音频、IoT 数据等，多模态数据包括图文数据、视频数据、思维链数据等等类型。单模态数据中，文本数据是非结构化的语言信息，用于自然语言处理的机器翻译、情感分析等场景以及语言模型的训练；图像数据是像素矩阵构成的视觉信息，用于计算机视觉的图像分类、目标检测，医疗影像分析以及自动驾驶等场景；音频数据是声波信号，用于语音识别、音乐生成、工业设备异常检测等场景；IoT（物联网）数据主要是传感器的实时流数据，例如温度、湿度、加速度等，用于设备状态的监控、智慧城市中交通流量的预测等场景。而多模态数据是指两种及以上模态数据的融合，通过模态互补提升模型的鲁棒性，用于图文生成、视频理解等场景。思维链数据则是一种特殊的文本形式，可以是单模态也可以是多模态，主要是通过分步推理解释模型决策，演绎从问题到答案的具体推理步骤，用于数学证明、逻辑谜题等模型的复杂推理，同时也提高人类对模型的信任度。

从流程阶段来看，可以分为预训练数据集、指令微调数据集和评测数据集。预训练数据集是用于大规模无监督或自监督学习的基础数据集，通过让模型从中学习通用特征和知识，为后续任务提供强大的初始参数。它是大模型训练的基石，其核心逻辑是“先通识教育，再专业精修”。其特点是海量、无需标注且来源广泛，包括网页内容、

书籍、学术文献、编程代码、平行语料库、社交媒体和百科全书等。指令微调数据集是用于进一步微调预训练的大语言模型，使模型能够更好地理解和遵循人类指令，从而增强大模型的能力。通常由一系列的问答对组成，问题一般是向大模型发出的请求或指令，答案一般是根据请求生成的响应。评测数据集是为了评估大模型在各种任务的表现，为大模型提供性能测量的标准。通过评测数据集，研发人员可以定量衡量大模型的性能，识别模型优化方向。

从数据应用来看，可以分为通识数据集、行业通识数据集和行业专识数据集。数据集作为开发和训练人工智能模型的重要支撑，不同类型模型所需数据集蕴含的通用知识、行业领域通用知识、行业领域专业知识的内容、范围和数量也不一样。通识、行业通识、行业专识三类高质量数据集，主要是通过数据集的知识内容、来源类型、时效性、标注人员类型、敏感程度、模型类型、主题范围等维度来进行划分。通识数据集包含面向社会公众、无需专业背景即可理解的通用知识，主要用于支撑通用模型落地应用；行业通识数据集包含面向行业从业人员、需要一定专业背景才能理解的行业领域通用知识，主要用于支撑行业模型落地应用；行业专识数据集包含面向特定业务场景相关人员、需要较深的专业背景才能理解的行业领域专业知识，主要用于支撑业务场景模型落地应用。

3. 高质量的内涵

为满足人工智能模型训练和应用的需求，数据集质量评估包含静

态评估和动态评估两种方式。在静态评估方面，一是要扩展规范性、完整性、准确性等传统数据质量“六性”指标的范围，比如数据集建设过程的完整性、标注的规范性、标注的准确性、多模态数据内容的一致性等；二是需新增多样性、真实性、合规性等面向人工智能应用的指标，包括数据领域分布的多样性、数据来源的真实性、数据内容的合规性（例如是否存在数据后门、符合隐私安全）等等。在动态评估方面，高质量数据集应能有效的提升模型性能，因此一是需引入合适的模型进行辅助评估，通过建立基准模型、基准评测数据集以及评估指标，通过基准测试，客观并量化模型性能提升的程度，明确高质量的要求；二是要搭建测试平台，统一评估标准和工具，确保不同数据集之间的公平比较，提高数据集质量的可比性和通用性。在实践方面，中国信通院组织编制技术标准《高质量数据集 数据质量评估方法》明确数据质量评估方法和评估指标，助力数据集高质量判定有据可依。

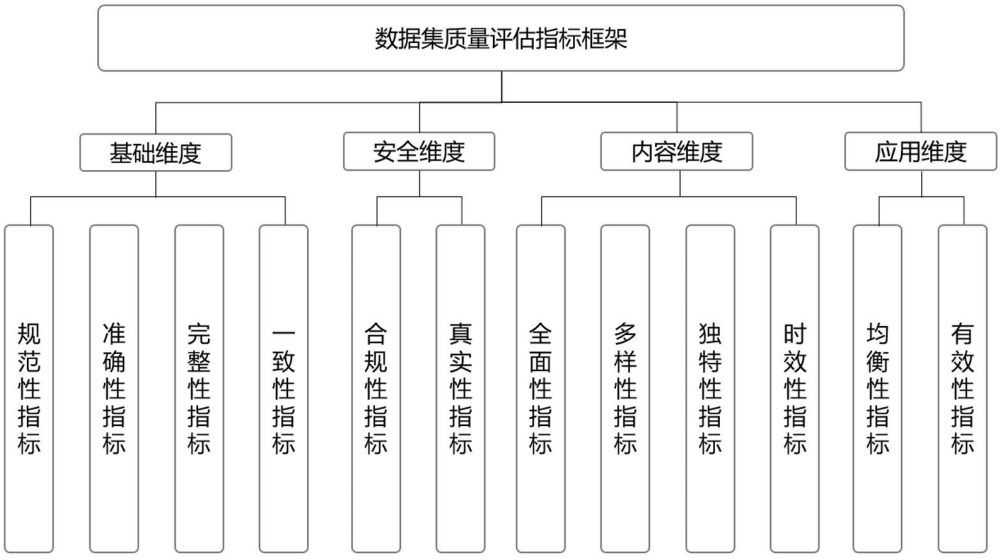


图 1 高质量数据集质量评估指标框架

（三）高质量数据集的关键问题

1. 高质量场景数据集存在较大供需缺口

我国数据资源总量优势明显，多样化数据资源丰富，为高质量场景数据集的高速增长提供现实基础。然而，我国高质量数据集供给能力、机制不足，海量数据与多样化场景优势的潜能仍有待进一步释放。

一是数据资源总量优势尚未转化，通用高质量数据集规模不足。《全国数据资源调查报告（2024 年）》显示全国数据生产量达 41.06 泽字节（ZB），同比增长 25%。但是仍面临开源数据集规模不足、数据处理能力有限、版权问题影响数据获取等问题。例如，英文开源图像-文字对齐数据对规模为 50 亿级对，中文仅为 5 亿级对，存在十倍差异。现有数据集中 85%的数据未深度处理，导致大模型类理解肤浅，易出现幻觉。版权问题尚未形成成熟的解决方案，书籍文献、影视音频等高质量数据处于“想用不敢用”的困境。

二是行业高质量数据集建设处在起步阶段，供需机制尚未形成。行业数据领域性和专业针对性较强，适合与行业大模型深度结合，但目前供给规模有限。大量模型企业虽有丰富数据需求，却常面临“无数据可用”的难题。例如医疗、金融等垂直领域，因数据获取门槛高、专业性强，数据量往往捉襟见肘，且因业务复杂、数据来源多样，数据质量管控难度大，标准化缺失、格式混乱、更新滞后等问题频出，导致难以满足模型高精度优化需求。然后，行业数据存在结构复杂、数据分散、质量不佳、完整性不足等问题，数据集加工难度大，企业

一般不具备专业处理能力。比如，我国数据标注智能化、专业化程度较低，专业数据处理人员队伍数量缺口较大，部分专业数据集无法规模化生产，难以满足专业场景需求；通过人工智能生成多样化合成数据的技术成熟度较低，难以满足大模型训练对于海量、多样化数据的需求。此外，行业数据具有私有性和垄断性，企业共享意愿低，数据流通应用生态缺失。当前公共数据集获取渠道不畅，数据发布格式多样，跨部门、跨地区数据共享程度不足，且《全国数据资源调查报告（2023）》显示，调研的27家交易所的数据产品中仅有17.9%实现交易。而迄今为止，针对高质量数据集的交易更是少之又少，由此也导致了单一企业建设高质量数据集的成本较高。

2. 高质量数据集建设路径缺乏实践指引

高质量数据集的建设包括需求分析、规划设计、数据采集、数据处理（如转换、清洗、合成、标注等）、数据质量检测与评估、数据运营等多个环节，流程长、技术要求高。目前产业还处于探索阶段，在人员、技术、标准、运营等方面存在诸多挑战，需进一步明确路径规划。

一是建设目标不匹配。在数据集需求分析和规划设计阶段，企业容易陷入“唯数据量论”的建设误区，出现业务场景需求与数据集建设目标之间割裂的现象，导致数据集与核心业务指标缺乏深度融合，致使无法有效提升模型性能或场景应用效果。**二是团队协作不高效。**团队需由数据工程师、算法工程师、数据科学家等角色组成，根据科

