

高质量数据集赋能人工智能产业发展研究

王 鹏 丁一桐

摘 要：在数智化转型加速推进的背景下，数据集作为人工智能（AI）大模型的关键要素，其高质量建设对于推动人工智能产业发展具有重大意义。文章基于数智化转型的历史方位，厘清了高质量数据集与人工智能的关系，结合实际，从顶层设计、地方和行业三个层面分析了高质量数据集建设现状，探讨了高质量数据集在数据资源、人工智能和应用生态三个领域的建设瓶颈，为破解高质量数据集建设难题，构建了一套涵盖技术路径、标准路径、供给路径、市场路径和存储路径五大路径的系统性建设框架，并提出构建体制机制、建立标准体系、精进技术工具、加强人才培养等对策建议，以推动人工智能产业发展。

关键词：数智化转型；高质量数据集；人工智能产业；建设框架

中图分类号：TP18;F49 **文献标识码：**A **文章编号：**1673-5706（2025）02-0036-09

引言

数智化是数字化与智能化深度融合的产物，代表着数字化发展到更高阶段的必然趋势。它不仅是数字化和智能化的简单组合，更是一种全新的发展模式或过程，通过现代信息技术如大数据、人工智能、5G 和区块链等，推动经济社会的数字化、智能化发展。党中央高度重视数智化转型及其视域下的人工智能产业发展，出台了一系列政策文件和纲领性文件，为推动人工智能产业发展提供了根本遵循。党的二十届三中全会通过的《中共中央关于进一步全面深化改革、推进中国式现代化的决定》中多处提及“人工智能”，对人工

智能产业未来发展作出了重要指引，如完善推动“新一代信息技术”“人工智能”等战略性新兴产业发展政策和治理体系，完善“生成式人工智能发展和管理机制”，建立“人工智能安全监管制度”等^[1]。习近平主席在向 2024 世界智能产业博览会致贺信中强调，人工智能是新一轮科技革命和产业变革的重要驱动力量，将对全球经济社会发展和人类文明进步产生深远影响。中国高度重视人工智能发展^[2]。习近平总书记在中共二十届中央政治局第二十次集体学习时再次强调，要充分发挥新型举国体制优势，坚持自立自强，突出应用导向，推动我国人工智能朝着有益、安全、公平方向健

康有序发展^[3]。高质量数据集是 AI 大模型实现更强的感知能力、学习能力、决策与推理能力、交互能力和多领域应用能力等的基础，因此高质量数据建设是人工智能产业发展的核心驱动力之一，其建设水平直接影响国家人工智能产业的发展格局。本文深入探讨数智化视域下高质量数据集推动人工智能产业发展的逻辑框架与发展思路。为推动人工智能产业发展提供理论支撑和实践参考。

一、数智化转型的理论内涵

（一）数字化与智能化的内涵关系

数字化、智能化概念是数智化概念的基础，是数字经济时代创新发展的运作逻辑，笔者从技术迭代、决策赋能等方面分析两者之间的联系。

化繁为简：数字化与互联互通。数字化是指将信息转化为数字形式，并通过数字技术进行处理、存储、传输和分析的过程^[4]。数字化持续应用数字技术，推动社会经济发展，表现为主体通过数字信息或数字化程序认识客体、掌握和作用于客体^[5]，为主体提供了新的价值创造机会。数字化实现了不同空间的相互连接，从而催生新的活动或对现有活动进行变革。其核心含义包括两个方面：一方面，利用数字技术联系数字空间与社会空间；另一方面，借助数字空间的力量，促使物质空间与社会空间之间以及内部的联系变得更加紧密^[6]。

运筹帷幄：智能化的决策赋能。智能化是依托大数据、云计算、人工智能等新一代信息技术，使系统自动学习、推理和决策，从而实现类似于人类智能的行为。智能化通过一定程度上模拟人类思维模式，产生类似人类智能的结果。其核心是智能系统利用现有知识，对大数据进行识别、分析、推演和整合，并通过挖掘数据相关性，提炼核心要素，从而转化为高度智能^[7]。在智能化决策过程中，人工智能不仅替代人类执行重要任务，更扩展至创造性解决开放性问题，推动决策科学从“经验驱动”到“算法驱动”的范式革命。

