CAICT 中国信通院 中国电信 CHINA TELECOM

集智 专题报告

# 数据标注产业发展研究报告

## (2025 年)

中国信息通信研究院人工智能研究所

中电信人工智能科技（北京）有限公司

2025年8月

# 版权声明

# 前　言

习近平总书记指出，数据是新的生产要素，是基础性资源和战略性资源，也是重要生产力。党的十九届四中全会首次提出将数据作为生产要素。新一代高水平数据标注在推动数据资源汇聚、提升数据质量、激活数据要素价值方面发挥着日益重要的作用，是支撑人工智能技术演进和应用落地的重要抓手。2024 年 12 月，国家数据局印发《关于促进数据标注产业高质量发展的实施意见》（以下简称《实施意见》），对数据标注产业高质量发展具有重要的意义。我们要紧紧抓住数据标注这个小切口，以服务国家战略大视野，着力推动产业高质量发展。

数据标注在人工智能产业结构中占据举足轻重的地位，是连接数据资源、算法模型与实际应用场景的关键桥梁，是人工智能高质量数据集的核心生产力。狭义的数据标注产业是指将原始数据标记人类知识转换成机器可识别信息的过程。广义的数据标注产业通常指对数据进行筛选、清洗、分类、注释、标记和质量检验等加工处理的过程。当前，以数据标注为代表的人工智能基础数据服务，连结上游数据来源方和下游人工智能算法研发方，其产业发展和服务水平直接影响人工智能应用效果和场景落地。

本研究报告首先回顾了数据标注产业发展的总体概况，全面总结了数据标注产业发展的六大核心要素，提出了当前数据标注产业发展面临的问题与挑战，分析了未来数据标注产业发展总体趋势，

提出数据标注产业下一步发展的建议，可为政策制定者、行业从业者及企业投资者等提供全面的行业洞察、策略建议与决策依据。面向未来，数据标注产业发展仍存在诸多问题与挑战，还需要产学研各界紧密合作，共同推进数据标注产业技术创新与产业发展，为行业高质量数据集的构建和大模型训练提供有力支撑。

本报告由国家数据局数字科技和基础设施建设司指导，中国信息通信研究院联合中国电信集团、沈阳市数据局等多家单位联合编制，撰写过程中得到了中国人工智能产业发展联盟数据委员会、数据标注专委会、人工智能关键技术和应用评测工业和信息化部重点实验室的大力支持。报告先后征求并采纳清华大学、北京理工大学、航天二院、赛迪网安所等多位专家意见，以及国家数据局综合司、政策司、资源司、数经司、国合专班意见，形成相关研究成果。

# 目 录

# 图 目 录



# 表 目 录

# 一、数据标注产业总体概况

数据标注作为数据治理产业中的重要环节，其核心任务是对数据进行精准的分类、标记和描述，以确保数据资产在全生命周期管控中的准确性和可用性。数据标注是连接数据资源、算法模型与实际应用场景的关键桥梁，是挖掘数据要素价值的关键环节，是人工智能高质量数据集的核心生产力。在当今信息化、数字化、智能化的时代，数据标注服务产业已经成为推动人工智能发展的重要环节。

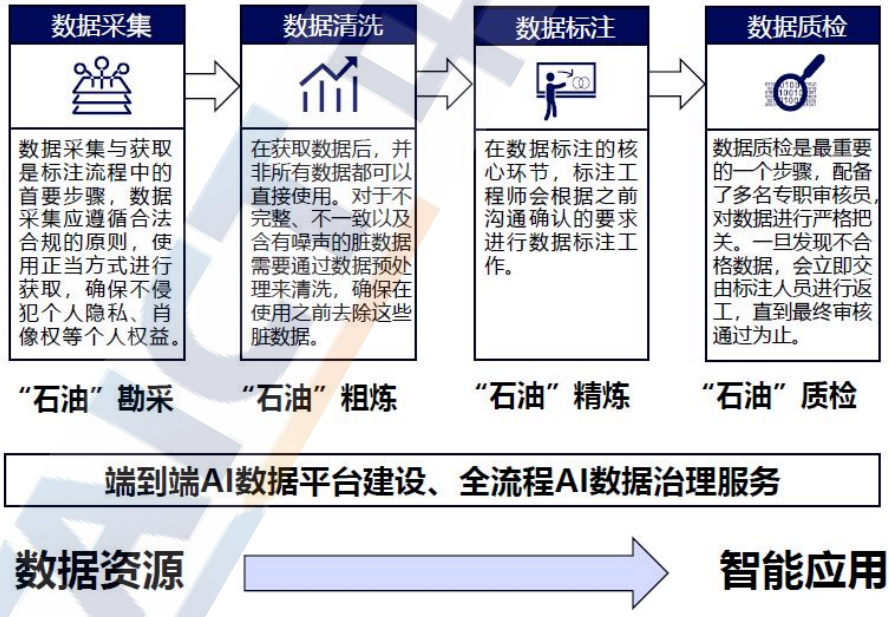
## （一）数据标注定义范畴

**从狭义角度来讲，**数据标注产业是指对未经处理的原始数据添加说明、解释、分类或编码的过程，以便数据可以被人工智能算法所理解和使用。这一过程主要是通过人工或半自动的方式，针对特定的数据集进行标注，以形成具有特定格式的结构化数据。通过高质量的数据标注，人工智能系统能够学习到更为丰富和真实的特征信息，进而提升其在各类应用场景中的表现力和泛化能力。狭义的数据标注旨在为人工智能提供标准化“教材”，助力机器实现更为精准和高效的处理与决策。

具体来说，数据标注包括文本标注（如分词、词性标注、命名实体识别等）、图像标注（如目标检测、图像分类、语义分割等）、视频标注（如行为识别、动作识别、目标跟踪等）、语音标注（如语音识别、语音分割、语音情感分析等）以及 3D 点云标注（3D 点云分割、3D 点云语义分割、3D 点云图像标注、3D 点云连续帧等）。

这些标注工作为机器提供了大量的高质量训练数据。通过学习这些标注数据，机器能够更准确地理解和解析人类语言、图像、视频和语音等信息，从而提升其在自然语言处理、计算机视觉、模式识别等不同领域的性能和应用效果。

**从广义角度来讲，**数据标注产业是指以数据标注为核心的人工智能数据服务上中下游产业链，涵盖数据服务的全生命周期,具体包括数据采集、数据清洗、数据标注、数据质检等多个环节.这些环节的协同发展推动了数据要素产业的持续健康发展，并为人工智能产业的快速发展提供了坚实的基础。**数据标注是对数据进行筛选、清洗、分类、注释、标记和质量检验等加工处理的过程，是提升人工智能算法、模型核心能力的关键环节。**




来源：中国信息通信研究院

图 1 广义的数据标注产业定义

广义的数据标注产业超越了单一的数据处理环节，包括从原始数据到加工形成高质量数据集的数据基础服务全流程,它涉及到数字经济发展的发展战略和数据资源的整体规划。这一产业不仅承载着推动数据资源汇聚、提升数据质量和盘活数据要素价值的使命，更是数字经济体系中不可或缺的一环。通过加强顶层设计和战略布局，优化数据标注产业的发展环境，可以进一步释放其潜力，助力数字经济实现更快速、更可持续的发展。

总结来讲，狭义的数据标注产业主要关注数据的标注过程和结果，不涉及数据收集、清洗等其他环节，它强调的是如何将人类知识转化为机器可理解的形式。广义的数据标注产业则不仅关注数据的标注本身，还涵盖了与之相关的整个人工智能数据服务产业链和生态系统，通过整合与优化各环节资源，推动人工智能技术的持续进步与广泛应用。

### （二）数据标注方式类型

**按照标注方式分类，数据标注主要分为人工标注、半自动标注与全自动标注，当前仍以人工标注为主。**人工标注是指全程手工标注，该种标注方式准确率较高但效率极低。半自动标注以人工标注为主，在标注过程中利用人工智能能力形成辅助工具帮助实现自动贴边、自动分割等功能，从而提高人工标注效率。全自动标注是指利用人工智能算法自动生成标注，该种方式标注效率较高但在复杂度和精细度要求较高的场景仍需要人工审核。