学的方法论将各个团队、人员、角色进行串联，形成“流水线”式的作业，提升团队协作效率。**三是标准规范不健全。**尽管现在有一些已发布或正在编制的国标、行标和团标，但是可以看到传统的数据治理体系和评价方法对非结构化数据已经完全不适用，缺乏可执行落地的标准规范作为企业实践参考，以及还需优化标准与具体业务场景的适配性。**四是技术工具不满足。**复杂人工智能场景对多模态数据处理能力、非结构化数据处理能力要求较高，传统大数据处理技术表现不佳，难以满足场景实时性、精准性的处理需求，还需进一步技术能力。**五是运营管理机制不成熟。**当前企业在战略层面普遍存在“重建设轻运营”的情况，不论是面向大模型应用还是面向业务场景，数据集的洞察和持续运营优化都是必不可少的关键步骤。

二、高质量数据集建设路径

（一）建设模式

高质量数据集的建设是一个覆盖数据集全生命周期的系统工程，业界通常存在两种典型的建设模式。**第一种模式是“场景驱动”的建设模式。**以明确的业务需求或场景为起点，通过“需求拆解-数据设计-数据采集-数据处理-数据质量检测-数据运营”的闭环，确保数据集对场景的智能化水平提升，避免“数据冗余”或“数据缺失”。**第二种模式是“数据驱动”的建设模式。**以已积累的大量、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。

第一种模式强调“先有需求或场景，再构建对应的数据支撑”，是目标导向型建设的典型代表。这种建设模式的优势是数据质量高、针对性强，能够有效支撑特定任务的模型训练和评估，易于形成闭环反馈机制，通过模型效果反向优化数据采集和处理流程。适用于垂直行业应用、科研实验型项目等场景。例如，在开发一个糖尿病视网膜病变筛查系统时，首先明确诊断目标和模型性能指标，再据此收集具有代表性的医学图像，并由专业医生开展标注，从而构建出具备临床价值的高质量数据集。

第二种模式强调“先有数据资产，再通过数据驱动需求升级”，是过程导向型建设的典型代表。这种建设模式的优势是能快速形成大规模数据资产，为后续模型探索提供丰富素材，一般更适合通用大模型、预训练模型等需要海量多样化数据的任务。适用于数据基础好但应用需求尚未完全明确的场景，包括通用人工智能模型（如大语言模型、多模态模型）训练、数据要素市场建设、政府开放数据平台等场景。

总的来说，“需求先于数据”是目标明确的精准建设，适用于业务方向清晰、需快速落地的场景；“数据先于需求”是数据驱动的价值挖掘，适用于数据积累丰富、探索新增长点的场景。**在实际建设过程中，两种模式相互交替、动态互补。一方面，以“场景驱动”的方式快速构建基础数据集，满足当前业务场景需求；另一方面，通过“数据驱动”的探索，挖掘数据中的潜在价值，为需求的进一步升级提供支持。最终形成了“场景牵引、数据赋能、价值反哺”的螺旋式发展**

动态。

（二）核心环节

为解决高质量数据集建设方法论缺失的问题，中国信通院联合40多家单位编制技术标准《人工智能数据工程能力要求》，涵盖AI数据的研发、交付、运维和运营等环节，涉及AI数据采集、AI数据处理、AI数据标注、AI数据合成、AI数据增强等多种技术能力，同时对全链路的数据质量和合规性管理提出了标准化规范。

依托此标准和业界实践经验，**高质量数据集建设需关注研发管理、交付管理、运维管理、运营管理4大核心环节**。这些环节构成了数据集从无到有、持续优化的闭环。在具体的实施过程中，各环节根据不同业务场景和建设需求顺序可调整，也可选择不执行。与此同时，高质量数据集建设的顺利推行，离不开相应技术工具的支撑。其中**关键的7项技术能力，包括数据采集、数据处理、数据管理、数据标注、数据合成、数据质检、数据服务运营等技术**，这些技术能力可以集成在一个或多个技术工具中。基于这些技术工具可以为高质量数据集的开发和运营提供完善的支撑。

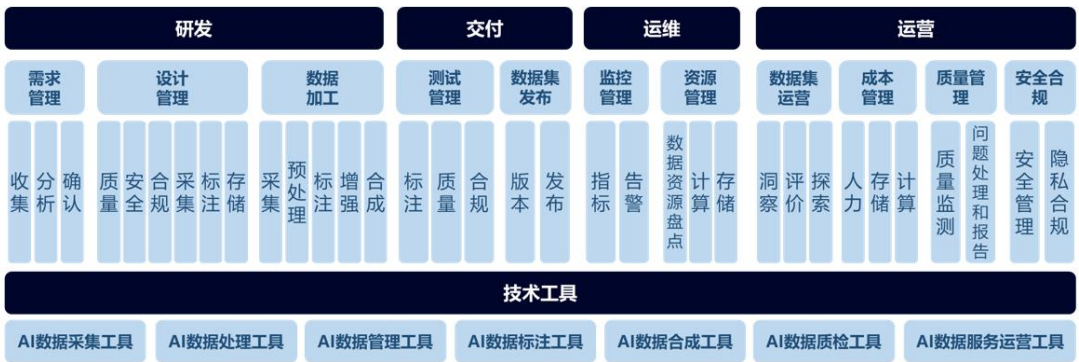


图 2 高质量数据集建设核心环节

1. 研发管理

研发管理是对数据集的生成流程进行管控，覆盖需求管理、设计管理和数据加工三个环节。需求管理是通过精准捕捉、分析和控制数据需求，确保数据集建设与 AI 模型目标严格对齐，即明确人工智能团队和业务部门对数据集有哪些需求，规范化需求的描述，收集需求并对需求进行分析，确认其优先级和合理性。设计管理是构建数据集标准、质量、安全、合规、采集、标注、存储的规范体系，先立规矩再开展具体的开发行为。数据加工管理是梳理数据集加工的整体流程（包括数据采集、预处理、标注、增强、生成等），明确对应的技术能力和管理要求。

2. 交付管理

交付管理是面向数据集的交付过程开展管控活动，主要有测试管理和发布管理两个环节。数据开发完成后需要对标注质量、数据集质量以及数据的伦理和合规性进行全方位测试，以保证开发完成的数据符合合规性、数据质量、AI 场景下的可用性等要求。测试管理是对数据集上线前的质量进行把关，开展质量和合规性的验证，包括标注测试、质量验证、伦理和合规性检测等。发布管理是建立发布体系（包含发布审批、API/接口管理、数据集管理等）将经过验证的数据集安全、高效、规范地转化为生产级服务，并对发布后的数据集版本变更进行管理，实现规范化记录、追踪和控制，保障数据在长期演化和协作过程中产生的可追溯性、一致性和可复现性问题。

3. 运维管理

运维管理是关注数据集的日常监控和维护，对过程中涉及的数据、计算、存储资源进行管理。其中监控管理需要对数据集质量、系统性能、安全合规等方向建立监控指标，开展日常的监控和告警活动，并开展应急处置。资源管理需要对数据集进行数据资源盘点和管理，对过程中涉及的计算和存储资源进行管控。数据资源盘点是摸清家底、激活价值的基础性治理，通过系统化梳理数据资产目录，破除“数据黑盒”状态。计算资源管理核心在于平衡效能与成本，通过优化任务调度策略、资源分配规则与弹性伸缩机制，在保证任务的前提下最大化集群利用率。存储资源管理是在成本、性能、可靠性三角中寻求平衡，通过分级存储策略、生命周期规则、压缩优化技术等实现存储成本的精细化控制。

4. 运营管理

运营活动关注数据集在用户端的使用情况，衡量投入产出的收益，关注数据集长期的质量提升。数据集运营旨在为用户提供数据探索和分析的入口，提供评价和持续迭代优化的机制，包括提供数据探索入口、提供数据集持续迭代优化的机制、提供数据集的探索分析工具等。成本管理涉及人力成本、存算资源成本和技术工具成本管理，核心在于设计成本的计算方法、预算制定、成本控制方法、内部结算方法等。质量管理、安全管理和隐私合规管理旨在提供一套持续的问题监测、报告、处理的机制和应急预案。

（三）成效评估

企业在构建和运营高质量数据集的过程中面临诸多困扰，诸如能力要求模糊不清、缺乏有效的评价标准、管理体系不够完善以及提升方向不明等问题。为助力企业精准衡量并定位自身在高质量数据集建设与运营方面的行业水准，中国信通院组织编制技术标准《高质量数据集建设运营能力成熟度评估模型》。基于理论研究与产业实践，总结出高质量数据集建设运营成效的评估方法，从组织管理、技术服务、数据安全、标准规范、运营管理、生态建设六大能力域进行能力拆解，并将能力成熟度划分为初始级（1 级）、可控级（2 级）、规范级（3 级）、优秀级（4 级）、卓越级（5 级）共 5 个等级。

