高质量数据集建设指引

2025年8月

前 言

党的十八大以来，以习近平同志为核心的党中央高度重视我国新一代人工智能发展。习近平总书记深刻把握世界科技发展大势，洞察人工智能的战略意义，在二十届中央政治局第二十次集体学习时指出，“人工智能作为引领新一轮科技革命和产业变革的战略性技术，深刻改变人类生产生活方式。”“我国数据资源丰富，产业体系完备，应用场景广阔，市场空间巨大。”这为把握智能化浪潮，释放数据要素价值指明了前进方向，提供了根本遵循。

随着大模型技术应用的快速发展，人工智能的研发重点正从“重点优化模型架构”转向“模型与数据协同优化”，其中高质量数据的作用日益凸显。数据作为人工智能发展的三大核心要素之一，已成为人工智能大模型训练的核心要素资源，决定了大模型的性能。加快人工智能高质量数据集建设，夯实人工智能发展数据基础，对于推动“人工智能+”场景落地具有重要意义。国家数据局联合各部门、各地区，构建起“部门协同、央地联动”的工作机制，联合施策、协同发力，积极引导做好高质量数据集建设工作，明确提出“‘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”。

由国家数据局指导，中国信息通信研究院、国家数据发展研究院、中国电子技术标准化研究院、国家信息中心、国家发展和改革委员会创新驱动发展中心、中国电子信息产业发展研究院等单位，在充分调研的基础上，编制《高质量数据集建设指引》，总结高质量数据集建设方法论，指导推进高质量数据集建设，力争为人工智能纵深发展提供有力支撑。

目 录

一、 高质量数据集建设背景

（一）高质量数据集的发展背景

在以人工智能为代表的新一轮科技革命和产业变革深入推进的大背景下，数据正日益成为国家基础性战略资源和关键性生产要素。高质量数据集作为人工智能发展的基础支撑，其重要地位不断上升，成为驱动技术创新、赋能产业升级和提升治理能力的重要抓手。当前，高质量数据集的建设正处于政策驱动与场景牵引同步发力、协同推进的关键阶段。

1．高质量数据集成为人工智能发展核心支撑

人工智能技术迈入大模型时代后，研发重点从“重点优化模型架构”转向“模型与数据协同优化”，其中高质量数据的作用日益凸显。主要表现在：一是将数据视为活的资产，不再是一次性收集、处理后就束之高阁的静态文件，而是一个需要持续投资、管理、监控和优化的动态、演进的战略资产。二是用自动化的、可编程的、可扩展的数据处理流程，取代手动的、一次性的数据处理工作，系统性处理海量数据，并能提升数据质量。三是整合领域专家，将拥有深厚行业知识的主题专家直接整合到数据处理的流水线中。专家知识被用来定义数据标准、标注复杂案例、识别数据中的细微偏差，从而将领域智慧注入数据。四是建立模型反馈闭环，将模型在实际应用中的错误作为诊断信号，用来发现数据中的问题（如标签错误、数据分布不均、边界案例缺失等），然后有针对性地改进数据集。由此就形成了一个“数据飞轮”效应，更好的数据训练出更好的模

型，更好的模型反过来帮助获得更好的数据。

大模型参数规模指数级增长与多模态能力的拓展，数据需求从“量级积累”转向“量质并重”。例如，以 OpenAI 为代表的国际领先科技企业正通过强化微调等技术手段，依托小规模但高度精准、精细化、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。而 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。

人工智能走进千行百业的具体场景，行业模型的性能跃升越来越依赖数据与场景的深度耦合，从“数据规模竞赛”转向“数据质量深耕”。越来越多的企业开始采用自动化的数据筛选、数据标注与数据增强等技术工具，以提升数据集的专业性和适用性。在这种趋势下，模型训练不再依赖于盲目扩充数据规模，而是更注重数据的代表性、多样性和场景适配性，为人工智能的可持续发展奠定坚实基础。例如，医疗领域中某肺结节检测模型训练数据集仅利用 1 万多例数据和亚毫米级病灶边界勾画的标注信息，使得早期肺癌筛查中的假阳性率大幅下降；工业质检场景中某企业通过合成数据等技术生成了 10 万种“极端缺陷样本”，弥补了真实生产中罕见缺陷数据不足的问题，使模型缺陷识别覆盖率大幅提升。

2．国家高度重视高质量数据集建设

党中央、国务院高度重视数据资源的开发利用与高质量发展，围绕构建数据基础制度、完善数据要素市场、推动公共数据开放、夯实智能技术底座等方面，陆续出台多项纲领性政策文件，为高质量数据资源体系建设提供了顶层设计和制度保障。

2022 年 12 月，中共中央、国务院印发《关于构建数据基础制度更好发挥数据要素作用的意见》，明确提出探索开展数据质量标准化体系建设。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）》，强调数据要素高质量供给与合规高效流通，提出打造高质量人工智能大模型训练数据集。2024 年 12 月，国家发展改革委、国家数据局等部门印发《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**”概念，将其作为人工智能与实体经济融合的核心载体，并对行业数据集建设提出具体要求。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》等政策均提出建设行业“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，加强统筹协调，深化部门协同，全力推动高质量数据集建设，高效赋能行业高质量发展，标志着高质量数据集建设进入系统化、规模化推进阶段。

（二）高质量数据集的概念内涵

高质量数据集是指经过采集、加工等数据处理，可直接用于开发和训练人工智能模型，能有效提升模型表现的数据的集合。高质量数据集主要服务于人工智能的实际应用场景，通常包括以下四个核心组成要素：特征、标签、元数据和样本。特征是模型训练的输入变量，用于描述每个样本的具体属性；标签是需要模型预测的目标输出；元数据记录了数据生成与处理过程的相关信息，如采集时间、地点、来源等；样本则是构成数据集的基本单元，由特征向量及其对应的标签共同组成。例如，机器学习中的经典数据集鸢尾花（Iris）数据集，包含 150 条样本，均匀分属三类鸢尾花，每类 50 条样本，使用花萼长度、花萼宽度、花瓣长度和花瓣宽度作为分类特征。再如图像识别领域广泛使用的 ImageNet 数据集，涵盖超过 1400 万张高分辨率图像，覆盖 2 万多个类别，每张图像均配有准确的类别标签，其中超过 100 万张图像还包含了物体边界框等精细标注信息。

高质量体现在规模“大”、安全“牢”、观点“正”、效果“好”、应用“广”等方面，可以采用静态和动态的质量评价方法来度量。静态质量主要关注数据本身的关键属性，在准确性、完整性、一致性、时效性等基础指标上增加多样性、真实性、合规性等维度，重点评估数据的领域覆盖、来源可靠性以及在隐私保护和安全合规方面的表现。动态质量则强调数据集在模型训练和应用中的实际效果，可通过引入代表性模型开展基准测试，结合基准评测数据集与量化指标，客观衡量模型性能的提升程度，从而明确数据集的“高质量”标准。同时，

还应建设统一的质量评估平台，规范评估流程与工具，增强不同数据集之间的可比性与通用性。**由于不同行业数据集的模态分布、标注需求差异较大，需根据行业特点应用不同的数据处理技术和方法，其质量评价也需要在通用的指标上进行定制加强。**例如，医疗卫生领域，以文本（电子病历）和医疗影像居多，侧重于文本解析、图文结合处理和专业标注等处理方式，更关注数据内容的合规性、安全性和标注准确性；工业制造领域，以时序数据、图像、图纸文档、仿真数据居多，侧重于时序数据处理、高精度合成和专业标注等处理方式，更关注数据内容的真实性、多样性和标注准确性。

（三）高质量数据集的分类

高质量数据集的类型和特性因应用场景、数据来源与模型目标的不同而呈现多样化，可以从**数据模态**、**模型阶段**与**行业应用**三个维度对当前主要的高质量数据集进行分类。

在数据模态方面，可以分为单模态数据集和多模态数据集。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 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。而近期新涌现的思维链数据则是一种数据标注方法或推理过程的表示方法，而非一种独立的数据模态，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。

在模型阶段方面，主要分为预训练数据集、微调数据集和评估数据集。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”，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。微调数据集是让模型“术业有专攻”的关键，其核心作用是让模型在特定任务、领域或场景中表现更优，引导模型聚焦特定任务的规律，强化与任务相关的知识，同时弱化无关信息的干扰，最终实现“通用能力+专项技能”的结合。它是连接通用预训练模型与具体应用需求的“桥梁”，相比预训练阶段使用的海量通用数据，微调数据集通常具有规模更小、针对性更强、标注更精细的特点。评估数据集是一类精心设计的数据样本，主要目的是为了相对客观地衡量模型的性能和泛化能力，具备独立性、代表性、时效性等特点。