**按照数据类型分类，当前人工智能领域数据标注可以分为文本、图像、语音、视频和 3D 点云标注。**

表 1 数据标注类型

| 标注类型 | 概述 | 标注子类 | 应用领域 |
| --- | --- | --- | --- |
| 文本 | 文本标注是对文本进行特征标记的过程，对其打上具体的语义、构成、语境、目的、情感的数据标签，标注好的数据可以使机器更加深入地理解人类语言。 | 词性标注<br>分类标注<br>情绪标注<br>命名实体识别<br>语义标注等 | 智能客服<br>智能病历<br>智能营销<br>知识管理等 |
| 图像 | 图像标注是将标签附加到图像上的过程，可以是给整个图像添加一个标签，也可以是给图像中的每一组像素分别添加多个标签。 | 拉框标注<br>语义分割<br>关键点标注<br>OCR 转写等 | 人脸识别<br>OCR<br>医学影像处理<br>自动驾驶<br>目标检测 |
| 语音 | 语音标注是指将语音中包含的文字信息、各种声音标记提取出来，再进行转写或者合成的过程。 | 语音切割<br>声纹识别<br>音素标注等 | 智能语音转写<br>智能语音合成等 |
| 视频 | 视频标注以图片帧为单位，对视频素材中的目标对象进行跟踪，对包括道路、车辆、行人等在内的目标物的特征信息、结构信息、语义信息等进行标记，形成训练数据集。 | 视频追踪<br>标签分类<br>视频打点<br>视频信息提取 | 目标跟踪<br>行为识别等 |
| 3D 点云 | 3D 点云标注是指在 3D 图像中，通过 3D 框将目标物体标注出来 | 3D 点云分割<br>3D 点云语义分割<br>3D 点云图像标注<br>3D 点云连续帧 | 自动驾驶<br>无人机<br>机器人 |
| 多模态数据 | 具身智能等领域所需的多模态数据集 | 视觉（RGB）、本体状态(机器人关节角度、末端执行器位置）、语言指令等 | 具身智能 |

来源：中国信息通信研究院

### （三）数据标注服务模式

**数据标注作为人工智能产业链中的关键环节，其组织服务模式对于推动整个行业的发展具有重要意义。**当前，数据标注产业的组织服务模式主要包括集中式、分布式以及混合模式等三种类型。

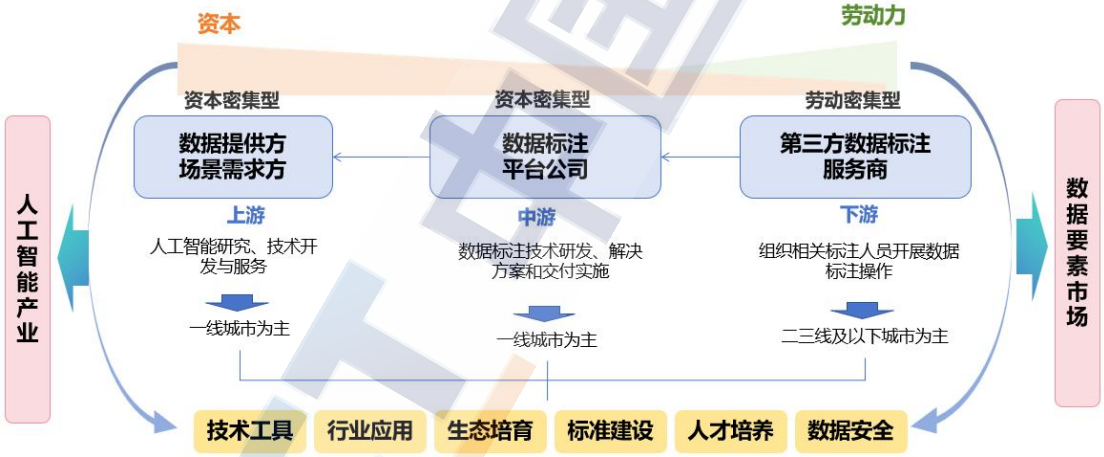
**集中式组织模式是指由大型企业或机构主导，通过集中资源和人力自建团队进行大规模的数据标注工作。**全职标注团队由专业数据标注员组成，经过专业有效的培训及质控手段，能够按照统一的标准和流程进行标注，能够提供较高的数据标注质量，业务匹配性较高，但是需要投入大量的人力和物力资源，总体标注成本较高。

**分布式组织模式是指通过众包、外包等方式将数据标注任务分配给多个团队或个人完成，具有较高的灵活性和效率。**该服务模式的优点是可通过利用互联网平台上的广大用户群体进行数据标注，能够快速收集到大量标注数据，总体成本较低，样本多样性较强。但是，由于分布式标注者的专业水平和责任心参差不齐，标注数据的质量可能存在较大差异。

**混合模式是指通过众包模式和集中模式相结合的方式，根据项目的具体需求，灵活选择标注人员，形成混合标注团队，以优化标注工作的整体效果。**混合模式既可以降低成本，又能很好地保证标注质量，目前，越来越多的人工智能数据服务企业在实践中选择采用混合模式，以充分利用众包模式和集中模式的优势，实现高效、精准的数据标注。

## （四）数据标注产业结构

**数据标注产业链呈现清晰的“需求-平台-执行”三层架构。**数据标注产业链上游是人工智能数据提供方和应用需求方，主要从事人工智能研究、技术开发与服务，根据自身业务提出数据需求，作为数据智能化应用需求的源头驱动产业发展；中游是数据标注平台公司，主要依据需求开展数据标注技术研发、制定加工实施方案和交付，众包、分包给第三方数据标注服务方，通过标准化流程连接供需两端；下游服务商依托人力资源优势完成具体标注任务，形成产业闭环。具体如下图 2 所示。




来源：中国信息通信研究院

图 2 数据标注产业链情况

**人工智能数据标注产业图谱呈现"基础供给-价值转化-生态保障"的立体化架构。**资源提供方提供原始数据，同时又是数据标注业务的场景赋能对象，其中通用人工智能企业与互联网企业兼具数据资源提供方和高质量标注数据使用者双重角色；科学研究、工业制造、

现代农业、智慧能源、交通运输等行业场景企业形成垂直领域数据生态化布局；公共数据作为开放性基础资源，通过非私密性标注赋能人工智能模型智能化升级，构建起覆盖多元主体与场景的数据标注产业体系。数据标注核心服务方提供数据标注技术服务、平台服务、交易服务和人力服务，技术服务方与平台服务方通过数据全生命周期处理技术及智能化管理平台，为标注流程提供底层技术支撑与流程优化能力；交易服务方与人力服务方构建起数据要素流通的交易枢纽和人力协同网络，形成从需求对接到质量保障的全链条服务价值闭环。配套支撑方聚焦行业可持续发展要素，通过标准应用机构建立基础性规范、人才培养机构打造专业人才梯队、生态培育机构推动产业协同创新，安全保障机构则构筑数据隐私与合规防线，形成涵盖质量保障与能力升级的支撑能力。

### （五）数据标注发展意义

数据标注通过提升数据质量、推动资产化转型，成为释放数据要素价值的核心引擎，同时作为人工智能技术落地的底层支撑，驱动垂直领域智能化应用突破。这一过程不仅强化了数据要素与人工智能的互促关系，更形成了覆盖技术、产业、标准的完整价值网络，为数字经济高质量发展注入持久动能。

**数据标注成为数据价值提升的核心驱动力。**数据标注通过标准化处理和语义赋予，推动原始数据向高价值资产转化，成为释放数据要素经济潜能的关键基础设施。在数据质量层面，标注过程通过

规范化、系统化处理，消除了原始数据的杂乱性和非结构化特征。例如，自动驾驶领域对道路物体（信号灯、行人）的精准标注，显著提升了数据集的可读性与机器学习效率，使数据从“可用”迈向“好用”。在资产化路径上，标注后的数据形成标准化产品，具备明确的商业价值。以医疗影像标注为例，标注后的 CT 图像数据可被用于疾病预测模型训练，其市场价值是未标注数据的数十倍，推动数据从“潜在资源”向“可交易资产”转变。此外，标注通过赋予数据多维语义，拓展了其在智能化场景中的应用能力。例如，金融领域的风险控制模型依赖标注后的用户行为数据，零售行业通过商品图像标注实现智能货架识别，数据要素由此渗透至各行业核心业务链条，形成从数据采集到价值释放的完整闭环。

**数据标注成为人工智能技术应用的核心支撑。**数据标注是人工智能技术从理论到实践的必经环节，为算法训练提供关键燃料，并通过行业适配推动技术纵深发展。在模型训练阶段，数据标注为人工智能系统提供“学习样本”。以法律文书智能分析为例，需对数万条司法文本进行案件类型、争议焦点等标签标注，机器通过反复学习标签特征，最终实现司法文书的自动分类与摘要生成。在垂直领域应用中，行业级标注推动人工智能深度适配复杂场景。智慧农业中，标注后的作物病虫害图像数据助力无人机识别病害类型；智能安防领域，视频流的人脸、行为标注使监控系统能实时预警异常事件，技术应用边界持续扩展。此外，数据标注通过建立标签与现实