图 3 高质量数据集建设运营能力成熟度评估模型

注：关于具体评估指标欢迎读者与工作组联系探讨，联系人：白玉真（18810275013）。

1. 评估框架

完善的组织管理体系是高质量数据集高效建设的前提条件。组织架构方面，高效的组织架构能够确保资源合理分配、职责明确、沟通顺畅，从而支撑组织的战略目标和业务发展。**项目架构方面，**合理的项目架构能够确保项目顺利进行，提高项目成功率，降低项目风险。通过建立明确清晰的项目架构、详细的规划设计，指导项目实施。建立完善的项目管理机制，保障流程顺畅，协作紧密。**项目管理方面，**高效的项目管理能够确保项目按时按质完成，提高项目成功率，降低项目成本。提高项目管理团队专业能力、执行效率高，并定期审核数据集建设成效，注重项目管理的持续改进与创新。**制度体系方面，**建立与数据集建设相关的制度体系及优化机制，以保障数据集建设职责分工明确、合规有序推进、数据质量稳定、数据风险可控、建设效能提升。

技术服务是高质量数据集高效、自动化及智能化建设的核心能力。数据采集阶段，明确数据采集的步骤和方法，根据数据来源选择合适的自动化和智能化工具，量化分析数据采集过程，构建数据资源地图、洞察数据采集方向。**数据处理阶段，**首先依托数据清洗对数据进行识别、处理错误值与缺失值、检测与删除重复记录等，遵循标准流程、优化策略、借助先进技术并实现智能化、自动化操作；然后通过数据标注、合成、增强等技术或工具完善数据的特征表示，以提升数据质量与可用性的功能组件。**质量检测与评估阶段，**根据数据集全生命周期的需求，建立系统化、可量化的数据质量检测与评估机制，以保障

数据集满足预期的使用要求和可信性。通过完善的数据质量指标体系、质量评估流程与评估工具，确保数据集建设的高质量水平。

数据安全是高质量数据集建设的生命线，需贯穿于数据采集、存储、传输、使用和销毁的全生命周期。一是数据采集安全。确保所采集数据的合规性、可靠性和可用性。建立完整的数据安全分类分级体系、数据安全审计体系以及采用智能化的工具提高效率。**二是数据存储安全。**保障数据在存储过程中不被未经授权的访问、篡改或删除。实施差异化的安全防护策略，构建多层次存储架构，采用先进的存储技术和方法，如分布式存储、数据冗余等，提高存储的安全性和可靠性。**三是数据传输安全。**保护数据的机密性和完整性，确保数据在传输过程中的安全。建立完善的传输安全政策，采用先进的加密技术和算法对数据进行保护，以及形成监控和审计机制，及时发现和处理异常。**四是数据使用安全。**建立完善的数据使用技术体系，数据使用规范，采用合适的智能化数据使用监测技术，建立数据使用风险预警和应急响应机制，以及监控和审计制度，保证各方如约使用数据。**五是数据销毁安全。**当数据不再需要或达到保存期限时，能够被及时、安全、彻底地删除或销毁。建立完善的销毁策略，采取先进的技术手段和方法，提高销毁的效率和安全性，并保证销毁过程记录完整能够进行合规性审查。

标准规范是实现高质量数据集规范化建设的必要条件。一是构建标准体系，在国家、行业、团体等标准指导下建立标准体系以保障数据集建设过程各环节的合规性、科学性和一致性。包括构建数据集数

据质量评估方法、数据集开发管理平台技术要求、数据标注平台技术要求、数据标注指南等内容。**二是标准实施管理，**建立监督检查机制，以确保数据集建设过程遵循标准，通过效果量化指标，提升实施效率与质量。**三是构建标准持续改进机制**，以确保相关标准能适应不断变化的法规要求、业务环境与技术发展，提升标准的科学性、适用性和有效性，从而推动企业在数据集建设中的规范化发展。

运营管理是高质量数据集推广应用的有效保障。一是建立运营管理体系。完善运营部门统筹能力和管理体系，采用管理平台对数据集的关键性能指标开展实时监测，运营管理工作与业务战略紧密结合。**二是搭建运维能力**。完善数据集运维机制，建立全面数据监控体系，运维团队使用自动化、智能化工具，保障复杂场景的自动化运维。**三是运营能力。**制定合理的运营计划，建立数据集运营监控指标评价体系，并能结合业务反馈及前沿技术发展趋势，对数据集进行内容迭代、功能优化或结构升级。开展数据集的联合运营和价值共创，拓展数据应用的边界和市场影响力。

生态建设是高质量数据集健康快速发展的持续动力。一是聚焦应用场景。数据集的建设需面向应用场景需求，持续验证和优化数据集，使其在特定的应用中具有高效性、可靠性和准确性。根据场景的需求变更，快速调整数据集结构和内容，确保在场景中的优异表现。构建最佳实践或典型案例库，形成可复用可参考的实施模板。**二是构建产业生态。**企业通过与产业链上下游各方的协作，构建开放、健康且可持续的数据生态系统，促进数据集的共建、共享和应用。在数据集建

设过程中具备开放、透明、可持续的生态合作机制，能够支撑数据集在跨行业、跨领域合作中实现广泛的共享和应用。**三是完善生态运营。**建立比较完善的数据集生态管理机制和运营流程规范，具备专门的生态运营团队、搭建统一的生态服务平台，建立高质量数据集生态健康度监测体系，实现多方的广泛认可和高效协同。

2. 评估等级

根据企业在各个能力域上综合表现程度，可以划分 5 个等级。

初始级（1 级）：数据集建设能力尚处于较为原始的状态，缺乏系统性和规范性。具体表现为组织管理模糊、数据建设过程未规范化、缺少必要的流程和工具、数据集质量依赖人工检查、数据安全保护缺失、数据集运营和生态未形成。

可控级（2 级）：数据集建设能力比较可控，并采取了一定的措施进行改进。具体表现为建立组织管理体系，但缺乏完善协作机制；制定建设标准和流程，但规范性不高；关键节点实现了半自动化工具，但仍需人工参与；具备基础的数据安全保护措施；初步建立数据集监控和管理机制；数据集生态环境逐渐开放但应用局限。

规范级（3 级）：数据集建设能力已经相对成熟，形成了较为完善的流程和规范。具体表现为建立较为完善的组织管理体系，形成覆盖数据集全生命周期的标准化流程，数据集建设的流程标准化；关键节点引入自动化工具，并能满足一定的扩展性和灵活性；形成较完善的数据安全保护体系；建立明确的数据集监控、更新和反馈机制，完

善的运营管理体系；开始关注数据集生态建设，生态开放活跃，数据流动性提高。

优秀级（4 级）：数据集建设能力已经达到了较高的水平，形成了较为完善的生态。具体表现为建立了完善的组织管理体系；技术服务体系完善，平台和工具的集成度高、自动化水平高；数据安全管理体系成熟；建立数据集投资回报评估体系，实现数据集建设全流程的精细化管理，且有持续改进的机制；数据集的共享与开放体系非常成熟，产业链各方形成了较为完善的协同创新机制，主导多方参与的数据集生态。

卓越级（5 级）：数据集建设能力已经达到了行业领先地位，形成了具有核心竞争力的能力体系。具体表现为：组织管理体系成为行业标杆，具有高度灵活性和适应性；主导国家、行业、国际标准，技术服务方面建立行业最佳实践，探索前沿技术创新应用；实现数据生命周期的智能化、精细化管理；实现数据资产化运营；数据集在跨行业、跨领域的合作中实现广泛的共享和应用；构建开放的数据生态平台，吸引上下游合作伙伴共建数据应用生态；具备强大的国际竞争力，能够引领行业变革和发展。

三、高质量数据集应用场景

（一）场景概述

日前国家数据局正在组织开展高质量数据集典型案例征集，为高质量数据集的加速建设注入了一剂强心针。重点面向科学研究、工业制造、农业农村、智慧能源、交通运输、金融服务、医疗卫生、教育教学、商务领域、人力资源、文化旅游、应急管理、气象服务、绿色低碳、公共安全、城市治理、低空经济、具身智能、智能驾驶、智慧海洋等 20 多个行业和领域开展征集。

本报告通过对已公开发布的高质量数据集名单、数据交易所上架的数据集产品、开源数据集等内容梳理，从供给角度和应用领域角度分析高质量数据集的应用情况。

从供给角度来看，高质量数据集大多集中于开源社区、数据交易所、数据服务企业以及数据标注基地。魔搭、飞桨、天池、帕依提提、超神经、智源、和鲸、启智、聚数力等社区平台提供多类型公开数据集，适用于基础的人工智能模型任务。北京、深圳、贵阳等大数据交易所陆续建立高质量数据集专区，汇聚多模态优质训练数据。以海天瑞声、数据堂等为代表的数据服务商在原有业务的基础上进一步拓展丰富高质量数据集产品。四川成都、辽宁沈阳等 7 个标注基地目前也已形成高质量数据集上百个。

从应用领域来看，高质量数据集应用当前在工业制造、医疗卫生、交通运输领域较为集中。其次，低空经济、具身智能等创新应用也因

产业发展驱动陆续涌现。下面以部分场景实践案例对高质量数据集的建设过程进行简要阐述，为业界提供参考。

（二）实践案例

1. 工业制造场景

工业制造场景数据集源于采集端存在大量设备采集的图片、音频、视频等非结构化数据，多应用于企业数字化转型的智能化场景，如生产制造、故障诊断、智能运维、设备状态监测等。