在行业应用方面，参考技术文件《高质量数据集 分类指南

（征求意见稿）》可以分为通识数据集、行业通识数据集和行业专识数据集。高质量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，例如百度百科；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，例如行业研究报告；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，例如医疗领域的电子病历数据集等。

二、高质量数据集应用需求

随着人工智能技术的快速发展，不同应用领域对高质量数据集的需求呈现出层次化、专业化的特征。根据 AI 系统能力的发展路径和认知层次，**高质量数据集的应用需求可以划分为基础认知层、场景理解层、行动规划层三个递进层次。**从建立世界的基本认知，到解析复杂场景关系，再到规划执行具体行动，每一层都承载着不同的学习目标和能力要求。深入探索这三个层次对高质量数据集的具体需求，将为建设主体提供清晰的数据集建设方向和路径指引。

（一）基础认知层数据集需求——建立世界的基本认知

基础认知层是人工智能系统的根基，主要负责建立对世界的基础表征和模式认知。这一阶段的核心目标是让 AI 系统掌握“这是什么”的基本判断能力，构建对物理世界和抽象概念的初步理解框架。基础认知层的能力直接决定了 AI 系统的认知上限——没有扎实的基础表征，就无法实现深层理解和复杂推理。

1．应用目标：建立全面的基础认知框架

基础认知层需要通过海量数据学习各领域的通用模式和基本概念，这种学习过程类似于人类儿童通过大量观察和体验认识世界的过程。在语言领域，系统不仅需要掌握词汇、语法、语义的基础表征，还要理解语言的统计规律和上下文关联模式，形成对自然语言的内在理解；在视觉领域，需要学习从低级特征（边缘、纹理、颜色）到高级概念（物体、场景）的层次化表征，建立视觉世界的认知地图；在跨模态领域，需要建立不

同模态间的基础对应关系，理解同一概念在不同感知通道中的表现形式。这种学习强调知识的广度覆盖和基础模式的充分学习，为后续的专业化学习和深度理解奠定坚实基础。

2．数据内容：海量数据支撑通用能力

基础认知层数据集的显著特征是规模庞大，通常达到 TB 至 PB 级别，这种规模需求有其深层的理论依据。大语言模型的预训练语料包含数万亿词元（Token），这种海量数据使模型能够捕捉语言中的长尾分布和罕见模式；视觉领域的大规模数据集如 ImageNet 包含超过 1400 万张图像，覆盖 2 万多个类别，确保模型能够学习到视觉世界的多样性。这种大规模需求源于模型需要从数据中学习通用表征，遵循尺度定律（Scaling Law）——随着数据规模的增加，模型性能会持续提升，且这种提升呈现幂律关系。更重要的是，海量数据能够提供足够的统计信息，使模型学习到稳定可靠的模式，而非过拟合于特定样本。

3．数据质量：覆盖广度与基础质量并重

基础认知层对数据质量的要求体现在多个维度的平衡上。首先是覆盖面的广度，数据需要涵盖多领域（科学、文学、历史、技术等）、多语言（主流语言及小语种）、多场景（正式、非正式、专业、日常等），这种广覆盖确保模型具备处理多样化输入的能力；其次是数据分布的合理性，要能够反映真实世界的特征分布，避免因数据偏见导致模型产生系统性偏差；第三是基础质量的保障，需要经过去重处理避免过拟合、去噪过

滤提高信噪比、内容审核确保安全合规。值得注意的是，相比后续层次，这一阶段对标注精度的要求相对较低，更多依赖自监督学习，从数据本身的结构中学习，这也是为什么能够利用互联网规模数据的关键所在。

4．典型应用：奠定模型基础能力

基础认知层数据集支撑了各类基础模型的训练，这些基础模型成为整个AI生态系统的基石。语言领域的GPT、BERT等模型通过大规模文本预训练，不仅学会了语言的表面形式，更掌握了语言背后的知识结构和推理模式，为各种下游任务提供了强大的语言理解能力；视觉领域的ResNet、Vision Transformer等通过大规模图像数据集训练，建立了从像素到语义的完整映射，使计算机视觉从特征工程时代进入深度学习时代；跨模态的CLIP等模型通过海量图文对数据，学习到视觉和语言的统一表征空间，实现了零样本图像分类等突破性能力。这些基础模型通过迁移学习和微调，能够快速适应各种下游任务，极大降低了AI应用的开发成本和技术门槛。

（二）场景理解层数据集需求——解析复杂场景关系

场景理解层在基础认知之上，负责理解复杂场景中的结构关系、语义逻辑和事件过程。这一层的核心是让AI系统能够深度解析“这里发生了什么”“关系如何”“为什么会这样”等需要理解的问题。场景理解层是AI从“看到”到“看懂”的关键跨越，它要求系统不仅能识别单个元素，更要理解元素间的相互作用和整体语境。

1．应用目标：实现结构解析与深层理解

场景理解层要求模型具备结构解析和关系推理能力，这种能力远超简单的模式匹配。在语言领域，模型需要理解篇章的层次结构、段落间的逻辑关系、句子中的隐含信息和言外之意，能够进行因果推理、类比推理和反事实推理；在视觉领域，需要理解多个对象的空间关系（上下、左右、包含、相邻）、功能关系（工具与使用者、容器与内容物）和场景的整体布局，从而推断场景的功能和可能发生的事件；在视频领域，需要理解时序事件的因果链条、动作的目的和结果、场景的动态变化规律，捕捉关键帧之间的语义连续性。这一层强调从简单识别到深度理解的能力跃升，要求模型具备类似人类的场景解析和情境推理能力。

2．数据内容：精细化标注与语义信息丰富

场景理解层数据集包含丰富而精细的标注，每个标注都承载着特定的语义信息。语言理解数据集如 SQuAD 不仅包含问答对，还标注了答案在原文中的精确位置和推理依据，要求模型理解问题意图并定位关键信息；视觉场景数据集如 COCO 包含多层次标注体系——从粗粒度的场景类别到细粒度的像素级分割，从静态的对象位置到动态的动作描述，形成了完整的视觉语义体系；视频理解数据集如 ActivityNet 包含时序动作的精确边界、动作类别的层次结构以及事件间的因果关系。数据规模通常在十万到百万级别，这种相对适中的规模反映了一个重要权衡：标注的精细度与覆盖的广度。每个样本都经过精心设

计和标注，确保包含足够的信息密度来训练模型的理解能力。

3．数据质量：语义完整性与逻辑一致性

场景理解层对数据质量有着严格的多维度要求。语义完整性要求标注覆盖场景的所有关键信息，不能有重要元素的遗漏——例如在图像描述中，不仅要标注主要对象，还要包括它们的属性、状态和相互关系；逻辑一致性要求不同层次、不同角度的标注必须相互协调，形成统一的语义表达——如对象检测的结果要与场景分类一致，时序标注要与事件描述对应，避免矛盾和歧义；标注精确性要求细粒度的语义区分，能够区分相似但不同的概念、动作或关系。这需要专业的标注团队经过系统培训，遵循详细的标注规范，并通过多轮交叉验证和一致性检查来保证质量。质量控制流程通常包括标注员培训、样例标注、批量标注、质量抽检和反馈改进等环节。

4．典型应用：支撑复杂理解任务

场景理解层数据集广泛应用于各类需要深度理解的 AI 任务。在自然语言处理领域，机器阅读理解系统能够回答关于文本的复杂问题，信息抽取系统能够从非结构化文本中提取结构化知识；在计算机视觉领域，目标检测能够精确定位和识别图像中的多个对象，语义分割能够理解每个像素的语义类别，场景图生成能够构建对象间的关系网络；在视频分析领域，动作识别能够理解人类的复杂行为，事件检测能够发现视频中的关键时刻。这些应用不仅要求模型能够识别“是什么”，更要理解“为什么”和“怎么样”，真正实现对场景的深度理解。

（三）行动规划层数据集需求——规划执行具体行动

行动规划层是AI系统的最高能力层，负责将理解转化为决策和行动，实现从认知到执行的完整闭环。这一层的核心是让AI系统掌握“怎么做”“为什么这么做”以及“这样做的后果是什么”，形成类似人类的决策推理能力。行动规划层代表了AI从被动响应到主动规划的质变，是实现通用人工智能的关键环节。

