数据工厂的构成、建设模式和运营机制研究*

涂　群[1]，耿贵宁[2]，张茜茜[3]
（1. 北京化工大学　经济管理学院，北京　100029；2. 三六零数字安全科技集团有限公司，北京　100015；
3. 北京物资学院　计算机与人工智能学院，北京　101126）

摘　要：高质量数据集是人工智能大模型训练的核心“燃料”。当前，高质量数据集构建主要由人工智能企业自行完成，呈现零散化、作坊式、非标化的特点，难以满足人工智能大模型快速发展的需求。借鉴水厂、电厂等资源型基础设施的发展规律，结合国内外高质量数据集设施化生产的典型实践，提出“数据工厂”概念，将其定义为面向人工智能大模型应用、设施化规模化构建高质量数据集的生产设施。系统阐述了数据工厂由“储备车间”“生产车间”“中试车间”构成的三级架构体系，分析了数据标注企业升级、数据存储基地转型、人工智能企业延伸和技术企业创新设立四种建设模式，提出了保障模式、定制模式、电商模式和结对子模式四种运营机制，为推动高质量数据集设施化、规模化供给提供理论支撑和实践参考。

关键词：数据工厂；高质量数据集；数据基础设施；数据要素
中图分类号：F49　　**文献标志码：**A　　**DOI：**10.19358/j.issn.2097-1788.2026.04.002

中文引用格式：涂群，耿贵宁，张茜茜. 数据工厂的构成、建设模式和运营机制研究[J]. 网络安全与数据治理，2026，45(4)：9-16.
英文引用格式：Tu Qun，Geng Guining，Zhang Qianqian. Research on the composition，construction models and operation mechanisms of data factories[J]. Cyber Security and Data Governance，2026，45(4)：9-16.

网络安全与数据治理

Research on the composition, construction models and operation mechanisms of data factories

Tu Qun[1], Geng Guining[2], Zhang Qianqian[3]
(1. School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China;
2. 360 Digital Security Technology Group Co., Ltd., Beijing 100015, China;
3. School of Computer Science and Artificial Intelligence, Beijing Wuzi University, Beijing 101126, China)

Abstract: High-quality datasets are the core fuel for training large AI models. Currently, the construction of high-quality datasets is mainly carried out by AI enterprises themselves, which presents the characteristics of fragmentation, workshop-style operation and non-standardization, making it difficult to meet the rapid development needs of large AI models. Drawing on the development patterns of resource-based infrastructure such as water and power plants, and combining domestic and international best practices in facility-based production, this paper proposes the concept of "data factory", defining it as a production facility specifically designed for the application of large AI models and for the facility-based, large-scale construction of high-quality datasets. The paper systematically expounds the three-level architecture system of the data factory, which consists of storage workshop, production workshop, and pilot workshop. Four construction models and four operation mechanisms are proposed, providing theoretical support and practical references for promoting the facility-based and large-scale supply of high-quality datasets.
Key words: data factory; high-quality dataset; data infrastructure; data element

*基金项目：北京市社会科学基金（23GLC058）

0 引言

算力、算法和数据是人工智能的三个关键要素，长期以来，高耗算力、模型闭源和数据短缺一直制约着人工智能大模型的应用普及。以 DeepSeek 为代表的人工智能企业，实现了 MOE 等关键技术重大突破并采取了模型开源策略，实现了“算力平权”和“算法平权”[1]，促进人工智能大模型向通信、互联网、汽车、能源、金融、医疗等各行各业加速渗透，人工智能大模型广泛应用的时代已经到来。与此同时，一条面向人工智能的数据产业链正在快速形成：上游是公域数据资源和基础大模型，中上游是非结构化高质量数据集和行业高质量数据集，中下游是智能体和垂域大模型两类应用工具，下游则是千行百业的智能化应用。垂域大模型通过在特定领域的深度应用，已在药品研发、金融风控、医疗诊断等领域展现出专业级能力[2]；智能体通过“感知—决策—执行”的闭环，在具身智能、工业制造、自动驾驶等场景实现实时交互与自主作业[3]。在这条产业链中，高质量数据集处于承上启下的关键位置：向上承接海量原始数据资源，向下支撑大模型训练和智能体运行。工具越成熟，应用越广泛，对高质量数据集的需求就越迫切。

然而，从原始数据到高质量数据集，还面临着三重困境。一是数据资源“供不出”。据 IDC 统计，全球非结构化数据占数据总量的 80% 以上[4]，这类数据格式各异、标准不一，加之大量涉及隐私或商业秘密的私域数据开放意愿不足，导致绝大多数数据难以直接流通。二是行业数据“存不好”。高价值的行业数据大多分散保存在各部门、各企业，据统计每年约四成数据从未被使用过[5]，大量潜在价值数据在沉睡中逐渐流失。三是高质量数据集“产不好”。当前高质量数据集的构建主要由人工智能企业自行完成[6]，生产方式原始、效率低下，投入产出失衡、标准规范缺失[7]。