的映射关系，增强了人工智能系统的透明性与可解释性。例如，医疗诊断模型通过标注数据可追溯病灶识别逻辑，金融风控系统据此展示贷款审批依据，显著提升高风险领域用户对黑箱模型的信任感，为技术规模化落地奠定基础。

**数据标注成为数据要素与人工智能融合的创新加速器。**数据标注作为数据要素与人工智能技术的连接器，构建起从数据采集到产业应用的完整生态闭环，驱动两者互促共生。产业链协作层面，数据标注串联起资源方、技术方与应用方。例如，公共数据平台（如城市交通摄像头数据）经标注后，由技术服务方提供给自动驾驶企业，形成“数据采集-标注加工-模型训练-场景应用”的全链条协作体系。跨行业创新层面，标注数据的开放流通催生跨界融合。智能家居企业联合人工智能芯片厂商优化语音识别模型，依赖标注后的用户语音交互数据；医疗健康领域整合基因数据与临床标注信息，加速精准医疗研发进程。生态基础设施层面，标注服务与配套支撑方共同夯实产业基础。标准应用机构制定的标注规范（如 ISO/IEC 数据质量标准）、人才培养机构输出的标注工程师（年均缺口超 30%）、安全保障机构的数据加密技术，构建起高效、安全、可持续的数据标注产业生态。这一过程中，数据标注不仅提升了数据要素的市场化水平，更通过技术与场景的双向赋能，推动人工智能与各行业深度融合，形成覆盖技术研发、产业应用、标准制定与人才储备的完整价值网络。

## 二、数据标注产业发展现状和机遇

数据标注产业在 AI 技术推动下快速发展，市场规模不断扩大，应用领域广泛。总体看，伴随着政策支撑体系的不断完善、产业生态的不断健全、自动化和智能化工具的普及、新兴市场的崛起、数据隐私保护需求的增加和数据标注市场的蓬勃发展等，都为行业带来了新的机遇。

### （一）“央地一体”的政策体系初步建立

数据标注作为数据产业发展的基础核心环节，其发展受益于国家大数据战略与人工智能战略的共同推动。近年来，我国各级政府在数据标注产业方面给予了较大的政策支持，全面、高质、快速推动数据标注产业的健康发展。

#### 1.国家层面，顶层设计不断完善

为抓住人工智能发展的重大战略机遇，构筑我国人工智能发展的数据先发优势，近年来国家政策利好频出，国家政策文件对激活数据要素潜能、加速释放人工智能技术红利做出新部署，政策中多次提及数据标注、流通、共享、交换、审核、验证，以及数据真实性、可解释性、准确性、公平性，这对于数据本身以及数据服务流程带来新的规范要求，需要从内部产品、外部合作、多方协同等角度保证数据服务合规性，并探索与新场景、数据资产化/数据要素的融合。

**一是总体谋划阶段。**国务院发文明确了发展数据标注产业的必要性和紧迫性，并列举了数据标注的多个关键处理流程。2017 年 7 月，国务院印发了《新一代人工智能发展规划》，明确提出人工智能作为国家战略科技力量的地位，政策的实施将对数据标注产业的发展产生广阔的市场需求和技术创新动力。2022 年 1 月，国务院印发《“十四五”大数据产业发展规划》，强调“强化高质量数据要素供给”“支持市场主体依法合规开展数据采集，聚焦数据的标注、清洗、脱敏、脱密、聚合、分析等环节，提升数据资源处理能力，培育壮大数据服务产业”。

**二是产业布局阶段。**为加快推动数据标注产业发展，2024 年 4 月 1 日，全国数据工作会议提出“探索建设数据标注基地”。国家数据局认真研究布局，确定了承担数据标注基地建设任务的省份，并由这些省份推荐，明确了所在省份承担数据标注基地建设任务的城市。2024 年 5 月 24 日下午，国家数据局党组书记、局长刘烈宏在第七届数字中国峰会主论坛上发布了承担数据标注基地建设任务的城市名单，分别为四川省成都市、辽宁省沈阳市、安徽省合肥市、湖南省长沙市、海南省海口市、河北省保定市、山西省大同市。

**三是全面推动实施阶段。**2024 年 12 月，国家发展改革委等四部门联合发布《关于促进数据标注产业高质量发展的实施意见》，从推动产业技术创新、加快培育壮大市场主体、培育良好产业生态、加强人才队伍建设、强化试点基地引领、健全安全保障体系等方面，

引导和规范数据标注产业健康发展，为数字经济和人工智能发展提供坚实的基础和动力。意见提出到 2027 年显著提升数据标注产业的专业化、智能化及科技创新能力，年均复合增长率超 20%，培育企业、打造创新载体、建设基地，完善产业生态，形成新格局。在**标准能力建设**方面，国家数据局规范高质量数据集格式和质量要求，明确数据标注的目标和对象。国家数据局以数据“供得出、流得动、用得好、保安全”为主线，推动数据领域标准化建设。2024 年 9 月，国家发展改革委等部门印发《国家数据标准体系建设指南》，充分发挥标准在激活数据要素潜能、做强做优做大数字经济等方面的规范和引领作用。经市场监管总局批复同意，2024 年 10 月，全国数据标准化技术委员会（SAC/TC609）获批成立，以加快语料领域标准体系顶层设计，统筹推进数据领域国家标准的制定实施和有序衔接，指导有关单位研制高质量数据集格式标准和质量标准。在**人才队伍建设**方面，2024 年 4 月，人力资源和社会保障部、国家数据局等九个部门联合印发《加快数字人才培育支撑数字经济发展行动方案（2024-2026 年）》，提出在人工智能方面实施数字技术工程师培育项目、举办职业技能竞赛活动、增设职称专业，促进数字人才在人工智能等领域创新创业，夯实产业发展的人才基础。国家数据局着力推动数字经济人才队伍建设，逐步解决数据标注高水平人才短缺的问题。在**产业发展**方面，国家数据局先后于 2024 年 10 月 22 日、2025 年 1 月 8 日、2025 年 3 月 20 日、2025 年 6 月 28 日组织召开四

次数据标注产业供需对接会，搭建政产学研用协同平台，推动数据标注产业供需精准对接，不断繁荣数据标注产业生态。2025 年 3 月 18 日-20 日，数据标注基地先行先试现场会在四川省成都市召开，总结数据标注基地先行先试工作开展一年以来的建设成效，着力培育数据标注新业态，大力推动高质量数据集建设，支撑人工智能赋能千行百业。2025 年 4 月 29 日，高质量数据集和数据标注主题交流活动在福州举办，围绕数据标注、高质量数据集等热点议题进行深入交流，共同探讨促进数据标注产业发展良策，共商高质量数据集建设路径。

2.基地层面，示范引领效应凸显

四川省成都市、辽宁省沈阳市、安徽省合肥市、湖南省长沙市、海南省海口市、河北省保定市、山西省大同市等七个承担数据标注基地建设任务的城市作为推动数据标注产业发展的先行示范区，主动作为，在数据标注产业方面出台一系列相关政策。

**技术创新是数据标注智能化、高端化发展的核心驱动。**沈阳市立足东北老工业基地，制定出台全国首个《数据标注科技创新指导意见》和《企业/园区数据标注能力等级评估及认定管理办法》，为数据标注领域技术创新和企业/园区认定提供遵循，差异化发展数据标注产业。

**高质量数据集建设是数据标注的重点目标。**保定聚焦打造京津冀数据标注高地，以深化京津冀协同发展为抓手，打造全国首个行

业高质量数据集评测平台，发布国内首个人工智能数据集质量评估体系，与高等教育出版社形成“结对子”的合作模式，持续构建区域协同重点向“京数保标”“京模冀用”的数据智能产业协同迈进的新范式。

**人才是数据标注产业发展的关键要素。**2025 年 2 月，保定市印发《保定市推进数据智能产业创新发展支持措施》，推出支持数据智能产业创新发展的措施，特别强调大数据/人工智能人才的引育，提供优惠政策，优先资助回国创新人才，并将数据标注等相关职业纳入政府补贴性职业技能培训项目，支持企业技能培训。