1．应用目标：掌握完整的推理决策链条

行动规划层需要学习从问题识别到方案制定再到执行验证的完整认知过程。在复杂推理领域，模型需要掌握多步推理的逻辑链条，能够将复杂问题分解为子问题，选择合适的求解策略，并验证每一步的正确性；在对话交互领域，需要理解用户的真实意图（包括字面意思和潜在需求），根据上下文选择合适的回应策略，维持对话的连贯性和目标导向性；在代码生成领域，需要理解需求的本质，设计合理的算法架构，处理边界条件和异常情况，生成高质量的可执行代码；在具身智能领域，需要将高层任务目标分解为可执行的动作序列，考虑环境约束和不确定性，实时调整执行策略。这一层强调思维的完整性、决策的合理性以及执行的可行性。

2．数据内容：包含完整推理链与决策过程

行动规划层数据集的核心特征是包含完整、可追溯的推理和决策过程。思维链（Chain-of-Thought）数据不仅提供最终答案，更重要的是展示到达答案的每一个推理步骤，包括假设的

提出、验证、修正的完整过程，使模型学会“如何思考”而非仅仅“记住答案”；代码数据集包含从需求分析、设计思路、实现细节到测试验证的完整软件开发流程，每个决策点都有明确的理由和权衡；强化学习数据集记录了智能体在环境中的完整交互历史，包括状态观察、动作选择、奖励反馈和策略调整，展现了试错学习的完整过程；人机对话数据集保留了多轮交互的完整上下文，包括话题的展开、转换和收束，以及对话策略的动态调整。数据规模相对精炼，通常在千到百万级别，但每个样本都是精心构造的“思维标本”，包含丰富的决策信息和推理逻辑。

3．数据质量：推理严密性与价值对齐

行动规划层对数据质量的要求达到了最高标准，因为这直接关系到AI系统的决策质量和安全性。推理正确性不仅要求最终结果正确，更要求推理过程的每一步都有充分的逻辑依据和事实支撑，避免“歪打正着”的伪推理；逻辑严密性要求推理链条完整、清晰、可验证，没有逻辑跳跃、循环论证或自相矛盾，每个推导步骤都能被人类专家理解和审核；价值对齐是最关键也最具挑战性的要求，需要确保AI的决策符合人类的价值观、道德准则和社会规范，不产生有害、偏见或违背伦理的输出。这通常需要通过人类反馈强化学习（RLHF）进行迭代优化，由领域专家进行多轮评估，甚至需要建立专门的伦理审查机制。数据的构建过程往往需要跨学科团队的协作，包括领域专家提供专业知识、标注员进行精细标注、算法工程师设计验证机制。

4．典型应用：实现智能决策与执行

行动规划层数据集支撑了 AI 系统最具挑战性和实用价值的应用。在科学研究领域，数学定理证明系统能够发现新的证明路径，科学假设生成系统能够提出可验证的研究方向；在软件开发领域，代码生成系统不仅能编写功能代码，还能进行代码优化、错误调试和文档生成；在人机交互领域，对话系统能够进行深度的知识问答、情感陪伴和任务协助，展现出类人的交流能力；在机器人领域，自主导航系统能够在复杂环境中规划路径，操作机器人能够完成精细的物体操控任务；在游戏 AI 领域，策略系统能够在复杂的游戏环境中制定长期策略，展现出超越人类的规划能力。这些应用代表了 AI 技术的最前沿，正在逐步改变人类的工作和生活方式。

通过对基础认知、场景理解、行动规划三个层次数据集需求的系统分析，我们可以得出以下核心洞察：

表 1 三类应用场景数据集需求对比表

维度	基础认知层	场景理解层	行动规划层
核心问题	这是什么	这里发生了什么	怎么做
应用目标	基础表征与模式识别	结构解析与深层理解	推理决策与策略执行
数据内容	海量数据、广泛覆盖	精细化标注、语义丰富	推理链完整、决策明确
数据质量	覆盖广度、分布合理	语义精确、逻辑一致	推理严密、价值对齐
典型应用	GPT/BERT 语言模型、CLIP 多模态模型	机器阅读理解、目标检测、动作识别	数学推理、代码生成、对话系统、机器人控制

这种层次化的数据集建设框架不仅反映了当前 AI 技术发展的实际需求，更揭示了智能系统能力提升的内在规律。从基础

认知到场景理解再到行动规划，每一层都建立在前一层的基础之上，同时为下一层提供支撑，构成了 AI 能力发展的完整路径。基础认知层提供了感知和表征能力，场景理解层实现了语义解析和关系推理，行动规划层完成了决策制定和执行规划。通过针对性地建设不同层次的高质量数据集，平衡各层次的发展需求，可以系统性地推动人工智能从狭义智能向通用智能演进，最终实现真正意义上的智能系统。

三、高质量数据集建设现状

近年来，全球高质量数据集建设进入加速阶段，呈现出政策引导、市场驱动与技术革新协同推进的态势。欧美等发达经济体在开放共享、标准体系、平台化建设方面走在前列，形成了较为完善的多模态、多领域数据集生态体系；我国则在国家顶层设计和多方协同推动下，高质量数据集建设体系逐步完善，区域与行业层面呈现并进发展格局。本指引通过分别梳理全球与我国的高质量数据集建设情况，分析当前面临的主要困难与挑战，为后续建设方法的探讨提供现实基础。

（一）全球高质量数据集建设现状

目前全球范围内，高质量数据集建设呈现出通识类与行业类并行推进的格局。通识数据集主要服务于通用人工智能模型的基础能力建设，强调广度和多样性；行业数据集则聚焦于特定领域的知识与场景，强调深度与专业性。

1．通识数据集建设

欧美国家作为通识数据集的建设主力，一方面全力支撑大规模预训练模型。例如，Kaggle Datasets 平台提供了超过 50 万个由用户和企业上传的真实数据集，覆盖 CV、NLP、金融、健康、社交等多个领域，美国人工智能公司 Hugging Face 托管超 4 万个开源数据集，涵盖文本、图像、语音等多种模态，数据规模超过 15 万亿 Token，已成为全球数据集开源托管的核心枢纽之一。美国非营利组织创建的非结构化、多语言网页开源数据集 Common Crawl，总数据量达到 PB 级别，是 OpenAI、

Meta 等科技巨头大规模语言模型训练的重要数据来源之一。德国非营利组织创建的数据集 LAION-5B 以图文对数据为主，是全球最大的多模态图文开源数据集之一，超过 58.5 亿个图文对，为 Stable Diffusion 系列模型训练和 AI 图像生成提供了重要的数据支持。加拿大多伦多大学开发者创建的数据集 BooksCorpus 是一个以电子英文书籍为主的文本类数据库，覆盖多领域多学科，超过 1 万本完整书籍，是 GPT 系列模型训练的重要数据来源。**另一方面，积极建设开放平台等基础设施提供统一服务。**例如，截至目前，美国国家开放数据平台（data.gov）收录超过 32 万个数据集，涵盖环境、健康、交通、海洋、能源等领域。欧盟统一数据门户（data.europa.eu）成为欧盟全域开放数据的中央访问节点，已收录 35 个国家的超过 195 万个数据集，覆盖行政、健康、环境、经济、科技等领域，成为欧盟发展数字主权、推动 AI 创新的战略基础设施。英国开放数据门户（data.gov.uk）作为官方公共数据集访问节点，收录超过 5.6 万个数据集，涵盖政府机构发布的民生、经济、环境等领域，聚焦高价值数据集成的 AI 训练生态系统。

2．行业数据集建设

一方面，欧美国家通过实施法案加速行业数据集的建设与开放。欧盟《高价值数据集实施法案》于 2024 年 6 月生效，强制要求成员国开放环境、地理、交通、企业、经济、气象等行业通识数据，要求机器可读格式，提升数据可用性，支撑环境监测、城市规划及欧洲共同数据空间建设，为跨行业 AI 模型提

供基础框架。英国《数据使用与访问法案》于2025年6月正式成为法律，旨在利用数据力量系统性地推进跨行业通识数据集的建设，构建标准化的数据共享框架，覆盖能源、金融、电信、零售等八大战略行业，致力于实现释放100亿英镑以上的经济价值目标。**另一方面，多领域各行业加速数据集建设。**欧盟EuroStat Industry Hub数据集包含所有成员国的完整工业统计数据，覆盖制造业、能源、建筑业等核心产业。美国医疗领域高质量数据集PubMed，含超3800万篇论文摘要，为医疗大模型提供术语体系与知识框架，支撑临床决策辅助系统、药物研发模型。美国证监会企业财务报告数据库SEC filings，收录超过1800万份文件，是目前全球最大、最完整的开源上市公司财务文本库之一，已广泛用于量化投资、自然语言处理预训练、合规监控与生成式 AI 决策系统。