从全球发展趋势来看，高质量数据集的设施化、规模化生产已成为共识。美国 Scale AI 自 2021 年获得美国军方 2.5 亿美元合同后，从数据标注外包公司发展成为专业的数据工厂，构建了覆盖通用、生成式 AI、公共部门、汽车等不同领域的专业化数据引擎矩阵[8]。美国 2025 年提出的“星际之门”项目整体投资 5 000亿美元，将高质量数据明确定位为“国家战略资产”[9]。欧盟 2025 年推出数据联盟战略，建设数据实验室作为人工智能工厂的有机组成[10]。在国内，帕西尼 2025 年建成全球规模最大的具身智能数据采集基地，库帕思构建了包含 403 个功能模块的语料工具链平台。借鉴水厂、电厂等资源型基础设施的发展规律，本文提出“数据工厂”概念，系统研究其构成体系、建设模式和运营机制。

1 数据工厂及其构成

1.1 数据工厂的概念

纵观人类社会发展史，水、电、燃料等基础资源的规模化利用都离不开资源型基础设施的支撑。这类基础设施通常包含三个协同环节：生产环节建设规模化加工设施，流通环节建设传输网络，利用环节承接多样化应用场景。数据基础设施作为数智化阶段新的资源型基础设施，同样需要三个环节的系统布局。在流通环节，2024 年 12 月《国家数据基础设施建设指引》已明确建设方向；在利用环节，场景应用实施方案持续推进。但在生产环节，面向人工智能的高质量数据集规模化生产设施仍是薄弱项，亟需加快布局。

基于此，本文将“数据工厂”定义为：集数据“采存算管用”功能于一体，面向人工智能大模型应用，设施化、规模化、标准化构建高质量数据集的生产设施。数据工厂是国家数据基础设施在生产环节的核心组成部分，具有三重功能定位：一是“生产设施”，提供数据存储、治理、加工、质检等全流程服务能力；二是“成本中心”，通过集约化运营显著降低数据归集、加工和管理成本；三是“供给枢纽”，推动行业数据和区域数据有序汇聚、高效转化。

1.2 储备车间

1.2.1 储备车间的定位

图 1 展示了储备车间的体系构成。储备车间定位于数据工厂内数据原料的“供应基地”，提供规模化收储、标准化预处理、体系化管理的数据原料；通过系统性地收集原始数据并转化为可量化、可管理、可追溯的数据资源，为后续加工环节奠定质量基石。

1.2.2 储备车间的功能

储备车间是数据工厂的原料仓库，是高质量数据集生产的前提和基础。储备车间的功能主要包括数据需求与规划、数据采集与汇聚、数据存储与管理等三部分。其中，数据需求与规划是根据特定人工智能应用场景，明确数据集的数据规范、质量、内容等，包括设计数据架构、制定质量计划、预估工作量。数据采集与汇聚是根据人工智能大模型训练和应用场景对数据的需求，确定数据来源、数据采集方式和数据汇聚方式，并将来自不同渠道的多源异构数据进行统一汇聚，或创新数据编织、数据虚拟化等新技术对多源