**优化产业布局是推动数据标注产业协同发展的重要路径。**2024 年 12 月，大同市印发《大同市数据产业发展三年行动计划（2024-2026 年）》，计划通过“数标扩容行动”，以建设国家级数据标注基地为核心，强化基础能力，依托煤炭等优势行业构建特色数据集，深化校企合作培养标注人才，并建设产业园区，形成全链条服务体系，推动数据标注产业规模化、专业化发展。2024 年 11 月，长沙市出台了《长沙市关于推进国家数据标注基地建设的工作方案》，以打造“全国数据标注产业第一城”为目标，计划到 2026 年，将推进工业制造、医疗健康、教育教学、文化旅游、地理信息等领域标注数据规模达 2500TB 以上，形成 8 个以上行业高质量数据集，并带动数据相关产业规模达到 100 亿元以上。

在 7 个基地的牵引带动下，**北京市、天津市、广东省、湖北省、**

**辽宁省、贵州省、陕西省、江苏省**等 20 多个省市，积极培育和发展数据标注产业，为地方数字经济发展提供新动能。

北京市海淀区正式揭牌全国首个高端数据标注示范基地，该基地致力于四大核心目标：引领数据要素产业生态示范，支撑数据流通以赋能产业创新，加速高质量数据集的开发利用，以及培育高级复合型数据人才。此举不仅是对国家数据要素市场化配置改革的积极回应，也为全国数据标注产业的高质量发展提供了新的路径和示范。

无锡市立足长三角数字经济核心区定位，以数据标注助力人工智能产业链发展，形成特色路径。市数据局将数据标注纳入政策规划，鼓励各区创新数据工作体系，大力吸引头部企业集聚。围绕本地特色产业集群，引培标杆企业，打造文心大模型（无锡）数据生态中心等载体，通过“地方-国家智库-人工智能企业”结对子模式，深化数据要素市场化改革，打造“无锡样本”。

武汉市数据局规划数据标注产业三年发展计划，积极争创国家数据标注基地，并在适宜城区建设产业园区。引培数据标注领军企业，培育细分领域标杆企业、高成长性的中小企业，打造一批数据标注成功案例，以促进人工智能产业链发展和数据要素市场化改革。

内蒙古自治区呼和浩特市新城区与百度智能云达成战略合作，双方将合作共建百度智能云（内蒙古）人工智能基础数据产业基地，将全力发展以 AI 数据标注为代表的基础数据服务产业，进一步释放

数据要素赋能效应，促进数字经济高质量发展。

河南省发展改革委根据国家数据局及省委、省政府的工作部署，于 2024 年 12 月 20 日公布了首批省级数据标注基地建设先行先试城市名单，包括洛阳市、鹤壁市、焦作市、南阳市、商丘市、信阳市及郑州航空港经济综合实验区。这些城市将作为先行试点，在数据标注领域的技术创新、行业应用、生态构建、标准推广、人才就业及数据安全等方面积极探索，旨在形成一套可实践、可复制、可推广的服务模式和建设机制，为河南省数据标注产业的蓬勃发展开辟新路径。

### 3.地方层面，产业生态加快建设

各级地方政府积极出台规划文件和扶持政策，以人工智能基础数据服务为切入点，寻求人工智能及数据标注产业的参与机会。

**建设数据标注服务平台，推动技术创新突破。**各地方推动数据标注产业化“人工”为“智能”，积极开展关键技术攻关，通过技术创新研发自动化和半自动化的标注工具，搭建一体化数据标注服务平台，大幅提升了数据标注效率和数据标注的准确性、安全性。在这一过程中，各地方还注重推动数据标注技术在不同领域的应用与推广，如自动驾驶、医疗健康等，以满足行业对高质量标注数据的需求，为人工智能与数字经济的高质量发展提供坚实支撑。

**建设行业高质量数据集，赋能行业发展。**各地方通过数据标准带动行业高质量数据集建设，赋能传统产业向数字化、智能化转型。

各地注重优化产业生态，通过加快数据标注龙头企业引育，构建完整的数据标注产业链、价值链和生态系统，带动数字经济产业发展。

**带动地方就业，培育高端标注人才。**各地方通过设立实训基地、举办标注职业技能大赛等多种形式，推动产教融合发展，培育高端化标注人才队伍，联合上下游产业链形成对就业的显著带动效应。同时，各地方围绕数据标注技术、行业和地方需求，引导企业积极参与标准编制工作，并积极推动数据领域相关标准在标注过程中的应用，确保标注数据的准确性和一致性，有效提高数据标注质量。

**完善安全体系，确保数据安全合规。**各地方把数据安全作为数据标注基地建设的红线，建立数据分类分级安全保护制度，搭建数据标注安全生产环境，构建数据安全风险防控体系，推动常态化、规范化的数据安全运营。在此基础上，各地进一步加强数据安全技术的研发与应用，如采用区块链、数字水印等先进技术，确保数据的可溯性、完整性和安全性，有效阻止数据篡改、损坏和泄露问题。

**出台数据标注政策，引领区域数据标注产业发展。**山西省省委、省政府高度重视数据标注产业发展，陆续出台《山西省数据标注产业发展规划（2019-2025 年）》《山西省加快数据标注产业发展的实施意见》《关于支持数据标注产业高质量发展的意见（征求意稿）》等系列政策文件，为数据标注企业提供强有力的政策保障。北京市在相关政策文件中明确提出“提升本市人工智能数据标注库规模和质量”，将“加强大模型训练数据采集及治理工具研发、数据清洗、

标注、分类、注释及内容审查等算法及工具”等内容。

## （二）大模型蓬勃发展带来新的数据标注需求

### 1.大模型数据需求呈海量增长

| 主体 | 时间 | 大模型 | 数据量 |
| --- | --- | --- | --- |
| OpenAI | 2018 | GPT-1 | 4.6GB |
| Google | 2023 | PaLM2 | 3.6万亿tokens |
| Google | 2023 | Gemini | 3.3万亿tokens |
| OpenAI | 2023 | GPT-4 | 约40000GB<br>13万亿tokens |
| Meta | 2024 | Llama 3 | 超15万亿tokens |
| 通义千问 | 2025 | Qwen2.5Max | 超20万亿tokens |

增长近1.4万倍

来源：OpenAI、Meta 学术论文及公开资料，中国信息通信研究院整理

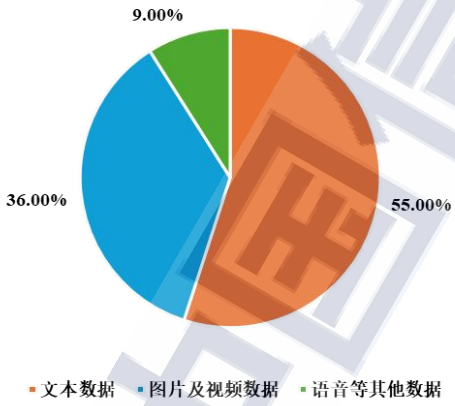
图 3 大模型数据需求海量增长

纵观国际主流大模型，训练数据规模增长近万倍。如图 3 所示，2018 年 Open AI GPT-1 大模型数据量为 4.6GB，2025 年 Qwen2.5Max 大模型数据量超过了 20 万亿 tokens，可见大模型数据需求增长近 1.4 万倍。2023 年 Google PaLM2 大模型数据量为 3.6 万亿 tokens，Google Gemini 大模型数据量为 3.3 万亿 tokens。2024 年 Meta LIama 3 大模型数据量超 15 万亿 tokens。

与传统人工智能相比，大模型在数据需求和数据维度上都有显著不同。首先，大模型通常需要大量数据来实现良好的性能，其训练所需的原始数据规模通常在 TB 至数百 TB 之间，但在训练之前，需将文本等原始数据进行 token 化处理。例如，2024 年 4 月开源的 Llama3，其训练数据集超过 15T token，是 Llama2 的 7 倍。

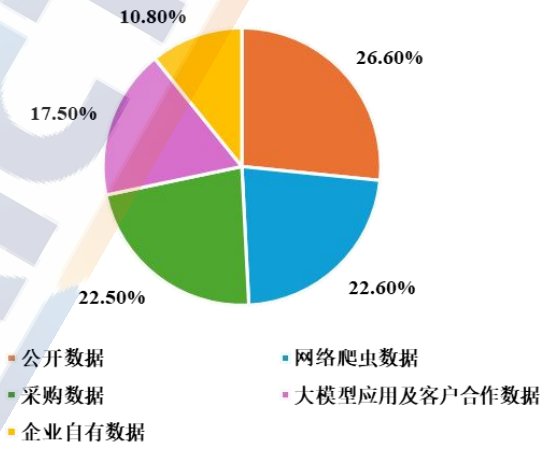
其次，大模型的数据来源极为丰富，涵盖文本、图片、音频和

视频等多种形式，包含海量的知识信息，涉及各类专业领域和多种语言。依托多样化的数据，大模型具备强大的通用能力和迁移能力，能够应对更广泛的任务和场景。像 ChatGPT、Claude、Llama 和 Mistral 等大模型的训练数据涵盖了文学作品、百科全书、新闻、社交媒体以及学术文献等各类知识信息，并且通常还包含图像、视频和音频等多模态数据。