（二）我国高质量数据集建设现状

在党中央、国务院的统筹部署下，我国高质量数据集建设成效明显。截至2025年6月，全国建设高质量数据集超3.5万个、总量超400 PB；数据交易机构挂牌高质量数据集3364个，作为交易流通中的关键商品，累计交易额近40亿元，规模达246 PB；国内多数模型使用中文数据占比达到60% - 80%。

我国高质量数据集建设总体呈现出“统筹规划、分层推进、多元协作”的鲜明特点。建设工作既在国家层面统一部署、总体谋划，又在区域与行业两个维度形成分工协作的格局。区域高质量数据集建设由国家数据局统筹指导，依托各地政策和资

源分层推进落地；行业高质量数据集建设则以中央企业和科研机构等为牵引力量，联合其他社会主体协同参与、优势互补。

1．区域高质量数据集建设

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。一方面，国家数据局统筹建设成都、沈阳、合肥、长沙、海口、保定和大同七大数据标注基地，充分发挥地方配套支撑作用，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，集聚龙头企业，促进区域人工智能产业生态发展，目前已建设行业高质量数据集524个，数据总规模超过29 PB，赋能163个国产人工智能大模型研发与应用，带动数据标注行业相关产值超过83亿元。另一方面，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政策体系、打造特色案例。例如贵州以专项资金支持重点行业领域，建设高质量数据集。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所引入高质量数据集567个，覆盖医疗、交通、能源、工业等20多个行业和领域。

2．行业高质量数据集建设

行业层面，中央企业、大模型技术企业、标准化组织、科研学术机构等多方主体正协同共建行业生态体系，形成了多元联动的发展格局。一是行业主体发挥数据资源优势，成为高质量数据集建设的重要力量，医疗卫生、工业制造、智慧能源等

领域建设活跃，低空经济、具身智能、生物制造等领域需求迫切。今年4月，国务院国资委发布首批10余个行业30项央企人工智能行业高质量数据集优秀建设成果。8月，国家数据局征集遴选出104个高质量数据集典型案例，涵盖科学研究、工业制造、智慧能源、交通运输、医疗卫生、教育教学等12个重点领域，以及低空经济、具身智能、智能驾驶、智慧海洋、生物制造等5个创新领域。二是全国数据标准化技术委员会等相关标准化组织积极协同企业开展高质量数据集建设和标准化研讨会，助力完善高质量数据集国家、行业、团体等标准体系，明确高质量数据集的建设路径，为业界实践提供兼具方向性和规范性的操作指引，推动行业数据水平提升。三是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注入动力。例如阿里巴巴发布中文问答数据集，为智能问答系统的研发提供了高质量的训练数据。智源研究院发布中英双语数据集IndustryCorpus1.0包含3.4 TB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。鹏城国家实验室开源百万规模标准化具身智能数据集，超过300万样本，覆盖258个系列任务和321064个具体任务实例。上海人工智能实验室开源数据平台OpenDataLab提供5500多个数据集，涵盖1500多种任务类型，总数据量达到80 TB以上，下载量超过百万次，为行业技术创新提供了丰富的数据支撑。

（三）高质量数据集建设困难与挑战

虽然我国高质量数据集建设在国家统筹、推进模式和应用场景方面具有独特优势，但在数据开放度、标准体系、关键技术及国际影响力等方面的短板，已经在实践中转化为数据供给、技术工具、标准规范、安全合规、商业模式等多重困难与挑战。

数据供给方面，结构性短缺与流通壁垒。高质量语料枯竭风险，尤其是专业领域数据储备量不足。数据孤岛与开放困境，跨部门、跨地区数据共享机制不健全，授权运营平台覆盖不足。

技术实现方面，处理能力与工具链水平薄弱。现有技术难以高效处理文本、图像、视频等混合结构数据，自动化清洗、智能化标注工具成熟度低。数据清洗、标注等环节仍依赖传统统计方法，人工智能驱动的智能治理引擎薄弱。

标准与治理方面，规范与协同机制待完善。标准体系规划仍需完善，如行业高质量数据集建设指南、分类标准、数据格式、质量评测等关键标准不充分，且标准应用与推广力度不足。数据合成、数据标注等关键技术也缺乏统一标准和规范性指导。

安全与合规方面，风险控制与开放平衡。隐私与安全技术瓶颈，数据脱敏、差分隐私等技术的规模化应用滞后，数据泄漏风险制约高价值敏感数据（如医疗、金融）的开放。权属规则不明晰，数据授权运营主体边界模糊。

成本与模式方面，商业闭环未形成。投入产出比例失衡，数据标注与治理成本占比高，但价值转化周期长；缺乏成熟的“数据—算法—应用”商业生态，难以支撑长效化可持续运营。创新模式探索滞后，数据交易所尚未形成规模化交易市场。

这些问题不仅制约了高质量数据集建设的速度与质量，也影响了数据要素价值的有效释放。为破解上述瓶颈，本指引将在第四章从建设方法与技术路径的角度，提出体系化、可操作的工程方案，并在第五章从建设运营体系的角度，探讨多主体协同、标准化治理与商业化运营的落地模式。通过方法论与运营体系的双重发力，形成覆盖数据集建设全生命周期的应对策略，推动我国高质量数据集建设走向高效、可持续、国际化的新阶段。

四、高质量数据集建设方法与实践

（一）典型建设模式

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性工程。当前业界主要采用两种典型的建设模式：**“场景驱动”的建设模式和“数据驱动”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”。这种模式强调“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。

第二种模式是“数据驱动”的建设模式。以积累的大量、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。这种模式强调“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。

当前，国家层面及各行业对高质量数据集提出了更为明确的建设目标与应用要求。因此，从实际成效出发，以需求为牵

引的“场景驱动”模式更契合高质量数据集建设的核心目标和发展方向。因此，本指引参考技术文件《高质量数据集 建设指南（征求意见稿）》提出高质量数据集建设模式，如图 1 所示。

高质量数据集建设应按照生命周期有序展开，包括**数据需求、数据规划、数据采集、数据预处理、数据标注、模型验证**等环节。其中，各环节主要按以上顺序逐步开展，同时，各环节会对其他环节进行反馈，或者会在其他环节反馈下进行迭代优化。