网络出全与数据治理

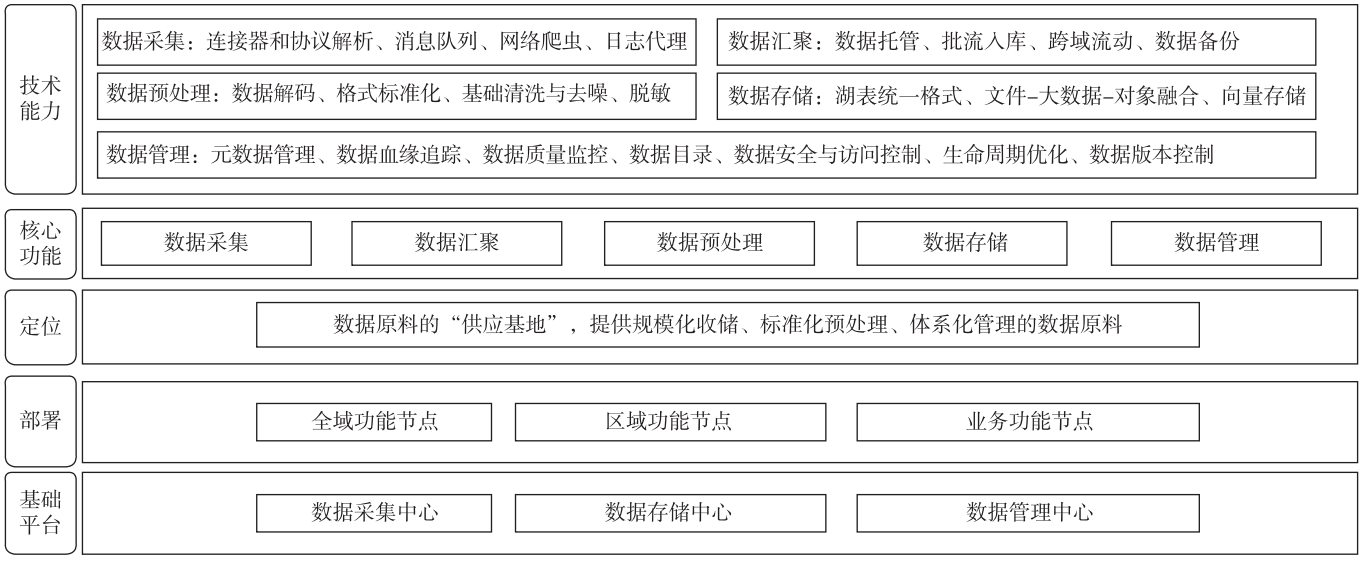

图 1 “储备车间”体系构成图

异构数据进行分布式整合。数据存储与管理是对集中汇聚或分布整合的数据进行预处理，包括数据转换、数据验证、数据清洗、格式统一等，转化为适用于生产车间加工处理的基础数据，构建统一的数据存储体系，并建立数据全生命周期管理机制。

1.2.3 储备车间建设内容

储备车间依托数据采集、汇聚、整合、存储、管理等多种技术，构建起数据采集中心、数据存储中心和数据管理中心三大功能中心，形成供给充沛的数据资源供应基地。

数据采集中心负责多源异构数据的采集、汇聚、整合和初步处理，建立多渠道数据获取体系。具体工作包括：（1）开展数据采集。运用实时采集、批量采集、断点续传等技术，配置网络爬虫工具、API 接口工具、传感器数据采集工具、批量数据导入工具、数据流动系统等数据集采集工具，支持公开数据集接入、网络数据抓取、自定义数据采集等多种方式进行数据采集。（2）开展元数据发现。通过文本列表、文件格式、标签信息、湖格式化等技术，实现数据的自动识别和分类。（3）开展采集质量监控。构建数据采集质量监控系统，配置敏感词检测工具、数据投毒检测工具、隐私智能检测工具等数据内容安全工具，实时监测数据采集的完整性和有效性。

数据存储中心负责将经过采集、汇聚/整合和预处理后的原始数据、中间数据和结果数据等各类数据进行安全、可靠、高效的持久化保存；构建一个集中/分布式的、可扩展的多模态数据存储池，以容纳海量多源异构数据，并根据数据的访问频率和价值，实施冷、热、温分层的存储策略，以平衡存储成本与访问效率；通过多副本、纠删码等技术保障数据的可靠性与可用性，并通过严格的权限控制和加密机制确保数据安全。

数据管理中心通过建立统一的数据标准、元数据体系和数据质量稽核规则，对数据进行规范定义、血缘追踪和质量监控，保障数据的一致性、准确性与完整性；实施分级分类和访问权限控制等数据安全策略，以管控数据风险；制定数据资产目录，实现数据的可发现、可理解与可便捷获取，提升数据共享与复用效率。

储备车间基于模块化设计，构建起涵盖数据采集、数据汇聚、数据预处理、数据存储和数据管理全流程的五大功能模块：数据采集模块，负责从各类数据源获取原始数据，支持实时采集、批量采集、断点续传等方式；数据汇聚/整合模块，负责将多源异构数据进行集中化/分布式汇聚或整合，形成规范的待加工数据流；数据预处理模块，负责对汇聚或整合后的原始数据进行清洗、转换与标准化粗加工；数据存储模块，负责数据的持久化、分级存储，并提供高效、安全的读写访问服务；数据管理模块，负责开展数据资产编目、质量管控与访问控制，实现数据高效共享与安全使用。

1.3 生产车间

1.3.1 生产车间的定位

生产车间体系构成如图 2 所示。生产车间定位于高质量数据集生产的“流水线”，承担面向人工智能应用的高质量数据集设施化、规模化、标准化构建，实现高质量数据集构建从分散化向设施化、从作坊式到规模化、从零散式向标准化的根本转变。