来源：艾瑞咨询

图 4 大模型的各类型训练数据投入构成




来源：艾瑞咨询

图 5 大模型的训练数据来源构成

### 2.大模型数据标注需求呈现新特点

**大模型的数据标注需求贯穿全生命周期，各阶段呈现显著差异。**在预训练阶段，标注需求侧重于海量弱标注或无监督数据的清洗与去噪，需通过文本分类、实体识别等基础标注技术构建高质量语料库，且需覆盖多语言、多领域内容以增强模型泛化能力。监督微调阶段要求高质量指令数据的精准标注，要求构建包含任务描述、输入输出对的精细化样本，标注过程需平衡专业性与多样性。强化学习阶段依赖人类偏好反馈标注，需通过对比排序、质量评分等复杂标注建立奖励模型，标注者需具备领域知识以评估回答的逻辑性、安全性及价值观对齐。持续学习阶段的数据标注更强调动态更新能力，需建立增量数据标注机制，实时捕获新场景、新术语并优化标注策略。

**大模型对标注数据质量要求不断提升。**原因在于模型规模扩大带来的误差放大效应。因此，高质量标注需满足四大核心标准：其一，事实准确性，要求专业领域数据由具备资质的标注员完成；其二，语义一致性，需建立跨场景的标注规范体系，确保相似语境下的标注标准统一；其三，价值安全性，需构建包含伦理审查、内容过滤的多层校验机制，特别是在涉及文化敏感、政治倾向等数据时；其四，场景完备性，要求标注数据覆盖长尾场景，如法律文书中的特殊条款标注需结合具体司法实践。为达到高数据质量，头部企业已采用“交叉验证+AI 质检”混合模式，且建立动态反馈闭环优化标

注规则库。

**大模型落地工程化需求对数据工程提出更高的要求。**大模型产业化落地催生了数据工程范式的根本性变革，推动标注体系向工程化、标准化演进。首先，标注系统需**支持超大规模并发处理**，通过分布式标注平台实现万人级协作，采用自动化流水线技术将标注效率提升 3-5 倍。其次，建立全链路数据治理体系，包含版本控制、血缘追踪等机制，确保从原始数据到训练数据的完整可追溯性。针对多模态场景，需开发跨模态对齐标注工具，如图文对位标注系统需支持像素级区域关联与语义映射。再次，**构建动态评估体系**。通过建立数据质量 KPI 看板，实时监控标注一致性指标、专家复核通过率等关键参数。最后，合规性工程成为刚性要求，需部署数据脱敏、权限分级等技术，并满足 GDPR、网络安全法等法规要求，部分金融场景的数据标注已实现全流程密态处理。

### 3. DeepSeek 开启数据标注的新范式

DeepSeek-R1 模型在后训练阶段大规模使用了强化学习技术。该模型团队在仅有极少数据的情况下，将数据标注视为提升模型性能的核心因素之一，深入到数据标注的每一个环节，确保每一条数据的精准和高效。DeepSeek 开启了数据标注的以下三个新范式:

**自动生成高质量数据集减少传统数据标注需求。**DeepSeek 模型训练采用自动化推理和数据生成技术，推动了智能化标注工具的发展，提升了标注效率和质量，也大幅提升自动化数据标注技术方式

占比，传统数据标注需求减少。

**数据蒸馏+人类协同技术提升数据标注质量和效率。**DeepSeek 通过数据蒸馏技术，从低质量数据中高效提炼生成高质量训练数据，同时采用自动化筛选和人类专家标注反馈机制保障数据标注质量，大幅提升数据标注质量和效率。

**强化学习新范式聚焦高质量推理型数据集。**DeepSeek 聚焦高质量推理数据，收集了大约 600k 的推理相关训练样本和 200k 的非推理训练样本，推理型数据与非推理型数据配比约 3:1，显著改变了监督微调数据的构成，大幅提高了推理型数据的比例。

## （三）数据标注行业与市场蓬勃发展

数据标注产业起源于 1984 年，旨在实现纸质内容电子化。2005 年以后，中国企业逐步涉足标注产业，尤其是 2010 年以后，随着人工智能产业的加速发展，新应用、新场景不断涌现，其海量数据需求持续为包括数据标注在内的人工智能产业链上下游企业带来巨大的发展红利。专业数据服务提供商和头部互联网等数据标注商，以人工标注方式为主，对文本、图像、语音、视频和 3D 点云进行标注，标注市场规模呈现出逐年扩大特点。

### 1.数据标注国内外发展情况

全球数据标注行业是伴随全球人工智能产业发展而生的。1996 年澳鹏（Appen）诞生并布局数据服务领域业务。2007 年数据标注

行业正式拉开序幕，始于斯坦福大学教授李飞飞等人的 ImageNet 项目，该项目要通过亚马逊的劳务众包平台 Mechanical Turk(AMT)来完成图片的标注和处理，得到的数据集供机器算法训练和学习。此后，全球开始涌现出众多的数据标注企业，全球数据标注行业也进入成长期。

**从行业供给情况来看，全球数据标注行业企业主要分布在北美、欧洲、亚太等地区，但具有一定规模的企业数量相对较少。**北美主要集中在美国，数据标注企业较多，突出的特点是技术驱动导向，数据标注服务供给能力和质量较高，代表性企业有 Scale 人工智能、Mighty 人工智能、Mturk、Supervise.ly 等；欧洲地区代表性企业有 Mindy Support 等，但近些年欧洲地区的数据标注企业逐渐将业务转移到人力成本更低的亚太地区和非洲地区等地。亚太地区的数据标注供给能力较为强劲，以中国、澳大利亚和印度为主，代表性的企业有海天瑞声（Speechocean）、澳鹏、Infolks、iMerit 等。中国地区的数据标注行业蓬勃发展，涌现出一批如海天瑞声、数据堂、百度众包、云测等人工智能基础数据服务企业。据企查查数据统计，截至 2023 年，数据标注行业相关企业数达到 1123 家，呈现出井喷的趋势。预计在未来，随着大数据产业的不断发展，数据标注相关企业数量将继续增长。

### 2.数据标注基地产出情况

我国七个数据标注基地分别位于四川成都、辽宁沈阳、安徽合

肥、湖南长沙、海南海口、河北保定和山西大同，据数据标注基地先行先试现场会数据显示，七个数据标注基地数据标注总规模达到17282TB，相当于中国国家图书馆数字资源总量的 6 倍左右；引进和培育标注企业 223 家；标注从业人员达 5.8 万人；带动数据标注行业相关产值超过 83 亿元。[1]截至上半年，7 个数据标注基地已建设数据集 524 个，规模超过 29PB，服务大模型 163 个。[2]

## 三、数据标注产业发展核心要素与实践

数据标注产业发展的核心要素包括从技术创新、行业赋能、生态培育、标准应用、人才就业、数据安全等六方面。在实践中，通过搭建高效的标注平台、培训专业团队、利用自动化工具、实施数据安全措施、制定行业标准、推动跨界合作和创新应用，可以有效推动数据标注产业的发展。


图 6 数据标注产业发展聚焦六大核心任务

[1] 数据来源：国家数据局公众号文章，《我国七个数据标注基地数据标注规模再创新高》

[2] 数据来源：人民邮电报，《国家数据局：我国 7 个数据标注基地已建设数据集 524 个，服务大模型 163 个》

## （一）技术创新

技术创新对于提高数据标注的效率、质量和准确性具有重要意义，是推动产业发展的关键驱动力。通过技术创新和应用，数据标注产业可以更好地满足人工智能发展的需求。当前，数据标注技术创新主要聚焦在自动化标注、众包标注、多模态标注、专家标注、数据预处理技术、模型评估与优化技术等多个关键技术方向。

**一是自动化标注技术。**自动化标注技术利用机器学习和深度学习算法，自动对数据进行标注。这种技术可以显著提高标注效率，减少人工参与，降低成本。例如，商汤科技通过大模型技术对自动驾驶的路测回流数据进行自动标注和重建。然而，自动化标注技术在某些复杂任务上可能无法达到手动标注的准确性，因此仍需要与人工标注相结合。