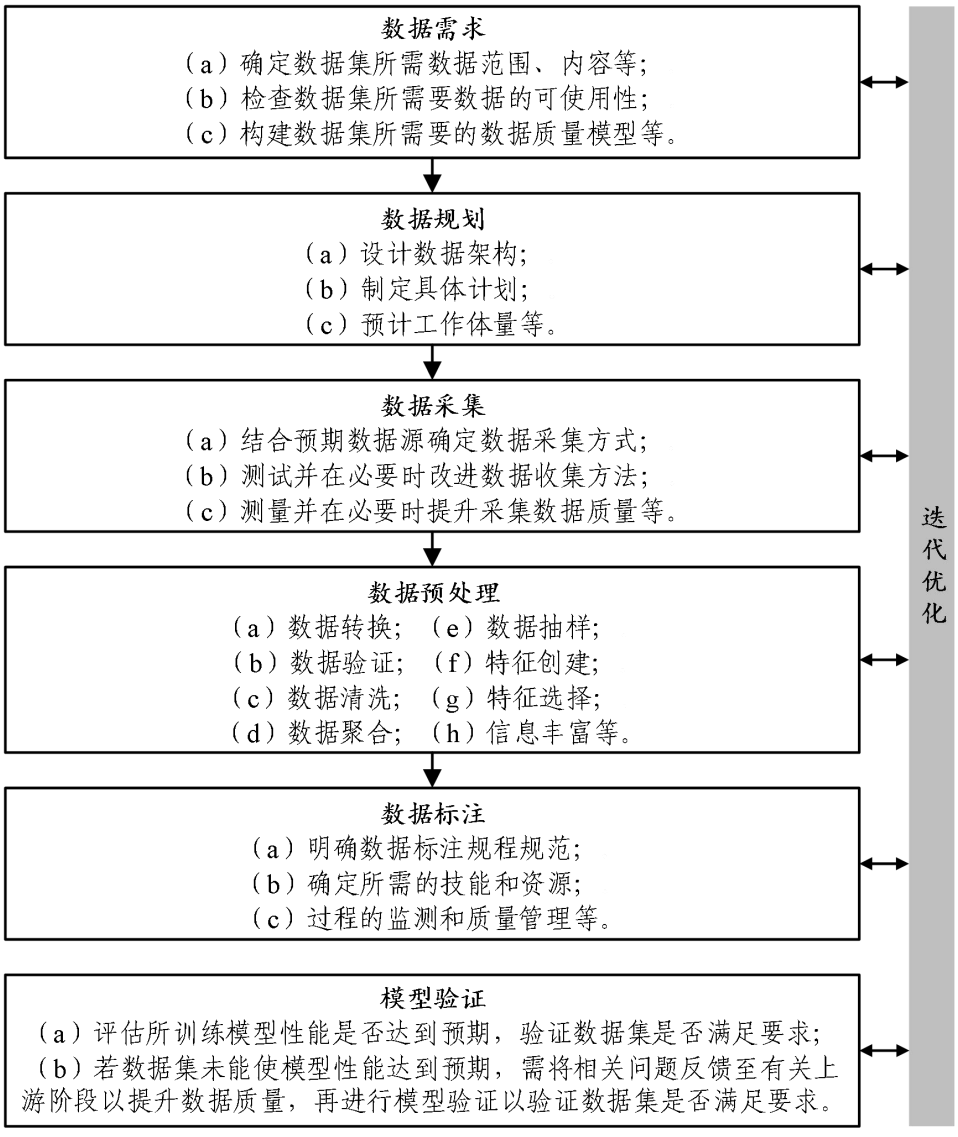

图 1 高质量数据集建设模式

（二）建设核心环节

1．数据需求

高质量数据集的数据需求环节主要是确定人工智能应用对数据的需求，即根据特定人工智能用途，明确数据集在数据范围、内容、可用、质量等方面的需求。在数据范围和内容方面，根据预期人工智能应用明确需要哪些具体数据，包括数据格式、统计特性和可分性等；在数据可用方面，检查数据集所需要数据的可使用性，即确认用于特定人工智能应用的数据是否可获取并使用；在数据质量方面，构建数据集所需要的数据质量模型，即实例化一个具有相关数据质量特征（例如完整性、准确性、一致性等）的数据质量模型。

2．数据规划

高质量数据集的数据规划环节主要是确保所用数据满足数据需求环节的要求，同时为使用这些数据完成人工智能应用的目标提供支持，主要涉及**设计数据架构**，即界定所需数据的全部属性、来源、范围等，以及如何使用这些数据；**制定具体计划**，即制定涵盖数据采集、数据预处理、数据标注、模型验证等环节的具体计划，包括各环节实施计划、数据质量计划等，以满足数据规范等方面要求；**预计工作量**，即预估获得和准备数据以支持特定人工智能应用所需的工作量，可能包括任何必要的数据重组、传输或收集的时间，以及为特定人工智能应用构建数据质量模型的时间。

3．数据采集

高质量数据集的数据采集环节主要是收集用于特定人工智能应用的数据，即从数据规划环节所确定的数据源收集的实时和历史数据。该环节主要涉及**结合预期数据源确定数据采集方式**，即根据所需数据是否已存在并可直接再利用、是否可转化现有数据来满足要求、是否可通过购买或许可获得数据、是否可以生成数据、是否需要采集新数据等情况，确定是以获取和组合现有数据集、生成数据（如模拟数据、合成数据等）、收集数据（如传感器采集、手动输入等）等之中何种方式采集数据；**测试并在必要时改进数据收集方法**，即如需收集新数据，则要测试数据收集方法，在必要时调整相关配置和参数设置、操作条件、传感器规格和安装位置等，以满足相关数据收集规范要求；**测量并在必要时提升采集数据质量**，以降低采集环节数据质量问题引入下游环节的风险，避免为下游环节增加不必要的工作量。

4．数据预处理

高质量数据集的数据预处理环节主要是将所收集到的数据处理成可供数据标注等后续环节使用的形式。该环节涉及以下可选过程：**数据转换**，以最小的内容损失，将数据从一种表示或空间转换为另一种表示或空间；**数据验证**，根据验证正确性、有意义、安全性、隐私性等数据质量特征，确保数据是正确的；**数据清洗**，检测错误数据或缺失数据，并通过替换、修改、输入或删除等方式修正数据；**数据聚合**，将两个或多个数据集以汇总的形式合并为一个数据集；**数据抽样**，从数据集中选择数

据，抽样可以替换或非替换方式进行；**特征创建**，创建比原始特征更能有效捕捉数据中主要信息的新特征；**特征选择**，使用可用特征的子集来降低数据的维数；**信息丰富**，链接各类数据源，并为数据增加额外的上下文语境。

5．数据标注

高质量数据集的数据标注环节主要是针对有监督机器学习的，其训练、验证和测试数据需要对单个或多个目标变量赋值。该环节可以涉及以下可选过程：明确数据标注规程规范、确定所需的技能和资源（如标注人员、工具、平台等）以及对数据标注过程进行监督和质量管理。

6．模型验证

高质量数据集的模型验证环节主要是将所准备的数据用于人工智能模型开发和训练，对模型性能是否达到预期进行评估，以验证数据集是否满足要求。若模型性能达到预期，则表明数据集已满足要求。若模型性能未达到预期，则可采取以下步骤。**一是**对于人工智能模型，确定数据集相比于算法，是不是致使模型性能未达到预期的根本原因。**二是**对模型验证环节所发现的数据质量问题进行分析，将对模型性能产生不利影响的数据质量问题反馈给上游环节，以改进相关环节的数据质量。**三是**重复数据规划、数据采集、数据预处理、数据标注等环节以提升数据质量；**四是**重建人工智能模型，对模型性能进行评估。

（三）建设核心技术

高质量数据集建设的每个核心环节均依托相应的核心技术

形成完整的数据集构建体系，主要包括数据采集技术、数据转换技术、数据清洗技术、特征选择技术和数据标注技术。本节将围绕这些核心技术，介绍其当前的发展趋势与典型实践。

1．数据采集技术

数据采集是指通过软硬件手段从多种来源中获取原始数据的过程，为人工智能模型训练、大数据分析和业务决策提供基础数据支撑，对应于高质量数据集建设中的数据采集环节。

随着人工智能、大数据等技术的不断发展，数据采集技术呈现出多源融合、自动化运行和边缘智能等创新趋势，多个行业和地方已在实践中取得初步成果，本指引介绍三种主流的技术。**一是多源异构数据融合采集技术**，支持对结构化、非结构化等多类型数据的统一采集和调度，广泛应用于工业、政务等复杂系统场景。例如，国家工业互联网平台在制造企业中部署多类传感器与控制器，实现设备层数据的高频融合采集，支撑工业模型训练所需的全流程数据获取。**二是边缘侧数据采集技术**，通过在数据源附近部署边缘设备，实现对实时数据的本地采集、预处理与上传，提升采集效率并降低网络压力。华为智慧园区解决方案采用边缘网关，实时采集环境监测和视频数据，保障数据的即时性和安全性。**三是生成模型辅助等数据合成采集技术**，针对数据稀缺或敏感的应用场景，利用仿真、统计、生成对抗网络（GAN）或扩散模型等技术模拟生成符合真实分布的高质量数据。例如，清华大学在医疗影像领域采用数据合成手段构建补充数据集，为模型提供多样化训练样本，解决实

际采集受限问题。

2．数据转换技术

数据转换是指以最小的内容损失，将数据从一种表示或空间转换为另一种表示或空间，旨在增强数据的一致性、兼容性和可分析性，对应于高质量数据集建设中的数据预处理环节。

近年来，数据转换技术正朝着自动化、智能化和标准化方向发展，当前主要包括以下几类主流和创新思路。**一是基于规则引擎的结构化转换技术**，通过定义转换规则，实现数据格式的统一和标准化。这一技术在医疗和能源行业得到广泛应用，例如中国移动医疗健康平台构建了规则驱动的数据转换系统，有效实现了医疗数据的跨系统标准化管理。**二是基于语义理解和知识图谱的数据转换技术**，利用自然语言处理和语义匹配，实现异构数据的语义映射和智能转换。京东在商品数据管理中应用此类技术，提升了多源数据的集成和智能化处理能力。**三是面向多模态数据的转换技术**，针对图像、视频、文本等不同模态数据，开发专用的转换框架和接口，支持跨模态数据的统一处理。深圳市智慧交通项目采用多模态数据转换技术，实现了交通监控视频、传感器数据与文本信息的高效融合应用。