相得益彰：数字化与智能化的协同作用。随着数字技术的迅猛发展和智能装备的广泛普及，“数字化”与“智能化”之间的联系愈发紧密^[8]。

智能化所涉及的大数据、云计算、人工智能等数字技术，相较于数字化更具有交互性和创新性。制造业的智能化并非无源之水，而是基于前期技术积累，并以人工智能和新一代信息通信技术等前沿技术为产业变革的关键驱动力^[9]。当前，一种以数字化模型为基础、智能化装备为工具，以“数字化+智能化”为核心内涵的新兴模式正在蓬勃发展^[10]。数字化的基本特征是大数据的分析与处理，智能化则以机器学习和人工智能为核心要素^[11]。智能化展现了数字技术应用的未来趋势，是数字化与信息化在实践中的深化与拓展，通过对人的主体性的确认，将数字智能技术与人的需求结合起来，发挥实际作用。同时，智能技术的运用也离不开数字化所实现的互联互通基础^[12]。

数字化是智能化的基础，它通过将信息转化为可处理的数据形式，实现数据的采集、存储、管理和分析，为智能化提供了必要的“原材料”，在数字化的基础上构建起完整的数据体系，从而支持智能化技术的运行。智能化则是数字化的高级发展阶段，它借助人工智能、机器学习、大数据分析等先进技术，赋予系统自主学习、预测和决策的能力，进一步提升数据的价值和应用效率^[13]。智能化不仅能够优化数字化过程中的数据处理，还能通过实时性和自动化的特征，实现更高效、精准和智能的业务流程和决策支持。因此，数字化与智能化相辅相成，数字化为智能化提供了数据支撑和基础架构，而智能化则为数字化赋予了更强大的功能和更高的价值，二者结合推动了各行业的转型升级，成为现代技术发展的重要驱动力（见图1）。

（二）数智化的特征

在数字化初期，互联网、大数据和云计算技术以数据为核心驱动力，广泛渗透到各行业。进入数智化阶段，智能技术的产业化进程加快，形成了“AI+ 数据智能应用”与“AI+ 服务创新”的协同模式。数字化转型持续推进，人工智能等智能化技术逐渐成为变革的核心动力，强大的数字化能力与人工智能技术相结合，匹配前瞻性的战略规划。因此，本文将数字化与智能化的融合简称为“数智化”。数智化是数字化与智能化的深

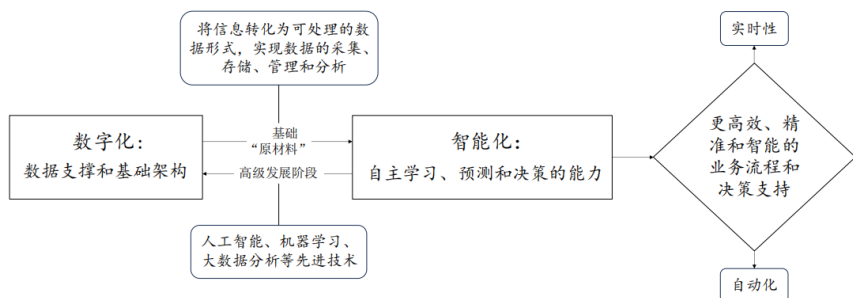


图 1 数字化与智能化的关系

度融合，强调通过数据驱动和智能技术的应用，实现业务流程的优化、创新和升级。数智化代表着数字化发展到更高阶段的必然趋势，是一种全新的发展模式或过程，通过现代信息技术如大数据、人工智能、5G 和区块链等，推动经济社会的数字化、智能化发展。数智化在信息数字化转化（转型初期）和流程数字化升级（成长期）的基础上，进一步开展业务智慧化转型实践（成熟期），体现了从基础数字化到高级智能化的进阶过程。从更深层次看，数智化的完整含义可以理解为：拥有数字智商的人运用数智技术获取和生产数据智能，以提供决策依据，从而实现数智化。数据智能、数智化、数智技术和数字智商是构成数智概念的四个核心要素。数智化的主体是人，客体是机器或物体，强调了人机协作与融合的重要性。