**二是众包标注技术。**众包标注技术通过引入激励机制、质量控制和任务分配策略，将数据标注任务分发给大量网络用户，从而提高标注效率。这种技术可以充分利用互联网上的闲置人力资源，但需要注意保证标注质量和一致性。

**三是多模态标注技术。**随着多模态数据（如文本、图像、音频和视频等）在人工智能应用中的广泛应用，跨模态数据标注技术变得越来越重要。例如，利用注意力机制等技术，关注多模态数据中的关键信息，提高标注的准确性和效率。

**四是数据预处理技术。**数据预处理技术在数据标注之前对原始

数据进行清洗、去噪、归一化等操作，以提高数据质量。良好的预处理技术可以降低标注难度，提高标注准确性，确保数据在进入标注环节前已经具备高可靠性和可用性。

**五是模型评估与优化技术。**模型评估与优化技术用于全面评估数据标注模型的性能，引入多样化的评估方法，如混淆矩阵、ROC 等定量指标，并结合定性分析方法，如交叉验证等，企业可以更准确地发现标注模型中的潜在问题。

**案例 1 多模态数据智能标注创新**

多模态数据智能标注平台致力于打破国外在 AI 训练数据方面的技术垄断，建立以可控数据自驱的多模态训练数据生产新业态，打造以数据为核心的新质生产力赋能千行百业。在视觉数据、语音数据、文本数据、多模态的数据标注领域，加速推动技术创新，建设了达到国际领先水平的多模态数据智能标注与管理平台。

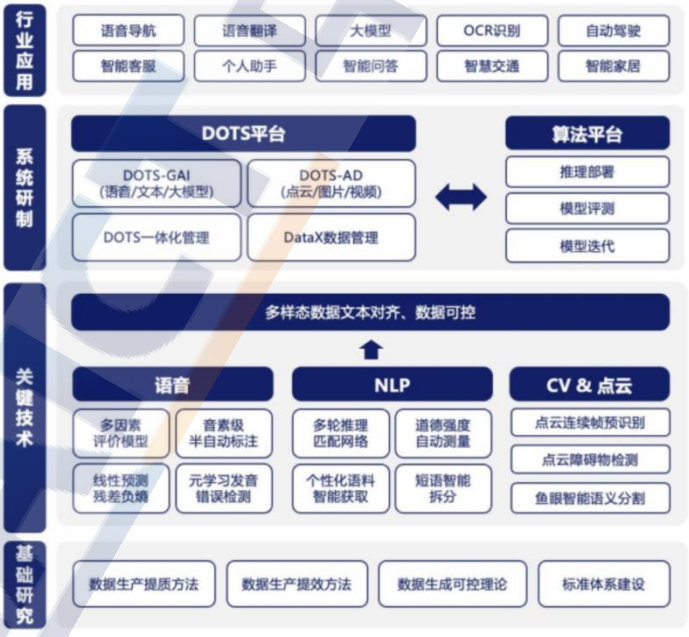



图 7 多模态数据智能标注平台总体架构

## （二）行业赋能

数据标注在不同行业领域的应用场景广泛且深入，为人工智能产业的发展提供了坚实的基础和强大的动力。行业赋能重点围绕科学、制造、农业、能源、交通、金融、医疗、教育、消费、互联网治理、人力资源领域、公共安全等行业领域典型应用场景，通过数据标注可形成一批面向人工智能产业应用的高质量训练和评测数据，为机器学习、深度学习、自然语言处理等算法提供有价值的训练数据，推动人工智能技术的不断进步和应用场景的多样化发展。

**科学领域。**汇聚实验数据、观测数据、计算模拟数据、文献与出版数据、基础学科数据、学科融合数据等关键基础数据，打造面向基础科学研究、科学计算、工程技术应用等典型应用场景的高质量数据集，推动科学各领域的进步和发展。

**制造领域。**深入到生产制造的各个环节，多维度、全方位汇聚生产过程数据、质量控制数据、物料与供应链数据、环境与资源数据等关键基础数据，打造面向智能制造、数字孪生、质量控制、供应链管理、故障诊断等典型应用场景的高质量数据集，助力企业实现更加高效、精确、灵活的生产模式，促进制造业的数字化转型升级。

**农业领域。**汇聚土地资源环境、气象环境数据、作物生长数据、农业生产数据、农业技术数据、农产品市场数据等关键基础数据，打造面向精准农业、智能养殖、气象预警与灾害应对、农业电子商

务、智能农机应用等典型应用场景的高质量数据集，助力国家农业科学领域的技术发展与产业升级。

**能源领域。**汇聚能源资源数据、能源生产与供给数据、能源消费数据、能源市场数据、能源基础设施数据等关键基础数据，打造面向智慧能源管理、智能电网、清洁能源与储能应用、能源交易典型应用场景的高质量数据集，更好地服务于能源系统的优化运行与社会可持续发展。

**交通领域。**汇聚交通基础设施数据、公共交通运营数据、交通规划数据、交通流量数据、交通事故数据等关键基础数据，打造面向城市交通智能调度与管理、公共交通优化、自动驾驶、智慧停车等典型应用场景的高质量数据集，实现对城市交通的全方位、精细化的管理和控制，提高城市交通的效率和安全性。

**金融领域。**汇聚市场基础数据、客户行为数据、风险与欺诈数据、监管与合规数据等关键基础数据，打造面向风险防控、精准营销、智能投顾、反欺诈等典型应用场景的高质量数据集，全面提升金融服务效能与客户体验。

**医疗领域。**汇聚政策监管数据、医疗机构数据、医疗管理数据、医疗业务数据、医学研究数据等关键基础数据，打造面向临床决策、医疗图像、药物研发、精准医疗、医疗教育教学等应用场景的高质量数据集，推动医学科研创新、医疗服务提升和健康管理智能化发展。

**教育领域。**汇聚课程与教学数据、教育资源与设施数据、教学质量与评估数据、教育管理数据等关键基础数据，打造面向智能教学、在线教育、互动课堂、个性化学习、智能辅导等典型应用场景的高质量数据集，为教育研究、教学改革、教育资源配置以及教育政策制定等提供科学依据和支持。

**消费领域。**汇聚消费者行为数据、产品销售数据、消费偏好数据、消费者满意数据、营销与促销数据等关键基础数据，打造面向线上购物、新零售业态、会员制营销、个性化推荐、移动支付等典型应用场景的高质量数据集，为市场的高效运作和体制机制创新奠定基础。

**互联网治理领域。**汇聚网络安全数据、网络行为数据、网络流量数据、内容数据、法律法规数据等关键基础数据，打造面向网络内容审查与管理、网络言论监管、网络安全防护、青少年保护、网络法治建设等等典型应用场景的高质量数据集，实现更加科学、透明、高效的互联网治理，促进互联网产业健康发展。

**人力资源领域。**汇聚人员基本信息数据、薪酬福利数据、招聘与选拔数据、绩效考核数据、职业发展数据等关键基础数据，打造面向人才招聘与选拔、员工信息管理、绩效管理、员工培训、职业发展规划等典型应用场景的高质量数据集，赋能企业制定合理的人力资源策略，提高企业的竞争力和效益。

**公共安全领域。**汇聚公共安全事件数据、犯罪事件数据、治安

监控数据、应急响应与调度数据、人口流动数据、灾害预警与应对数据等核心关键数据，打造面向智慧安防、应急响应、灾害管理、智慧警务等典型应用场景的高质量数据集，提高公共安全管理效率和准确性。

**案例 2 数据标注赋能医疗领域**

医学影像智能数据标注一体化解决方案在医学影像领域深耕，创新突破型医学影像分割大模型 MISM、人在回路等新一代数据标注关键技术，建设集数据、模型、工具、场景为一体的数据标注创新平台，实现医学影像标注工具的完全自主可控及国产化替代，破解医学影像标注工具严重依赖进口、金标准数据依赖国外数据源的瓶颈。多模态医学影像数据标注平台取得突破，完成医学影像分割大模型研发，搭建医学影像标注云服务平台，集成了 100 个预标注算法，同步完成人工智能训练场架构搭建，有效赋能卫生发展。


图 8 医学影像智能数据标注解决方案

## （三）生态培育

生态培育是数据标注产业发展过程中的重要环节，涉及整合各类产业资源，旨在打造健康发展的产业链和生态环境。它不仅确保数据标注工作高效、规范、可持续，还有助于提升整个产业的竞争力和创新能力，推动产业向更高水平发展。