3．数据清洗技术

数据清洗是指检测错误数据或缺失数据，并通过替换、修改、插入或删除等方式修正数据，旨在提升数据的质量和一致性，对应于高质量数据集建设中的数据预处理环节。

随着数据来源和类型的多样化，数据清洗技术不断发展，

向自动化、智能化和大规模处理方向迈进。当前，数据清洗技术主要包括以下几类主流和创新思路。**一是基于规则的自动化清洗技术**，通过预定义的数据验证规则和异常检测算法，快速识别重复、缺失和格式错误等问题，在传统金融和制造行业被广泛应用。例如，中国工商银行构建了完善的规则引擎体系，实现了对海量交易数据的高效清洗与校验。**二是基于机器学习和深度学习的智能清洗技术**，利用模型自动发现复杂数据中的异常模式和错误，提升数据清洗的准确性和适应性。阿里巴巴集团在电商大数据处理中应用了多种智能清洗算法，显著提升了数据质量保障能力。**三是面向大规模分布式环境的数据清洗技术**，结合云计算和大数据平台，实现海量异构数据的并行清洗和实时更新。华为云数据治理平台通过大规模分布式清洗框架，有效支持多个行业的复杂数据治理需求。

4．特征选择技术

特征选择是指从原始数据中筛选出与目标变量关系密切、信息量丰富且冗余较少的特征子集的过程，旨在减少计算复杂度，提升模型训练速度，对应于高质量数据集建设中的数据预处理环节。

近年来，特征选择技术正朝着自动化、智能化和高效化方向发展，当前主要包括以下几类主流和创新趋势。**一是基于统计和过滤方法的特征选择技术**，通过相关系数、卡方检验等指标快速筛选重要特征。此类技术在传统金融风控领域广泛应用，例如平安保险利用统计特征选择技术，优化信贷风险评估模型，

提高了模型准确性和稳定性。**二是基于嵌入式和包装方法的特征选择技术**，结合机器学习模型训练过程动态选择特征，提升选择效果。华为在通信网络故障诊断中采用该方法，实现了对海量网络数据特征的自动筛选和故障预测。**三是基于深度学习的自动特征提取与选择技术**，利用神经网络结构自动学习有效特征，适应复杂多样的数据场景。清华大学在智能制造领域开展相关研究，推动深度特征选择技术在设备故障预测中的应用。

5．数据标注技术

数据标注是对原始数据进行加工处理，形成可服务于人工智能模型训练、数据挖掘分析等活动必需的高质量数据集的关键技术，标注质量往往直接影响人工智能模型的训练效果和性能表现，对应于高质量数据集建设中的数据标注环节。

随着数据类型和应用场景的多样化，数据标注技术不断演进，逐步实现从传统人工标注向智能化、自动化标注的转变，主要集中在三大主流方向。**一是半自动化标注技术**，通过引入人工智能辅助工具，减少人工劳动强度，提高标注效率与一致性。例如，阿里巴巴智能标注平台利用自动标注预处理结合人工复核，实现了海量电商图像与文本数据的高效精准标注。**二是众包标注与分布式管理技术**，搭建规模化协作平台，整合大量标注人员资源，解决大规模数据标注的人力瓶颈问题。广东省政务数据中心借助此技术，构建了覆盖多部门的众包标注体系，有效支撑了政务数据的智能化应用。**三是主动学习与模型辅助标注技术**，利用模型预测指导标注优先级，提高标注资源

的利用效率。清华大学在该领域开展创新研究，推动多模态数据的精准标注和智能应用。

（四）数据集质量评价

为系统提升高质量数据集的建设能力与应用水平，必须构建科学规范的数据集质量评价工作体系。质量评价不仅是衡量数据集是否满足“高质量”标准的基本途径，也是推动数据集标准化建设、促进其可信流通与高效应用的重要抓手。为此，一方面，明确质量评价在数据集建设中的作用，理清其工作流程和关键环节；另一方面，构建覆盖数据全生命周期与实际应用需求的质量评价指标体系，确保评价结果的科学性与指导性。

1．质量评价定位与实施流程

数据集作为人工智能模型开发与训练的基础资源，其质量水平直接影响模型性能和实际应用效果。开展系统性、规范化的数据集质量评价，是判断数据集是否达到“高质量”标准的基本路径，也是推动高质量数据资源建设的核心抓手。通过“以评促建、以评促用”，可以有效倒逼数据生产和管理环节提质增效，全面提升数据资源的可用性、可信度与应用价值。制定统一的高质量数据集质量评价规范，明确评价工作的组织方式、指标要求和操作细则，对于提升优质数据供给能力，推动数据共享流通，强化人工智能模型训练支撑能力，具有重要意义。

从实施路径看，高质量数据集的质量评价通常包含若干关键流程与环节，形成闭环式管理机制，确保评价过程系统、规

范、可溯源。

一是评估准备阶段：在正式开展评价工作前，需明确数据集的基本信息、应用场景与评价目标，划定评价范围和对象类型，制定相应的评价策略和技术规范。同时，组织配备具备专业能力的评估团队，准备相应的评价工具和数据支撑环境，确保评价工作的规范性与一致性。**二是质量评估指标体系构建与实施阶段：**该环节是整个质量评价工作的核心，需要设计科学合理的质量评估指标体系，明确各项指标的评测标准和实施细则，结合自动化检测与人工核查等方法，开展全面系统的质量评估，确保评价过程规范、全面且具操作性。**三是综合评价与反馈应用阶段：**基于前述质量评价指标的评价结果，结合多维度指标体系进行加权汇总，形成定量化的质量评分与等级划分。同时，针对发现的问题，提出可操作的改进建议，形成评价报告，为数据集的发布、认证、共享流通及后续版本优化提供决策支撑与专业依据。

2．质量评价指标体系构建

为全面系统地评估数据集的质量水平，科学指导高质量数据集建设与应用，参考技术文件《高质量数据集 质量评测规范（征求意见稿）》构建涵盖**说明文档**、**数据质量**、**模型应用**三个维度的质量评价指标体系。该体系立足于数据全生命周期管理要求，兼顾数据的描述规范性、本体质量和应用效果，能够有效反映数据集的完备性、实用性与发展潜力。为增强评价的针对性和可操作性，以下重点介绍三个维度的评价细则与核心

要求。

说明文档指标主要评价数据集所配套的文档资料是否清晰、完整、规范，是数据集可理解性与可重用性的基础保障。该指标包含基本信息、内容特征、建设过程及应用说明等关键指标的完整性评价。在基本信息完整性方面，应包含数据集规模、格式规范、文件结构、访问渠道、技术支持方式等基本信息；在内容特征完整性方面，应包含模态类型、数据分布情况、标签类别统计、样本示例、局限性说明等内容特征；在建设过程完整性方面，应包含数据来源、采集方法、加工处理流程、标注规范、版本控制等建设过程；在应用说明完整性方面，应包含使用许可、目标应用场景、评估方法、基准测试结果、典型应用案例等应用说明。

数据质量指标直接衡量数据本体的质量是否满足人工智能模型开发和训练的基本要求。关键指标包括：格式规范性，数据集中数据的格式符合预定标准，可直接用于人工智能模型开发和训练；安全规范性：数据集中数据符合人工智能模型开发和训练的安全要求，应不包含违反社会主义核心价值观的内容、歧视性内容、商业违法违规、侵犯他人合法权益等非法内容；标注规范性：数据集中数据的标注符合预定的标注规范，遵循预先设定的规范化流程；结构完整性：数据集描述数据的元数据完整，不包含缺失值或缺失值应在合理范围内；内容真实性：数据集中数据真实可追溯。非合成数据能追溯到采集源头，能与采集源头保持一致，不存在未经说明的篡改；合成数据能追

溯到生成算法和过程，且能符合目标场景真实数据的分布规律；内容一致性：数据集中相关联的数据间内容一致，能在语义和表达上保持匹配，包括不同模态数据间的一致性和同模态数据间的一致性；类型一致性：数据集中数据符合其所属数据集类型的要求，通识数据集中数据应蕴含通用知识，行业通识数据集中数据应蕴含行业领域通用知识，行业专识数据集中数据应蕴含行业领域专业知识；内容干净性：数据集中数据经过严格清洗处理，不包含脏数据。

模型应用指标强调数据集应能有效支撑目标场景人工智能模型的开发和训练。该指标主要包括以下关键子指标：内容多样性：数据集的数据分布全面程度应满足目标应用场景人工智能模型开发和训练的要求；规模完整性：数据集的规模满足目标应用场景人工智能模型开发和训练的要求；内容时效性：数据集中数据的采集时间和更新状态满足目标应用场景人工智能模型开发和训练的要求；标注准确性：数据集中数据的标注能精准标记出目标应用场景人工智能模型开发和训练所需的所有信息；模型适配性：数据集是能有效提升目标应用场景人工智能模型的性能。

通过上述三个维度协同构建的评价指标体系，能够实现从数据文档规范、数据本体质量到模型应用效果的全过程质量控制与多维度系统评估，不仅为数据集建设单位提供明确的质量提升方向，也为评测机构、使用方等相关主体提供客观可依的评判依据。