数智化的特征主要体现在数据驱动、智能决策、流程优化、创新驱动和用户体验提升等方面。其核心逻辑是人机互融、虚实同构以及算法主导。首先，在大数据、云计算等数字技术的基础上，数智化进一步引入了深度学习、机器学习、知识图谱等新兴技术，加强并实现了技术多维衔接与融合发展。其次，数智化强调数据的深度应用，让数据反哺实践，赋能运行和管理，推动主体从数据驱动向智能决策的升级。此外，数智化还通过数据的实时分析和动态调整，进一步优化业务流程，提升企业的运营效率 and 创新能力。这种以数据为核心、以智能技术为支撑的模式，不仅能够帮助企业更好地理解客户需求，还能通过个性化服务和精准营销提升用户体验。在实际应用中，数智化技

术已被广泛应用于智能制造、智慧金融、智慧城市等领域，成为推动各行业转型升级的重要力量。

二、高质量数据集赋能人工智能产业发展

（一）高质量数据集与人工智能的关系

数据集：训练 AI 大模型的基础。数据集是指一组用于训练和评估 AI 模型的数据，其质量和多样性对 AI 大模型的性能和效果至关重要。（1）数据量：AI 大模型，特别是像 GPT、BERT、DALL·E 这样的语言和生成模型，通常需要大量的数据进行训练，甚至是数十亿到数万亿个数据样本。数据量的增加可以让模型学习更多的知识和规律。（2）数据种类：大模型通常需要多种多样的训练数据，比如文本、图片、音频、视频等。这些数据种类的多样性能够帮助模型具备更广泛的应用能力。（3）数据质量：数据的准确性、无偏性和代表性对大模型的效果有直接影响。垃圾进、垃圾出（Garbage In, Garbage Out）是数据质量差的模型常见问题。

AI 大模型：从数据中学习。AI 大模型的主要任务是从数据集中提取有用的信息和规律，可以通过以下方式学习数据：一是监督学习。数据集通常包含输入数据及其对应的标签（例如，图像分类中的每张图像和它的标签），模型通过大量样本的训练，学习从输入到输出的映射关系。二是无监督学习。无标签的数据集也可以用于训练模型，尤其是在特征学习和聚类任务中，模型通过寻找数据的内在结构进行学习。三是自监督学习。在一些大模型（如语言模型）中，数据集本身提供了部分标签，模型通过自我生成标签学

习，这种方法在处理大规模文本数据时非常有效。四是生成式模型。对于一些生成式模型（如 GPT 系列、图像生成模型等），它们不仅需要理解数据中的规律，还需要生成与输入数据相似的新数据，模型的生成能力依赖于数据集的广度和多样性。随着数据集规模的增加，训练大模型所需的计算资源和时间也大幅增加。使用数百万到数十亿的数据进行训练需要大量的计算力（如 GPU、TPU 集群）。AI 大模型的规模通常与数据集的规模成正比。更大的数据集可以支持更大的模型，反之亦然。因此，大规模预训练模型（如 GPT-3、GPT-4）需要海量的文本数据以实现优异的性能。

数据集和 AI 大模型之间的关系是相辅相成的。数据集为大模型提供了训练所需的信息，而大模型通过学习这些数据集中的规律、特征和结构，最终实现智能化的预测和推理能力。数据的质量、规模和多样性直接影响大模型的性能。因此构建高质量的数据集是训练有效大模型的关键步骤之一，高质量数据集是促进人工智能和机器人技术创新的“核心燃料”，高质量数据与人工智能的结合，将会进一步发挥数据和人工智能的倍增效应（见图 2）。

（二）高质量数据集建设现状

顶层设计层面。国家高度重视高质量数据集的建设，数据要素领域政策文件相继出台。如 2024 年 1 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）》，聚焦十二大重点行业领域，推动数据要素高质量供给，努力建设好高质量数据集，以发挥数据要素价值乘数效应。

地方层面。各省市积极响应国家号召，积极推动高质量数据集建设。湖北省发布《湖北省高质量数据集白皮书》，明确数据集标准体系、平台架构及治理流程，推动数据提质、联创和共享；首批认证 10 个高质量数据集，覆盖 11 个重点领域；探索数据交易平台，借鉴电商模式促进数据流通，鼓励企业共建地方特色数据集等。广东省提出“四步走”推动全省高质量数据集建设：一是深化数据空间建设，构建行业数据资源库；二是征集优质数据案例，赋能行业智能应用；三是加强政产学研合作，强化核心技术研发；四是完善数据交易机制，保障数据有序流通^①。