网络安全与数据治理

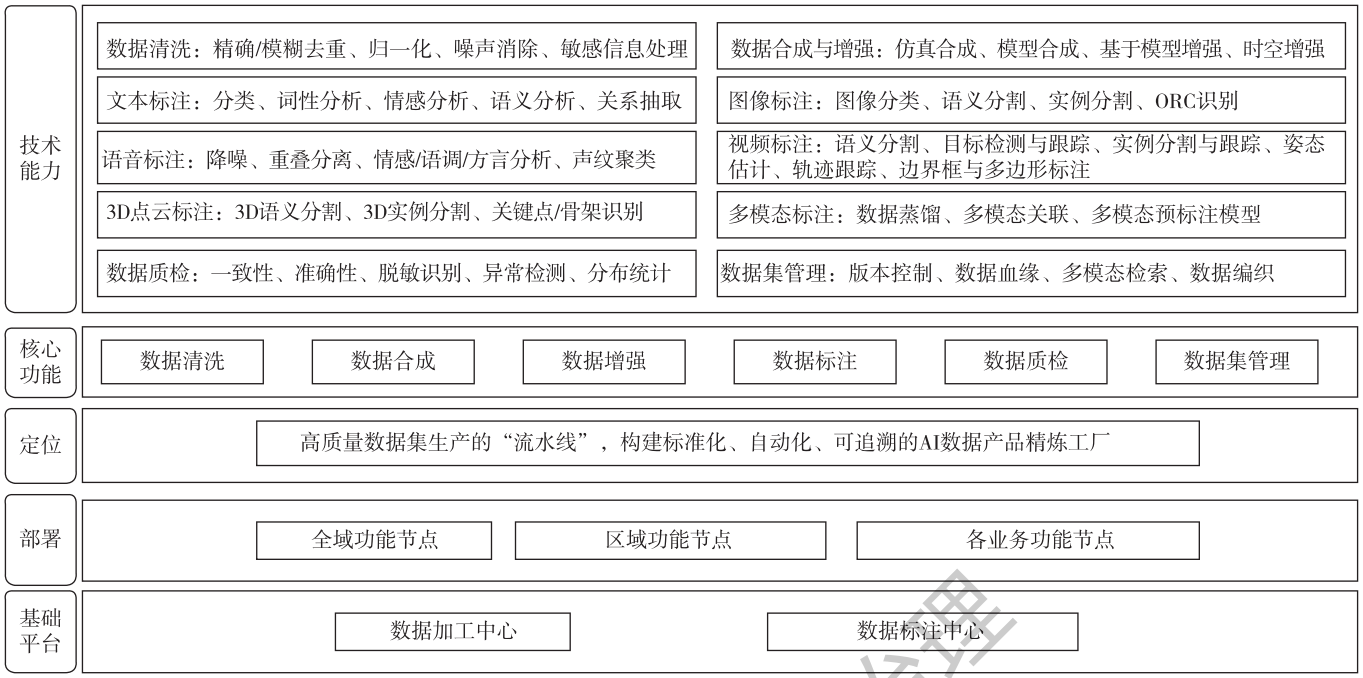

图 2　“生产车间”体系构成图

1.3.2　生产车间的功能

生产车间的主要功能是将储备车间的原始数据或初加工数据，加工处理成可供人工智能大模型训练、微调、推理等应用的高质量数据集，主要包括数据加工与合成、数据标注与质检两个阶段。其中，数据加工与合成通过数据清洗识别和处理重复、噪声、异常等问题数据，提升数据质量。通过数据合成生成符合特定需求和分布的新数据；通过数据增强在不改变数据核心语义的前提下，对现有数据进行可控变换与扩充，生成多样化的衍生数据，扩大数据规模、增加数据多样性。数据标注与质检指根据训练、验证和测试数据的需要，明确数据标注规范，对经过清洗、合成和增强的数据添加标签、语义注释和元数据，并对标注过程进行监测和质量管理。通过数据质检对各环节产出的数据进行系统性质量验证与评估，确保数据集的准确性、一致性、完整性；建立数据集全生命周期管理体系，确保数据集的可发现、可理解、可信任、可复用。

1.3.3　生产车间建设内容

生产车间依托数据清洗、转换、增强、标注、检测等各种技术，构建起数据加工中心和数据标注中心两大生产中心，形成规模化、自动化、智能化的高质量数据集生产线。

数据加工中心负责储备车间数据集的清洗、转换、增强等预处理工作，建立分层数据处理流水线，包括：（1）开展数据清洗工作。开发自动化数据清洗系统，配置错误 SQL 清洗工具、智能云隐私工具、智能去重工具等数据清洗工具，运用去乱码、去隐私、去重、去模糊、敏感内容过滤、声道处理、音量标准化等技术，开展识别和处理重复数据、噪声数据、异常数据等工作。（2）开展数据增强工作。构建数据增强平台，部署数据增强工具、自有数据增广工具等，通过分类、摘要、调色、调清晰度、视频去噪、冗余剔除、噪声抑制等技术手段，对数据进行提质与增强。(3)开展数据变换工作。运用文档解析、图像内容识别、音频转文字、视频内容识别等技术，提取数据的关键特征；运用数据集字符统计工具、数据样本精细化分析工具等数据特征分析工具，以及分布式计算框架，支持大规模数据并发处理；运用数据合成技术开展数据合成，扩充数据集规模和多样性。（4）开展质量检测工作。建立数据质量检测机制，确保加工后数据符合预定标准。