3．质量评价工作体系建设

构建统一的高质量数据集质量评价体系，是数据时代发展的必然要求。它能消除测评指标差异导致的“数据质量迷局”，让不同领域、不同机构的数据集在同一标准下接受检验，实现质量水平的横向可比，为数据共享、跨域应用扫清障碍。同时，统一体系可规范测评行为，提升整体测评能力，通过明确的标准和流程减少人为误差，让测评结果更具公信力。而有效的监督机制还能倒逼数据生产和管理方重视质量提升，从源头保障数据价值的充分释放，为数字经济、科研创新等领域提供坚实的数据支撑。构建统一的高质量数据集质量评价体系，需从多方面着手。

在构建原则方面，应遵循全面性、客观性、可操作性与动态适应性原则。全面性要求体系涵盖数据的准确性、完整性、一致性、时效性、安全性、可用性等各个关键质量维度，确保对数据集质量进行全方位考量；客观性即测评过程和结果不受主观因素干扰，依靠科学合理的方法与标准得出结论；可操作性意味着各项测评指标和流程在实际执行中切实可行，便于相关人员操作实施；动态适应性则使体系能够随着数据应用场景的拓展、技术的革新以及新数据质量问题的出现，及时调整优化测评内容与方式。

在指标设计层面，要打造层次清晰、结构合理的体系。设置一级指标，如准确性指标可通过对比原始数据与权威数据源，计算误差率、偏差度等量化指标进行评估，以反映数据与真实

情况的契合程度；完整性指标关注字段完整性、记录完整性以及数据一致性，统计缺失值、重复值和异常值的处理情况以及数据结构的完备性；一致性指标聚焦不同来源、不同时间数据在定义、格式、类型上的一致性以及跨系统、跨部门的数据同步情况；时效性指标考量数据更新周期、加载时间、同步时间等，衡量数据更新的频率与速度是否满足需求；可用性指标评估数据是否易于获取和使用。同时，针对每个一级指标进一步细化二级指标，形成完整的树状结构指标体系，以便更深入、细致地剖析数据集质量。

在评价流程方面，首先确定测评目标，明确是针对新构建数据集的质量评估，还是对已使用数据集的定期审查，抑或为特定项目筛选合适数据集等目标。然后制定详细的评价计划，依据测评目标选取适用的评价指标，确定样本选取方法、评价工具以及人员安排等。在数据采集阶段，确保样本具有代表性，能真实反映数据集整体特征。接着开展数据测评，综合运用统计分析法对数据进行量化分析，利用自动化检测工具进行快速筛查，组织专家评审从专业角度把关，收集用户反馈了解实际使用感受等多种方式。对评价结果进行深入分析，找出数据集质量问题所在，按照问题的严重程度和影响范围进行分类排序，为后续改进提供清晰方向。最后，根据结果分析制定针对性改进措施，跟踪改进措施的实施效果，形成闭环管理，持续提升数据集质量 。

在监督机制方面，国家数据局将会同相关部门建立专门的

监督机制，明确高质量数据集质量评价机构能力要求，指导制定数据质量评价管理规范，并负责对评价过程和结果进行抽查审核，确保评价的公正性与准确性。同时，将评价结果公开透明化，接受社会监督，对于评价过程和评价报告质量不达标的评价行为和机构依据管理规范进行管理，要求评价机构限期整改，并公布整改情况，以此推动全国范围内高质量数据集质量评价体系的建设与发展。

五、高质量数据集建设运营体系

（一）高质量数据集体系规划

体系规划是高质量数据集建设运营的前提，通过知识索引构建、数据资源盘点、标准体系搭建三大环节，分别驱动数据知识化、场景适配性、全周期标准化，为后续数据集的具体建设和高效运营提供清晰的蓝图与坚实的行动指南。

一是针对智能化需求，搭建行业知识索引框架。结合行业业务逻辑和模型需求，提炼核心知识节点，搭建层次化知识架构，将既有数据资源与知识索引精准匹配，实现数据的知识化归类。例如，医疗行业聚焦疾病诊断与药物研发，金融行业则围绕风险控制和客户营销。该索引为模型应用提供结构化路径，支持快速调用关联数据，有助于加速模型训练迭代，将数据资源高效转化为驱动业务创新的智能生产力，实现从数据到模型的价值跃升。

二是锚定智能场景，绘制行业数据集资源地图。通过深入分析模型应用的业务场景，全面梳理企业内外部数据资源，包括内部结构化数据、非结构化数据、半结构化数据，以及外部公开数据与合作数据，形成完整的数据资源目录。盘点数据资源目录清单，可视化呈现数据的分布、权属关系、质量状态及采集渠道、存储位置、更新频率等关键信息，形成“数据资源地图”，为后续数据采集与整合提供可操作性指引。

三是围绕高质量数据集建设运营环节，构建全链条、全行业标准体系。重点围绕基础通用、关键技术、质量控制、工具

平台、流通交易、行业应用以及安全保障等方面，建立健全高质量数据集标准体系，为模型开发方、数据运营方与管理方建立统一标准。在生产环节，制定数据采集、处理、标注等标准，规范数据生产流程；在质检环节，明确质量评估指标，规范自动化工具检测、人工抽检及模型反馈流程；在技术环节，统一数据清洗、标注、存储等技术工具的开发标准和要求，确保工具兼容性与易用性；在应用环节，制定数据与模型对接标准，规范资源运营与风险管理流程。此外，加速研制行业高质量数据集建设标准，规范各行业高质量数据集的建设流程、技术要求与质量评价体系。

（二）高质量数据集工程建设

工程建设是高质量数据集体系规划落地的实施阶段，涵盖研发、交付、运维等核心建设环节。

研发环节聚焦数据集生成流程的系统性管控，包含需求管理、设计管理和数据加工三个子环节。在需求管理环节，收集、分析、确认业务部门对数据集规模、模态、标注精度等的具体要求，明确其优先级和合理性；在设计管理环节，构建覆盖数据集质量、安全、合规、采集、标注、存储的全流程规范体系；在数据加工环节，梳理并管控数据采集、预处理、标注、增强、合成等方面的技术要求，并由此开展具体的研发工作。通过三环节协同，确保数据集研发目标清晰、流程规范、处理标准，为后续环节奠定基础。

交付环节是对数据集交付过程的规范化管控，包括测试管

理和发布管理两个关键阶段。测试管理阶段，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、场景下可用性等要求。发布管理阶段，建立包含发布审批、接口管理、数据集管理的发布体系，将经过验证的数据集安全、高效、规范地转化为生产级服务，并实施版本管控，规范记录版本更新内容、责任人及时间戳，保障数据在长期演化中的可追溯性、一致性与可复现性。

运维环节是确保数据集交付后的持续稳定运行，涵盖监控管理及资源管理两个方面。在监控管理方面，建立数据集质量、系统性能、安全合规等维度的监控指标，实施日常监控与告警机制。在资源管理方面，对数据资源、计算资源和存储资源分别进行盘点和调度。对于数据资源，通过数据资产目录厘清数据分布、权属与质量状态；针对计算资源，优化任务调度与资源分配策略，平衡效能与成本，最大化集群利用率；针对存储资源，实施分级存储与生命周期管理，在成本、性能与可靠性间寻求平衡。系统性运维可减少因数据质量下降、资源不足或安全风险导致的服务中断，保障数据集长期适配业务需求。

（三）高质量数据集运营管理

运营管理是实现高质量数据集可持续发展的核心，需围绕用户需求响应、成本精细化管理、质量与安全维护及生态协同发展四个方面构建全流程管理体系，在使用过程中达成**“需求响应及时、成本精准可控、质量安全可信、生态价值共创”**的

四大目标，反哺过程建设环节，从而“以用促建”。

用户需求响应，旨在通过构建用户友好平台、建立动态迭代机制、推动跨场景复用，实现从“数据可用”到“价值可见”。首先，通过提供可视化工具、接口及详实元数据，构建用户友好平台，支持按场景、模态、质量等多维度检索。其次，基于用户反馈与模型效果，联动研发团队补充缺失样本、修正标注偏差，建立动态迭代机制，并主动更新数据以保持时效性。最后，通过知识关联与格式适配，推动跨场景复用，打破“一数据集一模型”局限。例如，通识文本数据集可同时支撑预训练与情感分析微调；工业故障数据集可关联设备参数，延伸至预测性维护等场景。