**推动产学研用各方在产业链、创新链、价值链等方面的深度融合，**形成产业生态圈，自上而下规划产业发展目标和实现路径，加强产业园区建设，优化产业布局，提高产业集群效应。发挥产学研用各方的专业优势，为产业发展提供技术支持、人才培养、市场拓展、平台支持等方面的赋能，加快下游产业应用尽快落地，将历史沉淀知识性数据逐渐完成电子化及参与到人工智能训练，推动产业与金融、互联网、大数据等新兴产业的融合发展，提高产业的附加值。

**加快推动地方数据标注产业联盟建设。**区域数据标注产业联盟建设，是促进区域数据标注产业协同发展、提升竞争力的重要举措。加快推动地方数据标注产业联盟建设，需要政府引导、企业参与、资源整合、标准化建设、人才培养、技术创新、市场拓展等多方面的协同努力，构建高效、规范、可持续的联盟体系。

**案例 3 区域数据生态中心建设案例**

区域数据生态中心探索“1+N”发展模式，即一个数据生态中心为产业支点，N 个头部数据企业/数据标注基地多头并进，系统性推进以大模型精标数据为特色的数据产业发展。通过创新开展“地方政府+国家智库+头部 AI 企业”三方结对子的合作模式，政府侧发挥政策和数据优势，智库侧提供权威评估评测体系，头部 AI 企业提供大模型数据精标指导与大模型应用落地能力，打通数据从生产、治理、标注、评估到产业应用的全链条，解决政务、金融、工业、教育、医疗、法务等领域精标数据缺乏的问题，并最终通过行业应用落地实现数据价值的最大化利用


图 9 数据生态中心架构

## （四）标准应用

标准应用具体是指数据标注产业在发展过程中，遵循国际和国内的相关标准，提高数据质量，确保数据标注的准确性、一致性和可靠性，为人工智能的发展提供有力支持。同时，标准应用也有助于促进产业内的公平竞争和合作，推动整个产业的健康发展。

数据需求方、数据供给方及第三方等各方在数据保准标准制定、修订、推广等方面加强合作，推动产业标准体系的完善。从企业内部标准出发，向行业级、省市级以及国家级标准拓展，数据同类型数据加工生产具备相同或相似标准，构建标准完善的数据产品体系，建立产业标准信息服务平台，提高标准的透明度和适用性。

**数据标注标准建设的重点方向。**重点建设数据标注质量标准、数据标注安全与隐私保护标准、标注工具与平台标准、标注流程与协作标准、标注伦理与法律合规标准、标注人员培训与认证标准、

数据标注行业协作与共享标准、数据标注技术创新与自动化标准、国际化与跨领域应用标准。

**数据标注与应用评测能力建设。**围绕人工智能高质量数据集建设要求，搭建人工智能数据集质量评估标准化体系，涵盖标准体系，指标设计，平台开发，案例打造等核心组成部分。开展标注数据的分级分类评价，形成对数据质量的行业共识。

## 案例 4 高质量数据集数据标注标准体系

为贯彻落实《国家数据基础设施建设指引》《国家数据标准体系建设指南》等文件要求，加快规范和引领高质量数据集建设，高效赋能行业发展，在国家数据局指导下，以前期研究工作为基础，全国数据标准化技术委员会牵头研制了高质量数据集 4 项技术文件，包括《高质量数据集 建设指南（征求意见稿）》、《高质量数据集 格式要求（征求意见稿）》、《高质量数据集 分类指南（征求意见稿）》、《高质量数据集 质量评测规范（征求意见稿）》。在数据标注方面，指导编制《高质量数据集数据标注要求》，围绕高质量数据集标注全流程，提出了数据标注的流程要求、技术要求和管理要求等相关内容，明确数据标注技术要求和管理要求，帮助支撑数据标注产业专业化、智能化发展，促进数据科技创新能力提升，提高数据标注的效率和质量，支撑高质量数据集构建，推动人工智能、大数据等技术创新和应用，促进经济社会的高质量发展。

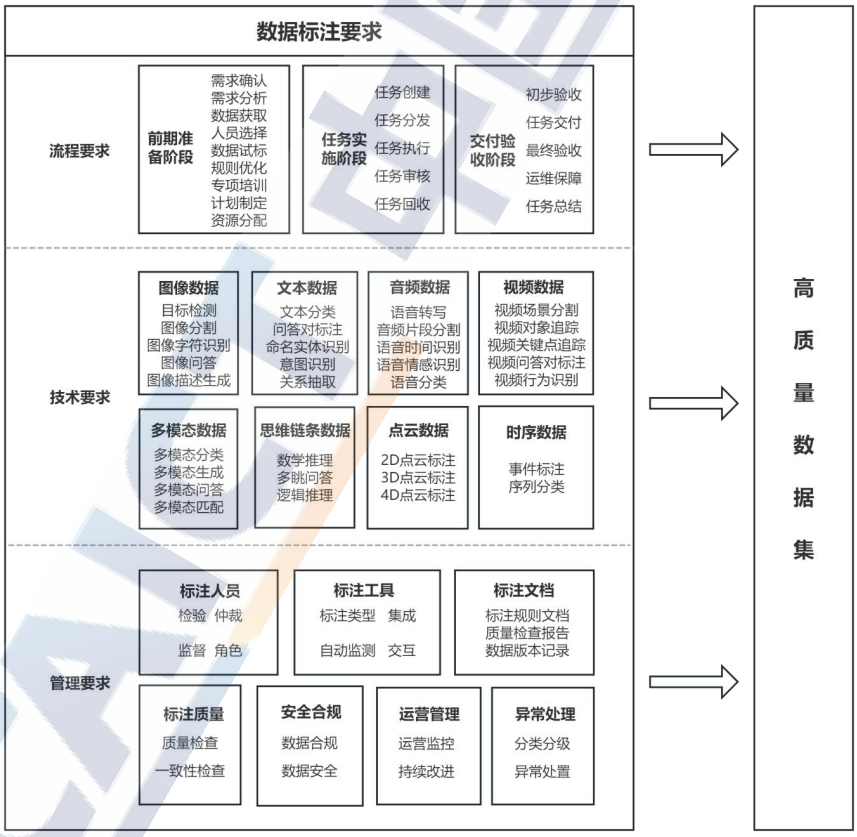



图 10 高质量数据集数据标注标准体系

## （五）人才培养

随着人工智能和大数据技术的快速发展，对高质量的数据标注人才的需求不断增加，数据标注人才的培养成为了推动数据标注产业乃至整个人工智能和大数据产业发展的重要因素。通过新一代高水平数据标注人才培养，数据标注产业可以培养出更多具备专业技能和知识的人才，为产业的发展提供有力支持。同时，人才培养也有助于提高整个行业的竞争力和创新能力，推动数据标注产业的持续发展。

通过开设相关课程、提供社会职业培训以及鼓励校企合作联合培养等，加大数据要素相关人才供给。加强产学研用各方在人才培养方面的合作，建立人才培养基地，开展人才培训、技能竞赛等活动，为从业人员持续提供在职培训和技能提升机会，并在生产过程中遴选优秀人才。完善人才引进、激励、流动等机制，吸引更多优秀人才投身产业发展，保持行业活力和平台竞争力。

**一是教育背景与专业知识。**数据标注人才通常具备计算机科学、统计学、数学等相关专业的教育背景，这些专业知识为数据标注提供了理论基础和技能支持。数据标注人才还应具备对数据的深刻理解能力，以便在标注过程中准确把握数据的内在含义和应用场景。

**二是技能培训和实践经验。**数据标注人才需要掌握数据预处理、标注工具使用、标注流程管理等方面的技能。企业和教育机构一般可以通过培训课程、在线教程等方式进行技能培训，确保人才具备

高效、准确的标注能力和一定的实践经验，以便更好地理解数据标注的实际需求和问题。企业可以通过实习、项目合作等方式为人才提供实践经验，帮助其积累实际操作经验，提高问题解决能力。

**三是持续学习与跨学科学习。**人工智能和大数据技术不断发展，涉及多个领域，所以数据标注人才需要具备持续学习的能力和跨学科的知识，以便在不同领域之间进行有效沟通和协作。因此，应鼓励人才拓展知识面，学习相关领域的知识，提高其在多领域的协作能力。