数据标注中心负责数据的标注、审核和质量控制，建立人机协同标注体系。研究表明，采用主动学习与人工标注相结合的人机协同标注模式，可在保证标注质量的前提下显著降低人工标注成本、提升整体标注效率[11]。数据标注中心工作内容包括：（1）开展数据标注工作。运用模型标注、智能标注、人工辅助标注等技术，结合自动标注工具和人工标注团队，提高标注效率和质量。（2）开展质量审核工作。构建多级质

网络安全与数据治理

量审核机制，包括自动质检、交叉验证、专家抽检等，确保标注的准确性和一致性。(3)开展标注管理工作。建立标注人员培训和管理体系，制定标注规范和操作流程；开发智能标注平台，集成预训练模型、主动学习等技术，实现半自动化标注。(4) 构建质量控制体系。通过规则校验、一致性分析、异常检测等技术自动识别标注问题；配置 AI 提取工具、AI 生成工具、AI 诊断工具、AI 决策工具、AI 分析工具等数据标注工具，支持文本分类、实体识别、图像分割、视频标注等多种标注任务；建立标注质量评估机制，确保标注的专业性和一致性。

1.4 中试车间

1.4.1 中试车间的定位

中试车间体系构成如图 3 所示。中试车间定位于数据工厂内高质量数据集交付投产前的“靶场”，衔接高质量数据集生产与人工智能大模型训推环节，负责评估数据集质量及模型适配性、排查问题、优化数据，确保数据集满足模型训练要求。

1.4.2 中试车间的功能

中试车间的主要功能是在高质量数据集生产完成之后，投入人工智能模型训练之前，对高质量数据集质量进行效果适配性评估，以确定数据集是否满足模型训练、微调、推理要求，主要包括模型评估、数据集维护和更新两项功能。

一是模型评估。通过模型训练评估，使用数据集训练基准模型，评估训练性能、收敛质量及资源效率，验证数据有效性；通过模型微调评估，对预训练模型进行场景化微调，评估微调效果、泛化性能及稳定性，验证数据集的领域适应能力；通过模型推理评估，测试模型在真实业务场景中的准确性、鲁棒性及业务适配性，反向识别数据缺陷。

二是数据集维护和更新。基于模型评估反馈驱动数据集迭代优化，将对模型表现产生不利影响的数据质量问题反馈给上游环节；通过问题溯源精准定位数据缺陷根源，重复数据采集、预处理、标注等环节以提升数据质量；建立版本化管控机制，记录每次迭代的内容变更与质量提升，实现数据集的可追溯、可比较与可回溯。

1.4.3 中试车间建设内容

中试车间构建起两大验证中心和四大功能模块，对高质量数据集进行训练前效果适配性评估和优化，确保形成合规、质量可靠、可用性达标的数据集。

两大验证中心包括模型评估中心和数据集维护更新中心。模型评估中心负责通过真实人工智能模型训练与推理验证，评估数据集在实际应用场景中的有效性。具体包括：使用标准测试集对模型的准确性、召回率等核心性能指标进行全面测评；结合实际业务逻辑与场景，验证模型输出结果的合理性、可解释性及业务价值达成度；验证高并发、大数据量等生产环境压力下的模型表现；检测模型是否存在潜在偏见或歧