行业层面。多个重点行业高质量数据集建设不断突破。截至 2024 年底，北京国际大数据交易所累计发布约 300 个高质量数据集，形成 10 余个应用领域数据资源地图，对接 100 多个市场主体参与高质量数据集的交易，并通过动态更新的数据资产图谱实现供需精准匹配^①。中国信息通信研究院率先构建我国首个面向行业的人工智能数据集质量评估体系，通过建立涵盖完整性、合规性、可用性等维度的量化评价模型，填补了行业数据质量认证体系的空白。由智源研究院主导研发升

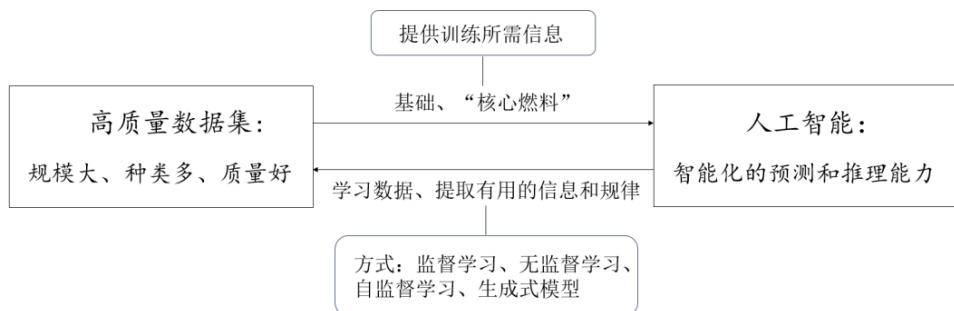


图 2 高质量数据集与人工智能的关系

① 数据来源：2025 北京数据交易成果报告会。

级的 IndustryCorpus 2.0 语言模型实现重大突破，相较于前代 IndustryInstruction 1.0 版本，新增了 12 个重点行业，在垂直领域数据储备和专业知识覆盖上实现质的飞跃^①。

（三）高质量数据集建设的难点

数据资源领域。（1）数据采集难题：许多企业面临着历史数据标准化和电子化的难度，相关数据的采集成本高周期长。此外，新生成数据的采集往往需要额外投资软硬件设备，投入较大。

（2）数据治理能力不足：由于行业内数据标准缺乏，企业的数据治理能力较弱，导致数据质量不高，开发和利用数据的成本较大。数据要素全国统一大市场尚未建立，制约了数据的共享和开放。

（3）跨区域、跨企业的数据流通与共享存在障碍：数据流通和共享机制的缺失使得不同企业间的数据共享效果较差。单一企业进行 AI+ 数据集的建设时，面临较高的综合建设和运维成本，而更大规模的行业或产业数据集可以有效提升大模型训练的效果和效率。（4）数据质量评价标准不健全：虽然我国现行相关法律法规已对数据问题作了原则性规定，但具体规则尚不明晰，缺乏明确的操作指南和实施细则。（5）行业差异导致的构建难题：一方面，领域专家需求难以满足，许多行业需要领域专家的参与，特别是在数据标注、模型验证、特征工程等方面。这种跨领域的知识融合增加了数据集构建的复杂性。另一方面，领域知识难以编码，AI 模型的训练往往依赖于领域专家对数据的理解，但很多行业的复杂知识难以直接编码到模型中，如何将这些复杂的专业知识转化为可用的训练数据仍然是一个难题。

人工智能领域。（1）算力集群建设困难：构建高效、安全的人工智能算力集群是训练 AI 模型的核心资源，单一企业要建设具有可控性的智能算力基础设施需要巨额的投资。（2）模型精度和有效性需要更多验证：目前，许多企业在场景化 AI 大模型的训练仍处于起步阶段，模型的精度和

有效性尚需大量的数据支持和多次验证，需要更多时间检验和完善。（3）智能模型深度融合企业生产运营：将 AI 大模型深度融入企业主要生产经营活动中，实现核心能力与经济效益的闭环，是一个长时间的发展过程。为了促进这一过程，更多行业资源需要进行统筹，避免行业内部的无效竞争。