专栏 1：紧固件失效案例数据集

案例背景：

航天产品保证是以技术风险识别与控制为核心，在航天器研制全过程进行的一系列有组织、有计划的技术和管理活动。以紧固件为代表的基础产品在装配及使用阶段必须满足航天产品及配套系统的安全、可用、可靠的要求，紧固件的选用控制除考虑其选用目录外，还应参考其失效案例，以历史经验教训为基础识别材料缺陷等风险薄弱环节，采取措施及时消减风险，提升任务成功率。

实践方案：

以紧固件为代表的航天基础产品保证数据集分为行业通识数据集和行业专识数据集两部分内容。行业通识数据集通过设置紧固件相关关键字使用爬虫在互联网数据集中获取，包括行业研究报告、标准规范、国内外专业书籍及手册、论文和专利等公开数据，用于在大模型中建立紧固件等基础产品相关的行业基本认知能力；行业专识数据集的数据来源有厂商研制过程中、紧固件产品试验验证阶段以及紧固件产品在轨服役使用阶段所发生的失效案例的记录，这部分数据涉及紧固件产品核心知识，需要采取恰当的数据安全防护手段来控制知悉范围。

紧固件产品保证的数据主要以非结构文本、参数表格及产品图片等形式存在。在数据处理阶段，采用 MinerU 等文本处理工具对数据进行格式转换，以行业标准 QJ 3050A-2011《航天产品故障模式、影响及危害性分析指南》为依据，对紧固件产品的故障模式、故障原因、故障影响、处置措施等数据特征进行总结，建立标签体系。以大模型的提示工程技术为代表方法，对紧固件相关的失效数据进行信息提取及标注，并同时辅以专家校验，保证数据的高质量。经处理后形成结构化的紧固件产品故障模式及影响分析表。

为了进一步分析紧固件失效产品间同材料、同阶段、同类别等关联关系，将紧固件 FMEA 表数据组织成失效案例知识图谱的形式。除此之外，利用图检索增强技术对新增的专业手册等非结构化进行处理，自动提取标签体系中的实体和关系，连接上下文的文本块。在数据存储方面，采用了 Neo4J 图数

据库、Milvus 向量数据库和 MongoDB 非结构化文本数据库三库并行的方式对图数据、向量数据和文本数据分别进行存储应用。

最后，将专家图谱与图检索增强生成的图谱进行实体对齐与知识融合，作为紧固件专业问答和选型的支持和依据，增加失效案例警示等数据标签，对上装的相似紧固件产品推荐可能发生的风险点及处置措施。

应用效果:

紧固件失效案例数据集包括国内外典型案例 300 个以上，图谱化节点规模达上千规模，有效应用在航天基础产品多维度场景中。

紧固件智能问答模型：融合紧固件失效案例知识图谱的紧固件智能问答模型准确性达到 95%以上，有效提升问答的准确性和专业性。

紧固件风险评估模型：以紧固件失效案例集为主要依据，通过分析紧固件连接结构在各个阶段存在的风险点进行分析，评估其风险等级，风险点覆盖率提升 20%以上。

紧固件失效分析功能模型：紧固件失效分析是紧固件研究的重要分支，用户主要包括专业紧固件失效分析机构（如实验室）和紧固件使用单位故障归零，而紧固件失效案例可以帮助失效分析专家开展紧固件的失效分析，推荐类似的失效案例，专家可以根据推荐案例设计失效分析试验，大大提升了失效分析的效率，减少技术归零所需的时间 10%以上。

紧固件智能选型模型：紧固件失效案例作为紧固件智能选型的重要支撑材料，可为设计师提供设计方案预警，避免有问题设计重复出现，并为紧固件预测性维护提供有效知识支撑，提升航天产品可靠性水平达 40%以上。

2. 医疗卫生场景

医疗卫生场景数据集是在传统专病数据集的基础上，结合影像数据、临床科研数据、医学术语、药品说明书等多模态数据，面向大模型应用的特定需求，进一步加工形成，用于辅助诊疗、辅助决策场景。

专栏 2：医疗健康数据集

案例背景:

随着人工智能在医疗健康领域的深入应用，高质量数据集成为推动 AI 模型发展和提升医疗服务水平的关键。中国联通致力于构建医疗健康行业高质量数据集，以满足辅助诊断、药物研发和智能监管等应用场景的需求。此项建设时间为 2024 年 2 月至 2024 年 12 月。

实践方案:

中国联通通过与多家顶尖医疗机构和国家药监机构合作，构建了多模态医疗健康高质量数据集，并在数据治理、安全合规、标注自动化和数据增强等方面形成了一套完整的实践方案 。

（1）数据集建设与来源:

中国联通建设了四大类高质量数据集，总规模达到 100TB 。

胸部 CT 影像数据集：联合北京胸科医院，完成 20000 余例影像标注，

用于辅助肺结核专病判定模型 。

耳部 CT 影像数据集: 联合全国知名耳鼻喉专科医院，完成 5000 余例高质量影像标注，用于早期听觉障碍及耳部异常智能识别模型 。

肾脏病慢病管理干预与临床科研数据集: 依托北京大学第一医院，完成 10000 余例患者全周期数据标注，用于智能化慢病动态管理及预警模型 。

药品说明书数据集: 联合国家药监机构，完成超过 58000 份药品说明书的文本与图像精准标注，用于智能解析与比对模型 。

各类数据均通过合法授权渠道获取，确保来源可溯、授权完整，并严格遵循《网络安全法》、《个人信息保护法》和《数据安全法》等国家法规进行去标识化处理，保护个人隐私 。

（2）数据治理与质量控制:

中国联通构建了从“领域问题、本体设定、知识规则、专家语义、论证评估、生产验证、入表入库”的全流程医疗数据治理工作思路 。

数据标准:联合专家团队制定 17 类胸部 CT 影像征象标签和标注标准 。

质量保障: 采用双盲标注方式核查数据标注一致性，一致性评估结果超过 95%; 通过专家抽样(10%)审核方式核查标注结果准确性达到 98%以上 。对于耳部 CT 影像，一级审核（自动预标注）和二级审核（主任审核）结果准确性达到 95%以上 。

安全合规: 建立数据沙盒监管机制，实现“零信任”数据安全合规体系，确保原始数据不出域、数据可用不可见。数据脱敏覆盖率达到 100% 。

（3）数据标注自动化与效率提升:

为了满足医疗行业数据标注的高效性和准确性需求，中国联通整合了文本、图像、语音、视频等不同模态的数据标注工具，并基于机器学习算法实现了自动化标注 。

影像标注: 通过开发标签阈值自动化适配工具、自动插值方法和预标注算法，将胸部 CT 影像标注效率从原始的 3-4 小时/例提升到 10-20 分钟/例 。

文本标注: 构建包含丰富医学术语的词典和规则库，并利用深度学习模型（如 BERT）进行语义分析，提高标注的准确性和一致性.

（4）数据增强技术:

针对医疗健康数据，通过数据增广、特征工程、去伪影等技术手段，输出更大规模、更多维度、更高价值的数据集 。

数据增广: 提供图像增强（旋转、缩放、翻转）、文本数据增量（同义词替换、数据扩充）、数值数据扰动等能力 。

特征工程: 基于 NLP 技术和多模态识别技术提取药品说明书文本和影像数据特征，优化模型输入 。

去伪影: 采用去伪影算法提升存在运动伪影、金属伪影的数据质量。

应用效果:

高质量数据集的构建直接支撑了多项 AI 模型的研发和应用，并取得了显著成效 。

胸部 CT 影像辅助筛查与诊断的肺结核疾病模型: 研发的辅助肺结核专病判定模型准确率超过 99.99% 。该模型填补了我国当前缺乏肺结核 CT 辅助检测类 AI 模型的空白，可用于医疗机构集中筛查、结核病防控机构主动筛查、区域“影像云”平台筛查等多种场景，有效提升结核病早期诊断水平与防控效果 。

慢性肾脏病早期筛查与干预专科模型：研发的智能化慢病动态管理及预警模型准确率突破 99.9% 。该模型旨在提高基层医疗机构慢性肾脏病诊疗能力，将防控关口前移，并通过生成式 AI 提供诊断建议和个性化综合诊疗意见 。

耳部疾病多模态辅助诊断模型：研发的早期听觉障碍及耳部异常智能识别模型准确率达到 99.95%以上 。该模型通过深度学习算法和影像组学技术，实现耳部病变的自动检出、解剖参数智能测量及病程动态追踪，显著提升诊断效率与一致性 。

药品知识库智能检索问答应用：基于国家药监局 5.8 万份药品说明书构建知识图谱，通过大模型解析能力实现说明书 18 类实体精准提取，准确率达 98.7% 。该应用为公众提供药品用法、用量、药物禁忌等智能服务，并为医疗医生提供合理用药建议和指引，助力构建“研发-生产-流通-使用”各环节的智能管理体系。