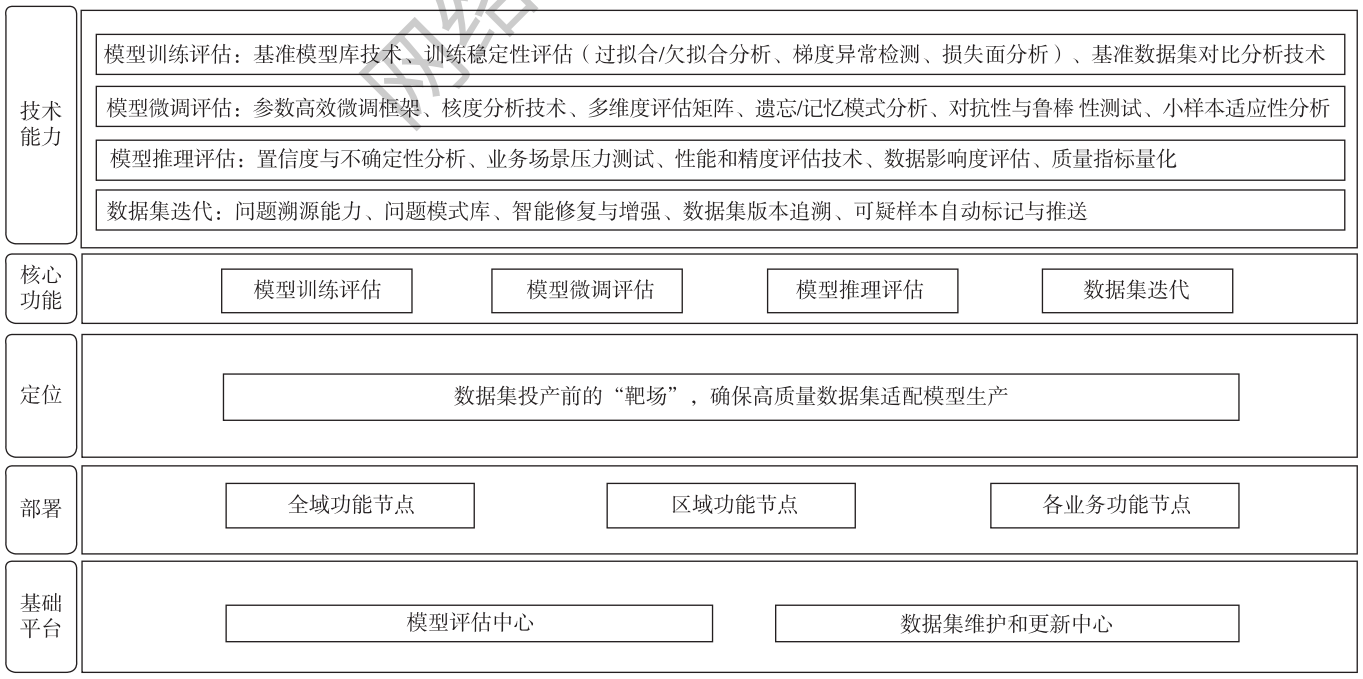

图 3 “中试车间”体系构成图

网络安全与数据治理

视，并验证其决策逻辑是否符合伦理规范及监管要求；通过特征重要性分析和错误样本诊断，反向识别数据缺陷，为数据集维护更新中心提供精准改进建议。数据集维护更新中心则根据人工智能大模型发展和新数据生成等情况，对高质量数据集进行定期维护更新和迭代优化，确保高质量数据集的长期价值。具体包括：开展标准化管理，记录数据内容处理逻辑变更，记录版本历史，支持数据回溯；开展数据补全，根据生产车间所交付数据集的变化，自动或手动触发数据集的增量更新与补全，保障数据的时效性；开展数据修复，接收并处理数据集质量反馈或异常报告，定位问题根源，协调并执行数据修复流程；开展数据优化，制定数据集归档、降级或下线策略，对不再活跃或价值较低的数据集进行优化管理，合理控制成本。

中试车间基于模块化设计，构建了覆盖模型训练评估、模型微调评估、模型推理评估和数据集迭代的四大功能模块。模型训练评估模块负责使用验证数据集训练基准模型，评估训练性能、收敛质量及资源效率，验证数据有效性。模型微调评估模块负责对预训练模型进行场景化微调，评估微调效果、泛化性能及稳定性，验证数据集领域适应能力。模型推理评估模块负责通过模型推理表现评估数据集质量，验证数据在真实场景中的准确性、鲁棒性及业务适配性。数据集迭代模块负责基于全链路测试反馈驱动数据集优化迭代，管理版本演进，促进高质量数据集持续迭代优化。

2 数据工厂的建设模式

根据建设主体的原有基础和条件，数据工厂可以通过数据标注企业升级、数据存储基地转型、人工智能企业延伸和技术企业创新设立四种模式建设。

2.1 数据标注企业升级

当前我国已形成一批数据标注企业，并正在建设四川成都、辽宁沈阳、安徽合肥、湖南长沙、海南海口、河北保定、山西大同七个国家数据标注产业基地。但是，这些企业普遍存在自动化程度低、设施化程度差、低端外包业务占比大等突出问题，亟需从零散式、手工化、低端化的劳动密集型数据标注企业转型升级为规模化、智能化、高端化的数据工厂。