应用生态领域。（1）安全合规能力不足：AI+ 数据集的开发和人工智能发展都需要坚实的安全合规专业能力的支持。目前，许多企业在数据资源流通和 AI 应用开发方面存在较为显著的安全合规短板。（2）数智化产业生态的不足：目前，许多企业在 AI+ 数据集建设方面存在产业生态尚不完善的问题。如何更好地融合地方资源，并找准更多 AI+ 产业应用场景，是亟待解决的关键问题。（3）人才短缺：AI+ 数据集建设需要既懂信息技术又懂行业经营的高端复合型人才，但目前许多企业在人员编制、薪酬待遇等方面存在一定限制，导致人才队伍建设的不足，影响了 AI+ 数据集建设和应用的效率。

三、高质量数据集的建设框架和对策建议

（一）高质量数据集的建设框架

高质量数据集已成为人工智能产业发展的核心要素，但当前高质量数据集的建设在数据资源、人工智能、应用生态等方面都存在着难点，制约人工智能产业的发展。破解这些难题，需要从技术路径、标准路径、供给路径、市场路径、存储路径等多个维度协同发力，构建系统化的高质量数据集建设框架（见图 3）。

1. 技术路径。智能化技术路径是实现数据质量提升与管理效能优化的关键手段。这一路径并非仅依赖于单一技术的突破，而是通过系统性地整合多种技术环节，逐步完善数据生态体系。在技术赋能的同时，它还需要与制度建设和市场机制紧密结合，以确保数据资源的流通和使用既高效又合法合规。

^① 数据来源：智源研究院数据与行业应用 Workshop 第三期。

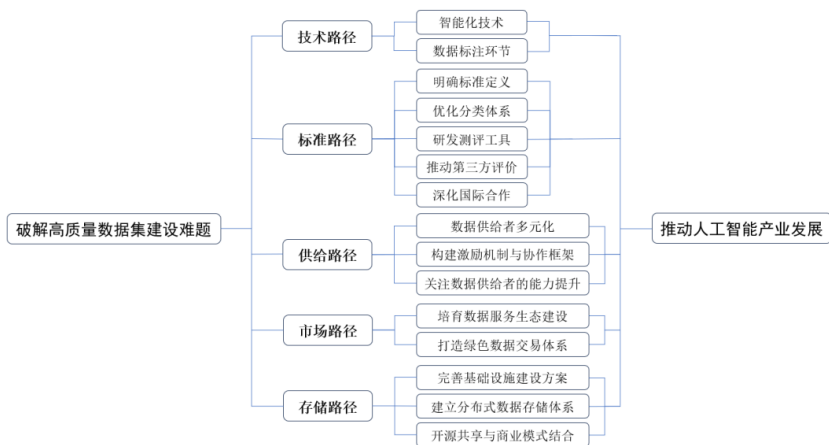


图 3 高质量数据集建设路径

智能化技术在数据质量提升中的作用尤为显著，通过一系列基于算法的清洗与预处理技术，数据的纯净性得以增强。例如，采用统计方法检测异常值，再结合哈希算法快速清除重复数据，不仅消除了数据冗余，也为后续处理铺平了道路。在预处理阶段，智能化技术进一步优化了多类型数据的处理流程。自动化分词工具能够精确分解文本数据，而图像数据则可以通过裁剪与归一化等技术实现标准化。这些技术措施并不仅仅是单一的改良，而是通过技术的协同作用，使得数据在一致性、适用性以及标准化水平上达到新的高度，为后续人工智能模型的训练提供坚实的基础。数据标注环节是技术路径的另一个重点领域。智能化标注技术依托深度学习算法，在高效完成大规模数据标注的同时，也展现了其对复杂任务的适应能力。预训练模型可直接应用于图像分类、文本实体识别等任务，而结合少量有标注数据的半监督学习技术，则进一步扩展了标注数据的覆盖范围。然而，标注的准确性不能完全依赖于自动化工具，智能化评估体系的引入弥补了这一不足，通过自动检测和校正标注结果，确保数据质量始终保持高水平。这种人机协作的方式，不仅提升了标注效率，还凸显了技术路径在数据治理中多层次应用的可能性。

2. 标准路径。高质量数据集的标准体系建设是数据生态中至关重要的一环，它为行业提供了

明确的方向和可操作的规范，同时奠定了数据在不同领域的可靠性和互操作性基础。通过科学严谨的标准化路径，不仅能统一行业认知，还能显著提升数据集的价值挖掘和产业转化效率。