3. 交通运输场景

交通运输场景数据集是通过对车载传感器、路侧设备、卫星定位、交通管理系统等多源数据进行加工标注而形成，以智能驾驶、智慧交通场景为主。

专栏 3：交通基础设施多模态三维构件数据集

案例背景：

传统交通基础设施建管养模式正经历智能化重构。虽然 BIM、神经辐射场（NeRF）、高斯溅射（GS）等 AI 技术已实现场景应用，但基于深度学习的交通基础设施三维对象语义解析能力仍是大模型技术链的关键缺口，严重制约行业智能化升级。为此，构建具备多源异构数据融合能力的交通基础设施多模态三维数据集成为破局核心。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。

实践方案：

本数据集构建遵循了顶层规划、数据采集、数据预处理和数据清洗、数据标注和质量控制的科学方法论，旨在确保数据集的高质量与高可用性，有效支撑行业智能化发展。

数据集建设初期即采用先进分层存储架构进行顶层设计，高效管理点云、图像、文本等异构数据。此架构具体分层为：底层依靠对象存储与关系型数据库进行稳健数据管理；中层负责数据清洗、点云和图像数据导出、格式转换及基于大模型的文本合成标注等核心处理；顶层则提供标准化接口，支持 BIM 平台集成与 AI 模型训练。

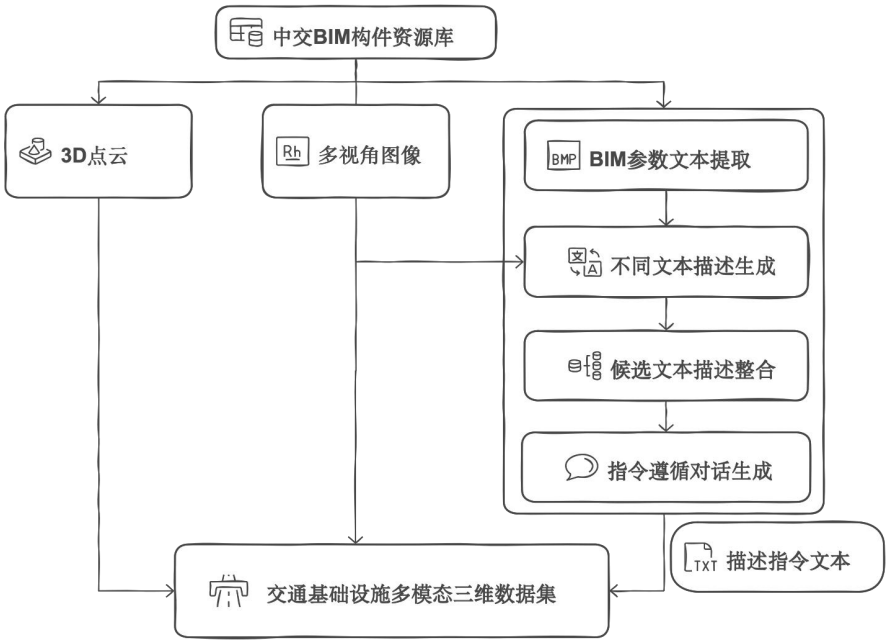
原始三维数据的获取遵循多元化与高标准原则。数据来源广泛整合了集团内部承建的重大工程项目（例如 CZ 铁路、平陆运河项目）所产生的 BIM 模型，并辅以外部专业构件资源库的采购以及合作伙伴的共享资源。在数据

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，依托中国交建在公路、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、8 个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对构件类型、材料属性等核心参数已经完成了精准标注，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。

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，是交通基建行业内规模最大的同类多模态数据集。该数据集以通用、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。



应用效果:

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，以及 CZ 铁路、平陆运河等项目，显著提升了模型建设效率，达到 60% 的提升率。

此外，基于该数据集的研究应用也取得了显著成果。一方面， 通过深度学习的 BIM 构件自动审核技术，精准分析 BIM 构件的图像特征，实现了对构件库资源的自动化审查，准确率高达 96.76%。与传统人工审核相比，在成本控制与效率提升方面均取得了指数级的进步。另一方面，中交 BIM 平台通过深度整合此数据集，并引入大模型、交互式建模等前沿技术，实现了 BIM 模型的快速创建，设计效率提升超过 30%，成果质量提升超过 20%，同时项目成本也随之降低了 30%以上。相关算法成果已发表于 EI/SCI 学术论文两篇，系统建设获得国家发明专利和 PCT 国际发明专利两项。

未来，该数据集以“资源-技术-生态”循环，有望支持施工运维深度优化，辅助构建全息场景支持自动驾驶与无人船舶，驱动车船路协同，增强机器人感知交互，并快速构建数字孪生，实现项目智能管理，助力交通基建数字化转型。

4. 低空经济场景

低空经济场景数据集通过对海量飞行数据、物流数据、地理信息数据等的分析挖掘，可以优化低空空域管理、提高飞行安全性、提升物流配送效率，为低空经济快速发展提供有力支撑。

专栏 4：低空经济场景数据集

案例背景：

四川省某自治州地处高原偏远地区，面积广阔、海拔高，自然环境恶劣。区域内生态保护、电力通信巡检、应急救援及旅游安全管理等任务复杂且由多部门分管。传统人工巡查方式受限于环境和人力资源，效率低、成本高、信息分散。为提升各部门作业效率，基于固定翼与多旋翼无人机实施统一基础数据采集与分部门专用数据采集，并通过构建高质量、可共享的数据集，依托人工智能技术实现了人工作业的无人化、智能化替代。

实践方案：

结合该地区地理环境特点与各部门业务需求，采取了数据集建设与智能模型开发一体的实施路径。整个实践过程分五个步骤推进，形成“统一数据采集+专有数据补充+数据标准化治理+高质量数据集模型应用+高质量数据集共享机制建设”的闭环流程。

（1）统一基础数据采集

考虑到该地区面积大、海拔高、地形复杂且环境恶劣，传统地面巡检方式效率低、风险高，因此项目率先部署固定翼无人机对全区开展高覆盖率的航测任务。无人机配备多任务载荷，可同时采集可见光影像、红外影像和高精度位置信息等，用于提取高程、坡度、坡向、水体分布、植被覆盖指数等关键地貌要素。作业累计完成航测面积约 1500 平方公里，采集地形数据共计 9.59TB，主要包括可见光影像 9.02TB，红外影像 0.1TB，激光雷达原始

点云 0.45TB，飞行记录数据（飞行轨迹、IMU、设备信息等）0.02TB。所有飞行数据回传至地面站后，经初步处理后进入后端存储，为后续各类专项任务提供一致的地理信息基础。

（2）专用数据采集与样本扩充

由于旋翼无人机具备高度灵活性，能够在复杂地形中低空悬停作业，各部门可根据自身业务需求，部署多套旋翼无人机开展针对性数据采集。采集对象包括：变电站、光缆、铁塔等设施状态；植被变化、栖息地破坏、水体污染等环境样本；滑坡泥石流等灾害隐患点；游客违规穿越行为；警示设施损毁情况等。当前已汇聚各类数据共计 1.61TB，主要包括可见光影像 1.39TB，红外影像 0.03TB，原始点云 0.18TB，飞行记录数据 0.01TB。这些数据将与统一基础数据融合使用，用于支撑各类业务场景。

（3）数据治理与标准化处理

在完成数据汇聚后，项目团队针对基础数据和专用数据开展了系统化的数据治理工作，重点解决原始数据异构、质量不一等问题，统一进行坐标系转换、冗余去重、噪点剔除、影像增强、图像拼接、三维建模、高程分析与图像标注等处理。同时，建立标准数据格式体系，并采用统一元数据结构记录采集设备参数等其它附属信息。经处理后，形成高质量的基础/专有数据集 4.74TB，主要包括正射影像 1.53TB、数字表面模型 0.82TB、数字高程模型 0.75TB、处理后点云 0.39TB、三维网格/模型 0.51TB 和各类业务对象（铁塔、灾害隐患点等）图像数据集 0.74TB。其中基础数据集为各部门任务提供统一基础数据支撑，专用数据集为专项模型功能定向优化和精度提升提供支持。

（4）高质量数据集模型应用

基于统一基础数据，项目构建地区通用的基础模型，涵盖地貌分类、道路识别、植被覆盖分析等基础智能任务。随后，各部门依据自身高质量专用数据对基础模型进行迁移学习，快速形成适配自身业务的专用模型。例如，引入变电设施图像样本，微调后形成隐患检测模型；基于植被信息构建退化监测模型；基于行为图像，训练违规识别模型。这种“基础训练—专用特化”的模型体系大幅提升了数据集的利用率，降低了训练成本，也提高了模型精度和落地适应性。

（5）高质量数据集成果共享机制建设

为打通各部门间的数据壁垒，项目建立统一数据共享平台，设计灵活可信的数据流通机制。所有基础高质量数据集对各单位开放，专有高质量数据集则根据授权授信使用。平台引入可信数据空间技术，控制和记录数据的生成、共享、使用与销毁全过程，确保数据可信任、过程可监控和责任可追溯。同时，设立数据评估与持续优化机制，结合现场人工巡检结果，对数据进行周期性修正，提升长期适应能力。