国家层面和有条件的省市应设立半集中式数据工厂建设和改造项目，支持传统数据标注龙头企业积极加大数据资源储备、数据生产加工、数据中试验证等方面科技投入和技术创新，尽快实现从数据标注企业到数据工厂的跃迁。推动国家数据标注产业基地的七个城市建设半集中式数据工厂，一方面应加快数据加工生产环节的技术改造和公共服务平台升级，另一方面要向上下游产业链延伸拓展，在上游加大数据资源汇集储备力度，下游拓展人工智能大模型训练和各行各业应用。支持新设高端化、智能化、设施化的半集中式数据工厂，逐步淘汰低端化、手工化、定制化的作坊式数据标注业态。

2.2 数据存储基地转型

当前以“八大枢纽十大集群”为代表的存储基地、数据中心和算力中心，普遍存在“有存储无数据、有数据无加工”的现象，存储设备、算力设备和数据资源闲置的情况不同程度存在，导致业务层级低、经营困难。在人工智能大模型对高质量数据集需求迫切的形势下，应充分发挥这些存储基地已部署大量存储设备、积累海量高价值数据资源、拥有超强算力配置等显著优势，推动转型升级为集中式数据工厂新业态。

在建设路径上，可采用存储（算力）厂商与数据源机构合作共建方式。即存储厂商和算力厂商提供数据工厂加工处理技术，数据源机构提供数据资源，合作共建集中式数据工厂，将长期存放在存储基地或数据中心的静态原始数据，加工成能为人工智能大模型训练应用的高质量数据集。

2.3 人工智能企业延伸

当前我国人工智能企业已达 6 000 多家，涌现出一批以 DeepSeek、Kimi 等为代表的全球领先大模型企业。这些企业拥有领先的数据采集汇聚、加工处理、清洗标注等方面的技术和完善的数据生产产业链，但其数据生产加工仍是“作坊式”生产，仅为自有大模型训练服务。

应采取项目支持、资金补助等政策措施，在全社会培育出付费使用数据的意识和氛围；支持人工智能头部企业建设独立运营的数据工厂，形成第二业务增长曲线；鼓励与地方数据标注基地合作建设区域数据工厂，与行业“链主”单位合作打造行业特色数据工厂。

2.4 技术企业创新设立

数据工厂的重要任务之一是将因涉隐涉密而无法流通的私域数据，采用“原始数据不出域”的技术手段加工成可供人工智能大模型训练的高质量数据集。从全球实践看，美国 Palantir 公司创新数据虚拟化和连接器框架等数据编织技术，将分散在不同区域的数据连接在一起，在数据原地进行处理，形成“逻辑统一，物理解耦”的分布式数据加工生产业态。我国迫切需

网络安全与数据治理

要出现一批技术创新型公司，通过技术突破构建分布式数据工厂新业态。

应在国家科技计划等项目中将数据编织、数据虚拟化等前沿技术列为重点攻关方向，鼓励龙头企业将数据编织技术作为重点开发方向，在国防、金融、医疗、交通等重点行业开展先行试点示范。

3 数据工厂的运营机制

数据工厂应根据不同类型和不同建设模式，采用不同的运营机制。从高质量数据集的需求类型来看，主要包括国家战略需求、公益科研需求、商业市场需求和产业协作需求四类，可相应采用保障模式、定制模式、电商模式、结对子模式四种运营模式。特别是要根据不同类型数据的市场化程度差异，引导数据工厂建设运营主体面向中小企业分类探索随取随用、按需供给的数据供应方式，降低中小企业数据获取门槛。

3.1 保障模式

针对国家重大战略需求，采用数据集供给“保障模式”运营。国家数据资源战略储备库或各行业各部门控制的重要数据资源应以市场方式无条件向数据工厂提供，数据工厂生产的高质量数据集应向国家人工智能训练场和人工智能行业应用基地等有条件无偿开放。该模式主要由国家级或区域级数据工厂承担，通过建立高质量数据集战略储备目录和优先供给清单，确保关键领域数据需求得到及时响应。数据存储基地转型的数据工厂宜优先采用该模式，国家应通过中央预算内资金等财政资金予以重点支持。

3.2 定制模式

针对社会公益和技术创新需求，采用数据集供给“定制模式”运营。该模式主要服务于高校、科研院所、公益机构等非营利性主体，由数据工厂根据具体科研课题或公益项目需求提供定制化数据集生产服务。运营成本可通过科研项目经费、公益基金、政府补贴等多元渠道分担，数据集使用须遵守约定的用途限制和成果共享要求。技术企业创新设立的数据工厂宜优先采用该模式。

3.3 电商模式

针对数据成熟度较高、市场化基础较好的重点行业领域需求，采用数据集供给“电商模式”运营。该模式依托数据交易平台或数据工厂自建交易系统，提供标准化数据集产品的在线展示、检索、试用、购买、交付等全流程服务。建立数据集分级定价机制，根据数据规模、质量等级、时效性、独占性等因素确定价格，同时提供质量承诺、售后服务和争议处理机制。数据标注企业升级和人工智能企业延伸形成的数据工厂宜优先采用该模式。