在制定高质量数据集标准时，首先要明确多维度的核心指标。这些指标涵盖数据来源、类型、规模、敏感度以及更新频率等关键维度。仅依赖某一单一指标难以全面评估数据集的适用性和质量。因此，通过建立综合性的指标体系，明确高质量数据集的定义和具体质量要求，能够有效指导行业内的数据构建单位实现标准化操作。例如，对于自动驾驶场景的数据集，动态更新与高精度标注是关键；而在智慧医疗领域，数据隐私的保护和医学影像的准确性则需要得到优先关注。为确保标准的广泛接受与应用，加强标准体系的宣传和推广尤为必要。通过组织行业研讨会、开展标准宣贯活动以及发布评估解读报告，可以有效提高行业对标准体系的认知度和参与度。同时，通过积极参与 ISO、IEC、ITU 等国际标准化工作，中国在全球数据标准化领域的贡献率将显著提升。这不仅有助于增强国内数据服务的国际竞争力，也能为国际数据生态的规范化发展提供宝贵经验。在“标准路径”的探索中，创新与实践的结合至关重要。通过明确标准定义、优化分类体系、研发测评工具、推动第三方评价以及深化国际合作，可以形成一套科学、高效且具有前瞻性的标准体

系。这种标准体系的构建，不仅为高质量数据集的产业化提供坚实基础，也为全球数据标准化工作注入新的活力。

3. 供给路径。在构建高质量数据集的过程中，上游数据供给者的作用不可替代。供给路径的核心在于如何通过深度协作与机制设计，充分挖掘和利用数据供给者的资源潜力，使其成为数据生态体系中的关键驱动力。这一过程要求对数据供给者进行科学梳理与分类，构建激励机制与协作框架，形成稳定、高效的数据供给体系。

数据供给者的多元化特性决定了其贡献数据的能力与意愿需要针对性挖掘。行业龙头企业作为关键数据持有者，往往掌握着覆盖面广、精度高、商业价值显著的数据资源。通过鼓励这些企业明确自身在人工智能模型中的应用需求，并将部分非核心数据开放共享，可以快速推动数据供给。例如，自动驾驶领域的头部企业可共享多样化的道路环境数据，形成行业基准数据集，从而加速相关模型的开发和迭代。这种共享行为需要以清晰的协议和边界保障为前提，确保企业核心竞争力不受影响，同时明确数据使用场景与范围，避免潜在的商业风险。在激励供给者参与数据共享的过程中，合理的收益分配机制是关键。数据供给往往需要供给者承担一定的开放成本和风险，因此设计激励机制以保障供给者的利益显得尤为重要。例如，通过建立数据交易平台，供给者可以基于共享的数据规模与质量获得直接的经济回报。与此同时，供给者还可以通过数据生态的合作模式获得更长远的非经济收益，如提升行业地位、拓展技术合作机会等。这种多层次的激励机制可以显著提升数据供给者的参与意愿。在长远发展中，与供给者的持续合作还需要关注其能力提升。通过系统化的培训与示范项目，帮助供给者认识到数据共享在行业发展中的重要性 and 潜在收益。例如，举办数据供给者研讨会、共享最佳实践案例，向供给者展示开放数据如何带动行业生态的整体跃升。在此过程中，尤其需要针对数据较为稀缺或传统行业的供给者，进行深入的沟通与价值梳理，确保其能真正意识到参与数据共

享的长远价值。

4. 市场路径。高质量数据集的市场路径建设，是推动数据资源价值释放与流通的重要手段。通过完善交易规则、培育服务生态以及促进跨境数据流动，可以打造透明、高效且具有国际竞争力的数据市场体系，为人工智能和相关产业提供持续的数据支持。

数据服务生态的培育，是市场路径中的重要一环。作为连接数据供给方与需求方的关键桥梁，数据服务生态需要提供多样化的服务来满足不同场景的需求。例如，数据服务机构可以提供数据清洗、标注、脱敏等增值服务，以提升数据的质量与适用性。与此同时，研究机构与企业也可以通过开发行业专属的数据分析工具与平台，为特定领域的数据用户提供个性化服务。为了激励更多主体参与数据服务生态的建设，需要在市场机制中嵌入利益共享模式。例如，通过数据服务机构将交易平台中介费部分分成给数据供给方，不仅可以增强其合作意愿，还能带动整个数据服务市场的健康发展。