通过以上五个步骤，实现了“基础数据统一、部门数据专有、多源数据融合”的数据集建设体系，切实缓解了高海拔复杂环境下的人工作业难题与资源重复投入问题，同时为跨部门智能化业务应用提供了坚实数据支撑。

应用效果:

统一的地形地貌高质量基础数据集和覆盖多个部门核心任务的高质量专用数据集，实现了“数据共享+模型定制”的智能化任务支撑体系。高质量数据集成果通过共享平台进行统一管理与可信使用，打破了原有“各自为政、重复采集”的局面，实现了跨部门的数据流通与智能协同，大幅提升作业效

率与响应能力。整体来看，数据采集成本降低约 45%，人工巡检人力投入减少近 60%，多部门信息汇聚效率提升超过 70%，数据复用率提升至 82%，数据处理周期缩短约 40%。以电力巡检为例，其通过模型实时识别光缆线路异常，故障响应时间较以往缩短 60%；生态监测则首次实现大范围林地变动趋势预警，植被异常检测准确率达 91%，有效支撑生态修复工作。此外，可持续的数据采集迭代机制，为后续在其它省市复制推广高质量数据集在低空经济场景下的应用提供了范式样板，标志着地区管理工作迈向数智化新阶段。

5. 其他应用场景

随着人工智能技术在应急管理、文化旅游、金融服务、自动驾驶等场景的应用深入，高质量数据集的建设需求也逐渐凸显。

专栏 5：视联网应急管理数据集

案例背景：

中国铁塔视联网业务飞速发展，超 22 万座“通信塔”升级为“数字塔”，在铁塔上搭载摄像头、无人机等设备与算法为应急管理、农业农村等 10 多个行业注智赋能。视联网应急管理数据集围绕综合风险监测预警能、指挥调度能力、自然灾害数据汇聚能力，实现数据全生命周期的管理，为应急行业大模型与小模型算法优化训练提供强有力的数据支撑。

实践方案：

中国铁塔采用标准引领，汇聚多源样本数据，构建 AI 数据处理工具链，以高价值业务场景与需求为导向，构建了视联网应急管理数据集。

样本数据采集与汇聚，汇聚全国中高点位视频监控样本数据。通过标准化的回传通道实现全国分布式视联平台样本数据的自动汇聚，不断提高样本数量；通过定制化采集上传稀缺性的样本，提供样本的丰富度。

需求分析与规划，基于算法训练优化需求检索数据，构建初步数据集。根据森林防火监控、自然灾害防治等应急管理关键业务场景和算法训练需求建立了 300 多个标签对样本数据进行自动化打标，通过语义检索、以图搜图和标签精细化检索等方式按需筛选所需样本数据，形成初步数据集。

数据多维度处理，提高数据集质量。在多环节对数据进行去重、去模糊等清洗操作，去除无效样本数据，提高有效样本数据的占比，支持对清洗后的数据进行恢复操作。按需通过数据增强，增加样本的丰富度。通过多模态数据对齐，建立“图像+拍摄时间+位置+PTZ+影像内外参数”等多维信息关联关系，形成图文数据对，实现应急管理中基于视频监控与算法告警的精准应急调度指挥。

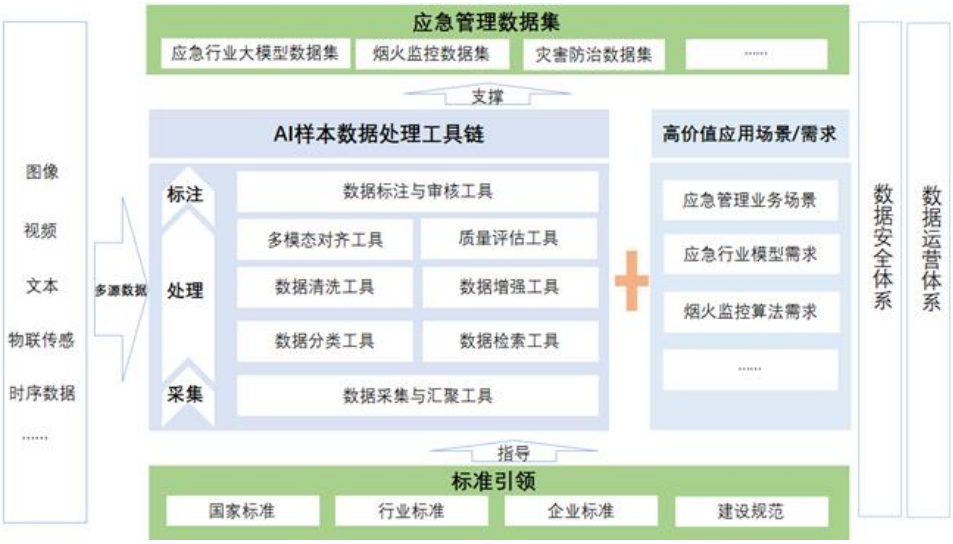
数据智能标注与审核，形成高质量数据集。依托铁塔样本库的智能标注工具快速完成数据标注，标注准确性超过 80%，自主设置抽检比例对标注结果进行人工审核和修正，形成算法训练所需的高质量数据集。

数据质量检测与评估，算法优化训练验证数据集质量，按需调整数据集。通过算法模型效果对数据集进行综合评价和反馈，并对数据集进行适当调整以符合模型预期。对数据集的调整包括重新评估数据需求，调整数据处理策

略，增加特定类别样本的采集比例等方式。

数据集运营，实现数据复利效应。洞察数据集规模、分布等构成，依托样本库系统实现数据集的可视、可管、可用、可追溯，维护数据集版本与上下架，实现数据集的全生命周期管理。根据算法训练优化需求灵活复用不同数据集，实现一次治理多方使用。

数据安全管控，贯穿数据集建设全过程。根据国家和公司数据安全管控办法，在技术上主要采用授权访问安全等级、数据沙箱安全审计、网络安全策略、数据加密与脱敏处理等手段；对于内外部使用数据制定规范制度，保证数据合法合规使用。



应用效果:

中国铁塔样本库以汇聚 4.86 亿中高点视频监控高质量视图样本数据，形成超过 25 亿个有效标签,为应急行业大模型构建超过 500 万的标注图文对，构建了一批服务于应急管理的烟火监控、火点识别等小模型算法的数据集。其中，自研应急行业大模型的 35 个下游任务在实际应用场景中准确率超过 96%；应用于中高点位复杂场景的烟火监控算法泛化能力显著增强，有效规避云雾、扬尘、灯光等目标干扰，准确率超过 95%，对应急管理事件实现了及时发现、准确告警。应急管理数据集除了应用于中国铁塔算法自主研发，通过安全管控手段还用于算法生态合作伙伴的 AI 算法研发，服务于与北京大学、清华大学、中国科学院等高校的产学研合作项目，实现了样本数据的有效复用。

专栏 6: 民航旅客流量预测模型的数据集

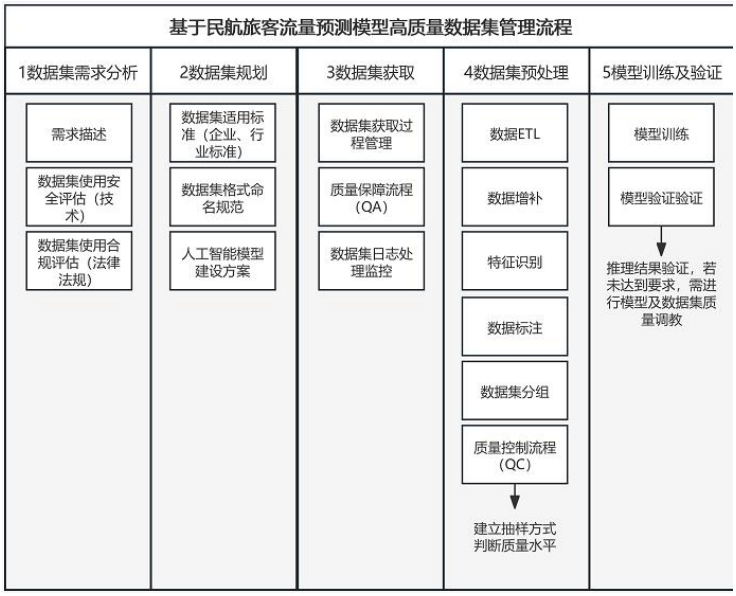
案例背景:

如何高效精准地对国际航班旅客量预测是目前民航亟待解决的技术问题，在进行不同航线旅客流量预测时，需要大量历史数据和先验知识构建预测模型，但受数据隐私保护政策约束，不同国家、区域航司之间记录的旅客流量数据无法实时共享，训练样本缺失，导致难以精准预测全市场航线旅客量。基于此，中航信数智建立了针对民航旅客流量预测的高质量数据集管理流程，对应用于旅客流量预测模型数据集进行修复、增补，并运用规则匹配、交叉验证和模型检测等一致性校验方式提升数据集质量，显著提升了模型预测准确性。

实践方案:

随着航空运输需求对旅客量预测的不断重视，需求预测理论与方法层出不穷，包括回归模型、重力模型、时间序列模型、神经网络、灰色模型等。在现有技术中，通过获取待预测日的历史旅客量特征数据和在待预测日的日期特征数据，计算间隔日，并将特征数据和归一化处理后的日期特征数据输入到训练好的旅客量预测模型，得到特定时间特征的旅客量预测值。

中航信数智基于特定航线航班及旅客三年历史样本数据建立民航旅客流量预测模型，用于航线及机场维度的全市场旅客量的预测。并建立高质量数据集管理流程，重点解决不同国家地区航司跨区域旅客量数据采集、共享不足，用于模型训练数据匮乏的问题；并基于数据全生命周期管理理念，在不同阶段都设置了数据质量提升计划。



在数据预处理阶段，针对训练数据不足的问题，利用航班历史承运旅客量、历史旅客量特征和时间特征等参数，对航班的历史承运旅客总量进行补充和推断，解决现有技术中跨系统隐私限制造成的预测准确性较差的问题。针对旅客流量数据中的非随机缺失，采用主动学习标注策略筛选高价值样本，通过标注变量标记人工修正记录，增强模型对数据可信度的识别能力，融合多任务学习的标注系统，自动化算法初筛疑似异常数据，人工标注复核后反馈至神经网络调整时序特征权重,最终实现对机场出入境国际航班旅客量进行预测。

在模型训练和验证阶段，根据质量保障流程重点关注数据集的一致性检查，即主要关注数据格式、数据结构和标签信息的一致性，通过规则匹配、交叉验证和模型检测等方式实现，确保数据不会因格式或内容不匹配影响模型训练和推理，并利用自动化检测与人工验证相结合的方式实施数据集的全面质量把控。

针对数据质量问题的后续管理，建立了分级处理机制，系统自动标记问题样本后，根据问题严重程度分别采取自动校正、人工复核或整批拒收等处理措施，并建立专门的不一致案例库用于质量分析。整个过程需完整记录检查日志、问题样本报告、修正措施等文档资料。

应用效果:

民航旅客流量预测模型的显著效果是将先验知识以数据标注的方式添加进数据补全过程，数据标注的方式完成样本缺失值处理，并通过平衡自动标注与人工标注比例，实现数据质量与模型性能的双重提升。

实践显示，通过该模型测算可以有效降低机场旅客流量预测失误率，对标注补全后的旅客流量数据采用策略回归与时间序列加权组合预测，某测试航线平均绝对误差较单一模型降低 21%，在大数据加处理的主动学习标注使百万级数据补全效率提升 3 倍。

此类高质量数据集管理流程可以应用于机场流量预测，旅游目的地民航旅客客流量、旅客价值分类预测等场景，并通过自动化标注算法有效降低民航数据治理成本。

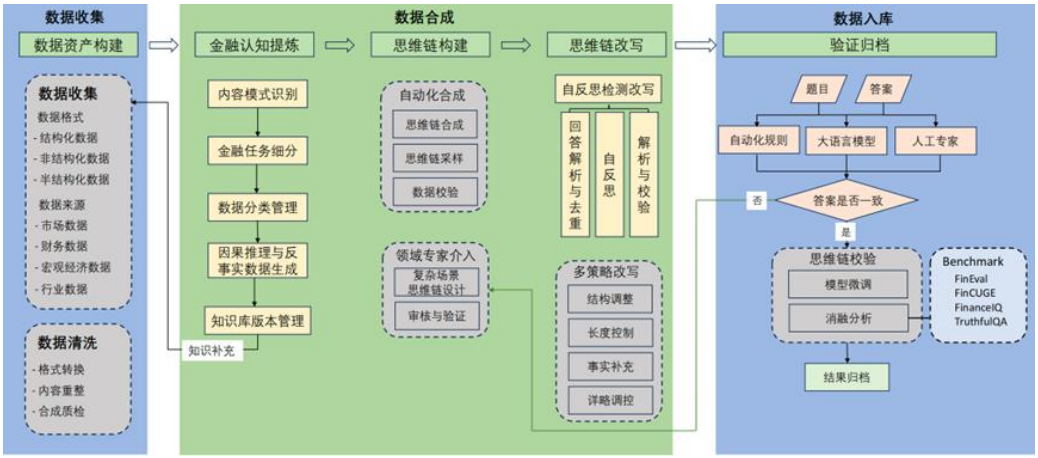
专栏 7：金融思维链推理数据集

案例背景:

随着人工智能技术在金融场景中的深入应用，如何在保障合规性与安全性的前提下，提升模型的智能推理能力，已成为行业发展的关键命题。在此背景下，构建高质量、可解释性强的金融领域专用数据集，尤其是金融思维链（Chain-of-Thought, CoT）推理数据集，逐渐成为推动 AI 金融应用升级的核心支撑。金融 CoT 锻造工程旨在通过数据引导模型生成具有逻辑性和透明度的输出，在风险控制、监管审查和决策溯源等方面发挥重要作用。

实践方案:

蚂蚁数科打造了 CoT 锻造车间用以支撑金融行业高质量推理数据集的高效建设，整体沿着数据收集、数据合成、数据入库三大流程进行。



多资产类别覆盖：包括股票数据、债券数据、基金数据、衍生品数据、其他数据等 5 大类目。

多金融任务覆盖：金融场景涉及多种复杂推理任务，蚂蚁数科基于行业认知将金融场景专有任务拆分为金融认知、金融知识、金融逻辑、安全合规与内容生成五大类，具体细分任务与任务描述如下表所示。

任务大类	任务名称	任务描述
金融认知	保险意图理解	文本意图理解
	保险槽位识别	保险实体识别
	研判观点提取	投资意图分类
	金融情绪识别	金融情绪分类
	金融意图理解	金融意图分类
	金融槽位识别	金融实体识别
金融知识	会计从业资格考试	资格考试试题
	保险从业资格考试	资格考试试题
	保险知识解读	名词解释
	基金从业资格考试	资格考试试题
	审计师考试	资格考试试题
	执业医师资格考试	资格考试试题
	执业药师资格考试	资格考试试题
	期货从业资格考试	资格考试试题
	注册税务师	资格考试试题
	理财知识解读	QA 选择
	证券从业资格考试	资格考试试题
	金融术语解释	名词解释
	银行从业资格考试	资格考试试题
金融逻辑	保险属性抽取	属性抽取任务
	保险条款解读	保险推理题
	金融事件解读	金融事件分析
	金融产品分析	金融推理
	金融数值计算	金融数值计算
安全合规	信息安全合规	安全拒答检测
	金融事实性	事实性风险检测
	金融合规性	金融合规性 Q 检测
	金融问题识别	是否属于金融问题检测
内容生成	投教话术生成	文本生成
	文本总结归纳	文本总结
	营销文案生成	营销文案分类
	资讯标题生成	标题生成

全自动数据清洗链路：原始金融数据往往来源多样、格式各异，并夹杂噪声、错误与冗余信息。因此，构建一套全面、高效、自动化的数据清洗链路，对于提升合成数据的质量、增强模型性能、确保最终投资决策的可靠性至关重要。该模块包括数据清洗与预处理、低质数据过滤、数据清洗与标准

化、数据去重等几大核心模块。

数据合成与标注：通过数据合成有效提升数据全面性进而对金融数据进行高效且专业的扩充已经成为行业共识。

基于因果推断的数据合成技术：反事实因果推断是提升模型稳健性与泛化能力的关键技术，尤其是在应对金融市场中的黑天鹅事件和极端情景时。通过构建反事实数据，训练模型在假设前提改变的情况下，依然能够进行准确的因果推断和风险评估，其效果在实际应用中表现显著。

思维链构建：金融场景下的 CoT 数据链路构建将专注于在 CoT 过程中构建模型的体系化思考能力，通过自然语言的输出增强模型思考过程的可解释性，增强人类用户对其的信任，同时融入实时动态数据与用户上下文信息，以确保模型输出的时效性与个性化。

自反思检测改写：在 CoT 合成过程中，由于问题难度与模型自身能力的问题，会导致部分的数据 CoT 合成失败，具体体现为：CoT 合成过程中的答案与真实答案不一致。为保证数据在采样过程中不因为结果错误被丢弃，而导致的数据量不足问题，在思维链构建过程后引入自反思检测改写机制，以确保更高的数据留存率。

人工标注与质检：对金融行业数据集进行人工标注，包括数据合理性、CoT 过程合理性、是否涉及个人隐私等；

数据入库从 4 个层面对数据集进行评估，包括基础质检：异常符号识别、长度检测、敏感词检测、去重等；难度检测：对数据进行难度分类分级；多样性：对数据进行细分领域标签确认；消融验证：通过下游 SFT、RL 实验并在评测集上与基线版本进行对比验证数据有效性。

应用效果:

高质量金融推理数据集- Ant-DeepFinance-1000K 能够有效模型在金融垂域场景能力。通过引入反事实因果推断技术，不仅成功扩充了 50%的问答对，还有效提升了模型对难例问题的处理能力，其中 43%的难例问题得到了精准检测与改写，大幅优化了模型在边界情况下的表现。在模型监督微调阶段，通过对齐真实业务需求和高质量数据，模型的回答准确率在多个权威评测集（如 Fineval）上均实现了 3%-10%以上的显著提升，尤其做到了金融合规任务、金融推理任务的有效提升，进一步验证了方案的技术优势。这些改进直接转化为实际应用中的效果提升：一方面，用户在智能投顾服务中获得了更精准、可靠的投资建议，显著提升了用户体验与信任度；另一方面，金融机构得以降低运营风险，提高服务效率，为业务增长提供了有力支持。

专栏 8：自动驾驶数据集

案例背景:

随着自动驾驶技术向 L3 及以上级别迈进，复杂路况与极端环境（如暴雨、隧道强光、夜间无路灯场景）下的感知精度成为技术瓶颈。传统数据集存在场景覆盖不足、标注标准不统一等问题，难以支撑模型应对长尾场景。因此，亟需构建具备多模态、高保真、标注精准的数据集，以提升自动驾驶模型在复杂环境下的决策可靠性。

实践方案:

柏川数据通过与多家车企和算法公司合作，构建了多模态自动驾驶高质

量数据集，并在数据采集规划、数据清洗流程、标注流程执行等方面形成了一套完整的实践方案。

(1) 数据采集规划:

确立需求目标：训练恶劣天气正常行驶能力。

确定采集场景：涵盖城市道路、高速路段、乡村小道等多种路况。

确定采集环境：区分晴天、阴天、风沙天、雨雪天、雾霾天、黑夜、逆光等不同环境条件。

确定采集装置：使用配备多种传感器（激光雷达、摄像头、毫米波雷达）的专业采集车辆，确保采集数据的多模态性。

规划采集所需各数据标签：将场景、环境进行组合搭配，确保满足各场景条件及 corner case 的需求。

规划所需数据量：针对各模型的训练需求，综合历史数据分析有效率，向上追加 cover 量，对不同场景定制采集数据量。当前模型确认雨天白天智能规避障碍物的水平较高，历史数据有效率 85%，采集量在需求基础上向上追加 20%，再结合各天气实际采集难度，最终输出如同下表的采集数量需求。

	常规天气		雨天		雪天		风沙天		雾霾天	
	白天	黑夜	白天	黑夜	白天	黑夜	白天	黑夜	白天	黑夜
城区非路口	9000	15000	3600	4800	1680	3600	2880	6000	3600	7200
城区路口	6000	10000	2400	3200	1120	2400	1920	4000	2400	4800
高速	5000	10000	2000	2000	2000	4000	2000	4000	2000	4000
乡村	10000	15000	4000	6000	1200	6000	3200	6000	4000	8000

备注
数字单位为帧，处理时需以1hz频率抽帧
白天定义：当日7点~18点
黑夜定义：当日18点~次日4点

（2）数据清洗流程:

确认数据质量标准：根据实际需求，商定数据清洗质量标准。单包数据至少连续抽取 60 帧，特殊天气场景中，单包需超过 45 帧拥有明显恶劣天气特征；常规天气场景，平均每帧车数量≥15，特殊天气场景，平均每帧车数量≥5。

制定清洗流程：运用自动化脚本与人工审核结合的方式，基于数据质量标准清洗数据。

1）通过自动化脚本，从采集数据中按照每秒抽 1 帧的频率，抽取出对应帧的各传感器数据，并对人脸、车牌等敏感数据进行脱敏处理

2）通过自动化脚本，智能识别过滤相机花屏错误、点云大片缺失数据

3）通过自动化处理，检测并修正数据格式问题，将不同传感器数据的时间戳处理统一

4）人工审核排查异常值，如激光雷达点云中的大片缺失，摄像头图像中的模糊、花屏，点云投影至图片时产生明显偏差等，确保数据的完整性与准确性

5）人工审核打标：筛选符合单帧目标物体量的数据，容差接受帧均 3 目标物；对场景中的指定特征进行标记，如分合流、路口、急转弯、信号灯、施工路段、天气时间、交通情况等，以便根据标签进行分类。

4. 标注流程执行:

制定标注规范：根据训练需求，对图像中的车辆、行人、路障等目标进行精确框选标注，对激光雷达点云数据进行三维目标检测和跟踪标注，设定

标注精度、参考尺寸，将规范量化、标准化，便于预标注可遵循指定规则预标、标注员批量生产。

适配预标注算法：根据训练需求，对预标注算法进行调整适配，匹配当前需求类型与生成规则，提升预标注准确率，降低人工修正成本。

人工修正：预标注技术自动识别常见目标，达到95%的识别准确率与80%的贴合精度，标注员对预标注结果进行复核、修正。

人工质检：实行多人交叉质检制度，对标注结果进行两轮~三轮审核，确保标注准确性与一致性。前期少量快速与训练需求拉齐，确立标准无误后，人员提效量产，达到大批量高质量交付的产出能力。

应用效果：

基于该高质量数据集训练的自动驾驶模型，在复杂场景下的目标检测与识别准确率显著提升。

在雨天夜晚场景中，对行人的综合检测准确率从 60% 提升至 82%，对车辆的综合检测准确率从 70% 提升至 87%，测试场景下极大降低碰撞概率。

在雪天场景中，基于采集难度较高、场景较难扩展的前提下，对行人的综合检测准确率从 40% 提升至 68%，对车辆的综合检测准确率从 65% 提升至 84%，有明显改善，且可持续优化。

在风沙、雾霾天场景中，对行人的综合检测准确率从 65% 提升至 77%，对车辆的综合检测准确率从 75% 提升至 85%。

模型泛化能力增强，能够更好地适应不同路况、不同天气的驾驶环境。并且，该数据集为预标注结果算法提供了丰富多样的数据样本，优化了预标注模型，使其在新数据上的预标注准确率提高了 8%，为后续标注工作节省了大量时间与人力成本。

四、高质量数据集发展趋势

高质量数据集是我国数据要素和人工智能产业发展的基石。随着“数据要素×”、数据标注、可信数据空间、数据基础设施等政策的落地实施，工业、交通、医疗、科研、金融、教育等众多行业领域已开展高质量数据集建设实践，良好的建设成效也开始驱动企业在已有业务中增强数据质量。面向未来更广阔的发展空间，在完善智能化数字技术和数据基础设施的驱动下，企业高质量数据集建设运营能力将持续提升，行业应用将更加丰富，产业生态也将更加开放，推动数据集源源不断形成，有效赋能人工智能产业和千行百业发展。

（一）建设运营能力逐步成熟

随着数智技术的飞速发展，高质量数据集的建设与运营方法论正逐步走向成熟，成为推动人工智能应用落地的关键支撑。在建设阶段，企业必须从战略高度出发，明确数据集的业务目标和应用场景，确保其构建与企业的核心需求紧密相连。随着数据开发、标注和评估工具的不断完善，大模型等智能化技术的加速发展，高质量数据集建设过程的数据标注、分类分级、质量评估的效率与效果将显著提升。在运营阶段，量化指标从数据质量扩展至经济成本、企业战略等更多维度，结合数据评估、更新与维护机制，企业将实时掌握数据集质量与企业运营情况，并通过开发丰富数据产品进一步释放数据价值，通过质量管理、血缘追踪、成效评估等实现数据处理、评估、应用全链条的闭环优化。

（二）多行业多场景加速落地

“人工智能+”行动到哪里，高质量数据集的建设和推广就要到哪里。**高质量数据集作为人工智能技术创新和产业发展坚实的数据基础，未来将在传统领域与创新领域加速落地应用。**一方面，高质量数据集可实现复杂、细分场景下的建模与智能决策，正加速赋能工业、交通、金融、能源等行业。如交通领域，将场景丰富的道路数据、大规模的车辆轨迹数据用于模型训练，可快速提升交通拥堵治理和应急调度优化水平。能源领域，完整准确的设备运行、传感监测数据集，可加速推动运维智能化与风险防控水平。另一方面，随着合成数据等新型技术不断完善、落地应用，混合真实与合成的高质量数据集，还将加速应用于低空经济、具身智能等创新领域。高度真实的模拟数据将能解决为无人机、机器人训练时的长尾问题，并提升机器在多场景中的适应性。

（三）基础设施推动协同生态

高质量数据集建设过程中，往往涉及与外部的交易、交互，如数据采买、数据标注、质量评估、利用外部工具进行数据开发等。**未来，随着可信数据空间等数据基础设施的落地，将推动形成高质量数据集协同生态。**一方面，协同生态将吸引更多数据提供方与服务方加入，促使更多潜在的数据源被挖掘和利用。这一过程不仅能推动高质量数据集与各行业的深度融合，还能将数据开发、质量评估等关键环节交由专业服务方承接，从而显著加速高质量数据集的构建进程。另一方

面，随着数据定价、数据权属、收益分配等机制的完善和生态内共识规则的明确，企业将获得更清晰的运营指引和更有力的激励机制，将推动企业高质量数据集持续运营，从而更好赋能数据要素市场发展。

联系方式：
大数据技术标准推进委员会
地址：北京市海淀区花园北路52号
邮编：100191
邮箱：baiyuzhen@caict.ac.cn
官网：www.tc601.com