3.4 结对子模式

针对数据资源禀赋较好但需要长期开发的重点领域需求，采用数据集供给“结对子模式”运营。该模式通过签订长期合作协议，明确行业龙头企业、大模型企业和数字化解决方案提供商在数据供给、数据加工、模型训练、应用开发等环节的权责分工和收益分配，建立“数据贡献—模型赋能—价值反哺”的可持续合作机制，推动数据集在实际应用中持续迭代优化。该模式适用于智慧医疗、智能驾驶、智能工厂等发展前景较好但当前技术和应用尚不成熟的行业领域。

4 结论

随着人工智能应用快速普及，高质量数据集的规模化供给已成为产业发展的核心瓶颈。本文借鉴资源型基础设施发展规律和国内外典型实践，提出了“数据工厂”概念，系统阐述了“储备车间—生产车间—中试车间”三级架构体系，分析了数据标注企业升级、数据存储基地转型、人工智能企业延伸和技术企业创新设立四种差异化建设模式，设计了保障、定制、电商和结对子四种运营机制。数据工厂的提出旨在填补国家数据基础设施在生产环节的空白，推动高质量数据集供给从零散化向设施化、从作坊式向规模化、从非标化向标准化转变。

展望未来，数据工厂的建设和运营需要政策、标准、技术、平台四方面的系统保障。建议加快出台促进数据工厂创新发展的政策文件，启动数据资源汇聚、数据集生产、质量评估、流通应用等全链条标准研制，突破非结构化数据采集、海量数据存储、智能化数据加工、数据集质量评估等关键技术瓶颈，建设数据资源汇聚、数据集生产、数据流通、人工智能训练等支撑平台，推动数据工厂成为国家数据基础设施的重要组成部分，为人工智能产业发展提供源源不断的“数据燃料”。

网络安全与数据治理

参考文献

[1] 谢新水. 智能跃迁，开源创新与主权 AI：DeepSeek 现象推动人工智能开源创新生态体系建设[J]. 电子政务，2025(3)：40－48.

[2] 苏美文，杨文爽，李博文，等. 推动人工智能与实体经济深度融合加快发展新质生产力[J]. 工业技术经济，2025，44(4)：32－59.

[3] 袁依格，何卓丰，李威，等. 具身智能驱动的智能制造应用发展研究[J]. 中国工程科学，2025，27(3)：67－82.

[4] REINSEL D, GANTZ J, RYDNING J. The digitization of the world from edge to core [Z]. Framingham: International Data Corporation, 2018.
[5] 张腾, 王珂飞. 推动数据产业高质量发展的路径研究[J]. 数字经济, 2025 (4): 7 - 10.
[6] 赵志君, 庄馨予. 中国人工智能高质量发展: 现状、问题与方略[J]. 改革, 2023 (9): 11 - 20.
[7] 姜春宇, 白玉真, 刘渊, 等. 构建企业级人工智能高质量数据集: 方法与路径[J]. 大数据, 2025, 11 (6): 47 - 56.
[8] Scale AI. Scale AI selected by U. S. Department of Defense to accelerate government's AI capabilities [EB/OL]. (2022 - 01 - 31) [2025 - 10 - 01]. https: //www. businesswire. com/news/home/20220131005304/en/.
[9] NIX J, O'BRIEN C. Trump signs AI orders, vows US will win race over new technology [EB/OL]. (2025 - 07 - 23) [2025 - 10 - 01]. https: //www. bloomberg. com/news/articles/2025 - 07 - 23/white-house-unveils-sweeping-ai-action-plan-to-boost-development.
[10] A&O Shearman. The European Commission publishes its AI continent action plan [EB/OL]. (2025 - 04 - 09) [2025 - 10 - 01]. https: //www. aoshearman. com/en/insights/ao-shearman-on-data/the-european-commission-publishes-its-ai-continent-action-plan.
[11] 徐拥军, 成徐慧. 建设面向人工智能的高质量档案数据集[J]. 中国档案, 2026 (1): 64 - 66.

(收稿日期: 2026 - 03 - 11)

作者简介:

涂群 (1989 -), 男, 博士, 副教授, 主要研究方向: 数据基础设施、数据安全。

耿贵宁 (1981 -), 男, 博士, 高级工程师, 主要研究方向: 网络安全、数据安全。

张茜茜 (1990 -), 女, 博士, 副教授, 主要研究方向: 数据要素、数据基础设施。

网络安全与数据治理

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.com