在长远规划中，数据市场的建设需要进一步注重可持续发展。通过打造绿色数据交易体系，确保数据交易的透明与合规，避免因滥用或不正当交易引发的不良影响。同时，制定市场激励措施，如对跨境数据交易提供税收优惠，对符合公共利益的数据流动给予政策支持，可以更有效地推动数据市场生态的稳健发展。

5. 存储路径。高质量数据集的存储路径是数据生态建设中的基础性环节，其目标在于通过系统化布局高效、低成本的存储与治理设施，支撑数据资源的长期保存与高效利用。这一路径不仅需要技术上的创新突破，还需在架构设计、治理标准和使用规范方面进行全面规划，以确保数据存储的安全性、可靠性和扩展性。

构建高效的数据存储体系，首先需要完善基础设施建设方案，打造一个全国性的数据集“算力一张网”。这不仅包括优化存储硬件的布局，还涉及算力管理的制度设计。在此过程中，合规有序的算力分配机制尤为关键，应在公益、科研

和商业应用之间实现资源的合理调度,确保各类需求都能得到满足。同时,通过统一的全国数据存储网络,可以实现分布式存储与集中治理的有机结合,为不同领域的数据资源共享提供高效的存储支持。分布式数据存储体系的建立是应对大规模数据存储需求的核心策略。通过“1+N”模式,构建一个核心节点和多个分支节点的存储架构,不仅能够提升存储系统的弹性与可靠性,还能显著降低大规模数据存储的成本。例如,核心节点可以负责国家级关键数据的集中存储与安全管理,而分支节点则聚焦行业级、区域性数据的存储和治理。这种分布式存储体系,还需要辅以明确的治理与评价标准,为数据存储设施的质量与运行效率提供系统性评估框架。创新打造高质量数据集治理工具链,是存储路径实现全面优化的关键举措。这一工具链涵盖从数据采集到清洗、增强、标注、融合与合成的全流程,并通过自主研发确保技术的可控性与适应性。例如,针对多模态数据的复杂处理需求,可以开发专用的处理模块以提高数据的一致性与扩展性。治理工具链的开发还需要注重标准化,编制统一的使用规范与标准,以推动工具链在行业内的普及与共享。为了实现存储路径的可持续性,开源共享与商业模式的结合不可或缺。通过开放部分基础工具链模块,降低行业内数据存储与治理的技术门槛,同时通过有偿提供高端服务模块,形成开源共享与商业合作的良性生态。例如,在公益与科研领域,可以免费提供基础工具的使用许可,而在高附加值的商业应用场景中,则可以通过收费机制支持工具链的持续迭代。

(二) 建设高质量数据集的对策建议

高质量数据集的建设与人工智能产业的发展是相辅相成的。作为训练高效大模型的核心要素,数据集的质量、规模与多样性对于大模型的性能具有极为重要的意义,二者深度融合将释放显著的协同倍增效应。当前,在数智化转型背景下,我国高质量数据集建设正从多维度协同推进,形成自上而下的系统化发展格局。这种全方位、多层次的推进模式,为高质量数据集建设注入了强

劲动力,为人工智能产业发展筑牢了数据根基。但目前,高质量数据集的建设在数据资源、人工智能、应用生态等层面仍存在多重瓶颈,制约产业进一步发展。为建设高质量数据集,进一步推动人工智能产业发展,本文从以下四个方面提出对策建议:

1. 构建体制机制。一是构建跨部门协同机制。AI大模型的高质量数据集建设涉及多个部门和行业,应建立跨部门的协调机制,形成协同推动的合力,共享行业数据资源,避免重复建设,提升资源利用效率。促进产学研合作。建立政府、企业、科研机构 and 高等院校的合作机制,推动人工智能基础研究、技术研发与数据集建设相结合,实现理论研究和实际应用的良性循环。上海人工智能实验室联合多家单位成立了大模型语料数据联盟,通过整合多方资源,加速数据要素建设,为人工智能的产学研合作提供了有力支撑。三是形成开放共享机制。推动公共数据集和行业数据集的共享与开放,鼓励企业和科研机构在确保数据隐私和安全前提下,逐步实现数据资源的共享和流通,避免数据孤岛。在这方面,成都市创新性构建的“管住一级、放活二级”数据资源开发利用模式,通过统一授权和集约化运营,开发多个一级数据产品,并通过“蓉数公园”等平台,汇聚各类数商,推动数据要素流通和应用,具有借鉴意义。