成本精细化管理，旨在通过成本核算、成本优化以及建立内外部成本结算机制，实现从“粗放投入”到“精准管控”。其一，量化核算人力、存算资源及技术工具成本，并基于历史数据与业务需求制定预算。其二，实施成本优化策略，应用自动化工具降低人力成本、优化资源调度、清理冗余数据，并推动技术工具跨场景复用以节约技术工具成本。此外，建立成本结算机制，内部按调用次数和样本下载量分摊成本，外部合作按数据质量、稀缺性及应用价值制定定价标准。

质量与安全维护，旨在通过建立质量监控体系和安全管控体系，实现从“交付合格”到“持续可信”。建立全生命周期质量监控体系，要求实时跟踪完整性、标注一致性、时效性等核心指标，通过自动化扫描与人工复核等手段处理异常，有效

保障数据集质量水平。建立覆盖数据集全生命周期的安全管控体系，严格遵循《中华人民共和国网络安全法》《中华人民共和国个人信息保护法》等法规，实施分级安全管控，如敏感数据去标识化处理。此外，为确保数据集长期满足模型训练的准确性与安全性要求，需规范版本控制，记录迭代内容、责任人及时间戳，支持历史回溯。

生态协同发展，旨在通过制定行业数据集共享、流通、共建与价值分配机制，实现从“单一运营”到“生态共赢”。一是制定分级共享策略，基础数据集可以通过数据交易所或开源社区开放，专有数据集可以通过可信数据空间等数据流通基础设施在授权范围内共享。二是遵循国家与行业标准规范数据格式、接口及权属界定，推动标准化流通。三是建立共建与价值分配机制，协同产业链研发工具，共建行业基准数据集与评测体系，按数据量、标注工作量等贡献度分配联合建设收益，拓展数据应用边界和市场影响力。四是完成生态运营，通过完善的数据集生态管理机制和运营流程规范，专业的生态运营团队和服务平台，建立高效生态健康度监测体系，实现多方的广泛认可和高效协同。

六、高质量数据集建设推进思路

作为一项系统性工程，高质量数据集建设工作需要政府、企业、科研机构等各方协同参与，从制度设计、技术攻关、生态培育等方面多管齐下，在各级数据主管部门统筹下凝聚共识、形成合力。我国高质量数据集建设的推进思路是以体系化思维优化高质量数据集建设布局，以设施化手段促进高质量数据集流通利用，以生态化环境保障高质量数据集可持续发展，构建覆盖全流程、贯通各环节的高质量数据集建设格局。

（一）体系化布局高质量数据集建设

完善的工作体系是推进高质量数据建设的基本前提。围绕建设主体、建设方向、数据来源、标准规范和技术能力，回答谁来建、建什么、怎么建等关键问题，形成层次清晰、体系完备的整体布局。

一是加强分工协作。理顺政府和市场的关系，明确各类建设主体的角色定位，构建协同体系。充分履行主管部门在政策引导、资源统筹和监督管理方面的职能，发挥企业在技术创新、市场运营和应用推广方面的优势，释放高校和科研机构在理论研究、人才培养和成果转化方面的潜力，调动行业协会、数据交易机构等在平台搭建、产业服务和生态聚集方面的作用，促进政产学研用分工协作。

二是突出建设重点。挖掘地方产业特色，有效利用资源禀赋和产业优势，因地制宜制定建设清单，避免无序竞争和重复投入。立足行业发展特性，建立产业图谱，整合行业资源，形

成适用于本行业的建设路线，促进各领域数据集均衡发展。聚焦典型应用场景，建设具有共性需求的高价值场景数据集，确保建设成果赋能智能化转型，形成示范效应。加强具身智能、低空经济等新领域高质量数据集前瞻布局，推动新兴产业与人工智能融合发展。

三是畅通数据供给。拓展多元化数据供给渠道，为建设高质量数据集提供充足“原料”。加快推进公共数据资源开发利用，建设公共数据授权运营平台，加大民生领域和重点行业数据开放共享。推动企业数据有条件开放，探索数据互换、数据联盟、数据交易等多种供给路径。鼓励“链主”企业与中小企业开展合作，整合上下游分散的数据资源，促进产业链数据融合。

四是完善标准规范。围绕数据集全生命周期，完善数据集标准规范体系，推进建设指南、格式、分类、质量评测等国家标准验证发布，组织制定数据标注、数据合成、建设运营能力评估等新一批国家标准。推动重点行业标准制定，结合行业特点细化数据采集、处理、加工、标注等流程和标准，为高质量数据集建设、流通和应用提供依据。加强标准解读和宣传，推动标准施行应用。

五是加强技术攻关。鼓励科研院所、大模型企业、数据标注企业等开展联合攻关，重点布局跨语言、跨领域、跨模态语义对齐、数据合成等攻关项目，支持专家标注、多模态标注、众包标注、质量评测等智能化工具研发和平台建设，提升数据

标注自动化水平。推动先进技术推广、普及和应用，为建设高质量数据集提供自主可控的核心技术能力。

（二）设施化推进高质量数据集应用

可靠的基础设施是高质量数据集流通利用的重要保障。通过构建**“平台+数据集+模型”**的一体化服务设施，降低数据集应用门槛，推动数据集市场化流通和规模化应用。

一是充分发挥各类数据基础设施效能。统筹国家数据基础设施与高质量数据集建设，在数据基础设施的全域功能节点、区域/行业功能节点和业务功能节点，部署不同应用范围的“数据生产车间”和“数据中试车间”，实现高质量数据集的规模化、标准化、体系化生产运营。利用可信数据空间的建设成果，通过隐私计算、区块链等技术，支撑数据资源整合，赋能数据集可信流通。依托各类数据交易服务平台建立高质量数据集交易专区，完善数据集交易功能，制定数据集上架审查、权属确定、价格形成等规则。

二是建立全国数据集统一目录体系。鼓励搭建“国家”和“地方/行业”两级数据集管理服务平台，实现数据集合规汇聚、高效检索、样例下载、质量评测等功能，绘制全国数据集资源地图，建立质量动态评价机制，促进数据集供需对接。鼓励地方建设数据集管理服务平台，基于地区和行业数据集提供个性化服务，并与国家级平台互联互通，促进数据集在供需主体间安全流通。

三是探索建设数据集集成应用平台。鼓励应用企业、模型

厂商、科研院所等主体联合建设数据集创新应用平台，面向重点行业提供覆盖数据集开发建设、应用服务、价值运营的全链条支撑能力，集成数据集建设工具、评测工具、流通环境和人工智能模型，构建整体解决方案，探索数据集利润分配机制，完善商业运营模式，激发社会主体创新活力，加速数据集应用落地。

（三）生态化赋能高质量数据集发展

良好的产业生态是高质量数据集可持续发展的动力来源。通过制度创新、产业协同和人才培育，构建多方共赢的生态体系，着力破解建设成本高、共享意愿低、创新动能弱等瓶颈。

一是搭建合作平台。成立产业共同体，建立完善政府、企业、科研院所等相关主体的合作机制，实现理论研究和产业发展相互促进、良性循环。推动各方建立和壮大数据标注产业联盟，打造数据清洗、标注、交易、应用的闭环产业链交流合作平台。鼓励政府引导与专项资金支持，探索数据集资产化路径，通过项目补贴等政策激励企业参与，形成多元投入的可持续运营形态。鼓励金融机构参与数据集交易市场，提供质押融资等金融服务，活跃数据集交易市场，促进高质量数据集流通与价值实现。

二是完善制度机制。加快构建数据要素基础制度体系，填补制度空白、破除机制障碍。探索数据集交易定价策略，形成以市场供需关系和模型训练效果为核心的价格形成机制。完善收益分配机制，鼓励商业模式创新，加快形成高质量数据集价

值循环体系，实现数据集的可持续开发和应用。探索建立委托授权、模型训练知识产权保护豁免机制，试点行业间、地区间联合共建共享开放交流机制。

三是加强人才培养。引导高等院校、职业院校加强学科建设，开设数据治理、数据分析等专业课程，为高质量数据集建设提供充足的人才储备。深化产教融合，鼓励校企合作建设实训基地，提高在校学生的实践技能，建设高素质的人才队伍。吸引数据科学家、工程师等行业专家参与重点领域高质量数据集建设，培养具有专业背景的高端人才。完善国家数据标注等职业认证体系，开展专项培训，规范数据标注人才发展，提高从业者的数据治理与跨领域协调能力。

四是推动共建共享。鼓励社会力量支持和参与开源生态建设，推动开源社区发展，促进数据集开放共享。建立揭榜挂帅、技术竞赛等机制，针对关键领域和重点行业智能化转型需求公开征集解决方案，打造一批高水平、有特色的行业数据集融合应用示范案例。开展数据集建设和应用案例征集等活动，营造社会各界共建共享的良好氛围。