2. 建立标准体系。一是制定行业数据集标准:建立统一的数据集标准体系,包括数据格式、标注规范、数据质量控制等标准,确保数据集在行业中的通用性和互操作性。二是推动数据质量标准化。制定数据质量评估标准,明确数据标注的规范和质量控制机制,提高数据集的质量和准确性,采用自动化工具和人工智能算法提高数据清洗和标注的效率。三是统一隐私保护标准。制定隐私保护的标准,确保数据集的使用不会侵犯个人隐私并符合相关法规要求至关重要。隐私保护标准应当明确数据收集的透明度和边界,可以参考DeepSeek的隐私政策,其强调用户数据的收集必须遵循“最小必要原则”,并明确区分“必需信息”与“可选信息”,尤其在敏感信息处理方面,DeepSeek明确声明不主动收

集生物识别、金融账户等敏感数据。

3. 精进技术工具。一是加强数据处理技术创新：推动数据清洗、数据增强、数据标注等核心技术的创新，提升大规模数据处理和分析的效率，降低数据准备阶段的成本。发展自动化标注、数据质量检测等技术，提升数据集的构建效率和质量。二是发展去中心化技术。采用区块链、同态加密等技术，解决跨地域、跨行业的数据隐私保护问题，为大规模跨行业合作和数据共享提供技术保障，促进数据集的安全、高效流动。三是推动数据集多样性增强技术。发展多模态数据集建设技术，支持图像、文本、语音、视频等多种数据类型的结合和应用，提升 AI 大模型的多任务学习能力，增强模型的泛化能力和鲁棒性。

4. 加强人才培养。一是加强 AI 数据集相关专业教育。高校应开设 AI、数据科学、大数据分析等相关专业课程，在课程中增加对前沿技术的讲解和应用实践，使学生能够掌握最新的技术和方法，从而培养更多专业人才，提升其在数据采集、标注、清洗、分析等方面的能力。二是促进跨学科人才培养。培养跨学科复合型人才，有助于提升数据集的专业性和应用价值。高校可以联合不同学科的院系，共同开设跨学科课程，打破学科壁垒，促进学生在不同领域的知识融合。三是建立行业专家队伍。通过政府、企业和科研机构的合作，培养和吸引行业专家，尤其是在关键领域，通过协同攻关制定数据标注标准，建立动态评价机制，提高数据集的专业性和质量。

参考文献：

- [1] 中共中央关于进一步全面深化改革 推进中国式现代化的决定 [M]. 北京：人民出版社，2024.
- [2] 习近平向 2024 世界智能产业博览会致贺信 [N]. 人民日报，2024-06-21.
- [3] 坚持自立自强 突出应用导向 推动人工智能健康有序发展 [N]. 人民日报，2025-04-27.
- [4] 林军. “数字化”、“自动化”、“信息化”与“智能化”的异同及联系 [J]. 电气时代，2008，(1).
- [5] 刘少杰. 从实践出发认识网络化、数字化

和智能化 [J]. 社会科学研究，2022，(2).

[6][12] 张子涵，陈子韬. 中国社会治理数字化转型顶层设计的核心特征 [J]. 河南大学学报（社会科学版），2025，65（1）.

[7] 于跃. 智慧政府的生成与演进逻辑 [J]. 电子政务，2019，（7）.

[8][10] 杨成延. “数智化”助力精益供应链构建与运营 [J]. 物流技术与应用，2019，24（11）.

[9] 李廉水，石喜爱，刘军. 中国制造业 40 年：智能化进程与展望 [J]. 中国软科学，2019，（1）.

[11] 罗斌元，陈艳霞. 数智化如何赋能经济高质量发展——兼论营商环境的调节作用 [J]. 科技进步与对策，2022，39（5）.

[13] 金艳艳. 贸易企业财务数字化与智能化转型研究 [J]. 上海企业，2025，（2）.

[14] 珠海市人民政府服务和数据管理局. 广东四步走推动全省高质量数据集建设 [EB/OL]. https://www.zhuhai.gov.cn/zwfwglj/gkmlpt/content/3/3782/post_3782850.html#1678, 2025-04-02.

作者：王 鹏，北京市社会科学院副研究员、博士
丁一桐，北京市社会科学院研究助理

责任编辑：周修琦