**四是职业认证体系。**为了提高数据标注人才的专业水平和社会认可度，可以推广职业认证制度。通过职业认证，人才可以获得专业认证证书，证明其在数据标注领域的专业能力。同时，政府和行业组织可以制定统一的职业认证标准，确保认证体系的权威性和公信力。

**案例 5 数据标注产教融合实训体系**

通过打造数据标注产教融合实训平台和搭建数据标注人才培养实训基地，为高校与企业提供一个交流与合作的桥梁。通过实践教学的方式提升学生的数据标注全产业链专业实践技能，并为企业输送高质量的数据处理和分析人才，实现从事人工智能数据工作人员持续培训、职业能力不断提升，为区域提供高质量的人工智能基础数据服务人才梯队。


图 11 数据标注产教融合实训平台设计

## （六）安全保障

数据标注产业是人工智能数据信息处理的重要环节，通过数据安全保障，不仅可以确保数据的安全和隐私，还能为客户提供高质量的数据标注服务，提升客户满意度和信任度，推动整个行业的持续发展和创新。

鼓励科技企业参与到数据运营安全生态的建设中，对行业数据的存储、传输、利用等环节创新透明、可记录、可审计、可追溯的技术手段，促进建立安全可信、管理可控的数据交易环境，提升数据商进场交易的积极性。

**一是合规性遵循。**合规性遵循是数据安全保障的基础，为后续措施提供法律依据，数据标注产业必须遵循国际和国内的相关法律法规，如《网络安全法》、《数据安全法》、《个人信息保护法》等，确保业务开展在合规的前提下进行，保障运行环境安全。

**二是数据加密与访问控制。**在数据传输和存储过程中，采用加

密技术对数据进行保护，保障数据安全，防止数据被泄露、损坏、非法窃取或篡改。实施精准的访问控制策略，确保只有授权人员才能访问敏感数据。同时，对访问人员进行身份验证和权限管理，防止内部人员滥用数据。

**三是数据脱敏与安全审计。**对于包含个人隐私或敏感信息的数据，进行脱敏处理，以保护个人隐私和商业秘密。并定期对数据标注过程进行安全审计，检查是否存在安全隐患和违规行为，帮助企业发现潜在的安全问题并及时进行整改。

**四是安全培训与意识提升。**加强员工的安全培训和意识提升工作，使员工充分认识到数据安全的重要性，自觉遵守安全规定，共同维护数据安全。通过定期组织数据安全培训，让员工了解数据安全的基本概念、法律法规以及在日常工作中防范数据安全风险。

## 四、数据标注产业发展趋势

数据标注产业作为人工智能发展的基石，当前呈现出高技术含量、高知识密度、高价值应用的“三高”特征，预示着其未来发展的广阔前景。同时，也应看到，数据标注产业仍然存在顶层设计尚需完善，高水平人才供不应求，技术创新能力有待提升，专业平台能力不足等问题，制约着产业生态的进一步完善。

### （一）高技术含量

数据标注产业在技术创新方面，呈现以下智能化标注技术深化应用、人机协同标注优化升级、合成数据技术创新突破三大趋势。

**智能化标注技术深化应用。**智能化标注技术不断取得突破，如自监督学习、主动学习、弱监督方法等，能够显著减少对大规模标注数据的需求，提高标注效率和质量。例如，自监督学习通过使用未标注的数据进行预训练，然后在少量标注数据上微调模型，从而降低标注成本。

**人机协同标注优化升级。**人机协同标注模式日益成熟，标注工具更加智能和自动化，标注员更多承担关键决策角色。通过构建持续反馈循环，标注员可以实时纠正模型错误，并将改进反馈给算法，促进其自我优化。这种模式不仅提高了标注效率，还保证了标注的准确性。

**合成数据技术创新突破。**合成数据技术作为新兴领域，正受到广泛关注。它能用人工智能算法生成数据而非真实产生，可替代真实数据来训练、测试和验证大模型。合成数据可以补充更多边缘、长尾场景数据，有效解决大模型时代下的“数据鸿沟”，并自然规避数据隐私安全、合规等问题。

**数据标注平台能力加速发展。**目前，头部大模型企业和数据服务企业均建立了数据处理平台和工具，企业高质量数据集平台化处理能力以及核心数据处理技术显著提升，标注平台的可靠性提升，服务数据的采集、交互、处理、标注和流通等全流程。当前的数据标注平台采用智能化辅助标注工具，配套协作工具和质量控制机制，已具备处理大规模数据集的能力，以满足高效率、高质量的数据标

注需求。目前，数据标注平台的信创国产化水平显著提升，采用国产硬件、操作系统、数据库等先进技术，提升平台的整体性能和稳定性。

## （二）高知识密度

**从业者素质要求的提升。**随着大模型的发展，数据集的评判标准变得更加复杂，要求标注者具备更深层次的理解和分析能力，以及更高的逻辑思维和知识体系要求。同时，处理复杂、多模态数据时，专业技能和学术素养变得尤为重要，数据标注行业对从业者的专业素养要求越来越高，高学历背景和多学科融合成为从业者的基本特征。在大模型时代，数据标注从劳动密集型向知识密集型转变，从业者从高职高专为主体转变为本科及以上学历、多领域专业人才聚集。例如，百度组建的数据标注团队中，学历层次全部达到了本科。

**跨学科知识的融合应用。**数据标注工作涉及多个学科领域的知识，如计算机科学、数学、统计学、语言学等。在自然语言处理领域，标注员需要具备语言学知识，才能准确标注文本的情感倾向、语义角色等信息。同时，随着数据标注应用场景的拓展，还可能需要融合医学、金融、法律等特定领域的专业知识，例如医疗影像标注需要专业知识以识别病灶，自动驾驶领域则侧重于对道路场景的高精度标注。

**人才培养与职业发展的专业化。**数据标注行业将加大对相关人

才的培养力度，提高标注员的技能水平和综合素质。同时，数据标注师的职业发展路径也将更加清晰，可以发展成为算法工程师、数据分析师等更高层次的职位。一些高校和培训机构已经开始开设相关课程和专业，为数据标注产业培养更多高素质人才。为了有效吸引并留住高水平、专业化的数据标注人才，政府、企业亟需构建一套完善的激励机制和福利待遇体系，如具有竞争力的薪酬、舒适的工作环境以及明确的职业发展机会等要素，从而激发人才活力，支撑行业的持续健康发展。

### （三）高价值应用

**应用领域的多元化拓展。**数据标注的应用领域不断拓展，从传统的互联网、安防等行业，逐渐扩展到医疗、金融、教育、制造等多个行业。在医疗领域，通过对医学影像数据的标注，可以帮助医生进行疾病诊断和治疗方案的制定；在金融领域，对文本数据的标注可以用于风险评估、客户信用分析等。

**领域场景的专业化深耕。**在一些特定领域，数据标注呈现出专业化深耕的趋势。例如，在自动驾驶领域，需要对大量的道路场景数据进行精细标注，包括车辆、行人、交通标志等，以提高自动驾驶系统的准确性和安全性。在法律领域，对法律文本的标注需要专业的法律知识，以确保标注的准确性和合法性。

**质量高标准化的推进。**数据标注行业越来越重视质量的高标准化，通过建立统一的数据标注标准和规范，提高数据标注的一致性

和可靠性。质量高标准化包括数据收集标准、分析监控项目过程标准、质量评估标准和审计标准。高质量的标注数据能够更好地服务于人工智能模型的训练和优化，提高模型的性能和泛化能力，从而为各行业的智能化发展提供更有力的支持。同时，质量高标准化也有助于提升数据标注行业的整体竞争力，促进产业的可持续发展。

## 五、推动数据标注产业发展的建议

### （一）不断加强数据标注技术创新能力

高自动化、智能化的数据标注工具，作为推动数据标注产业快速发展的关键支撑，正引领行业发展迈向新高度。建议各地与行业头部企业联手共建联合实验室，持续加大在数据标注工具与机器学习等智能算法融合方面的研究力度，致力于提升标注工具在效率、质量、精度和稳定性等多方面的性能指标。同时，积极开展产学研合作，与高校、科研机构携手共同开展前沿技术研究，加速科技成果向实际应用的转化，持续推动数据标注技术的创新与发展，为产业升级注入源源不断的动力。

### （二）持续提升数据标注行业赋能水平

高质量的行业数据集为传统产业的数字化、智能化转型提供了坚实支撑，有力推动了行业整体发展水平的提升。为了实现这一目标，应深入挖掘各行业的数据标注需求，支持公共数据在多领域的标注与开发利用，并积极推动数据标注服务纳入政府采购范围。同时，鼓励企业加大对数据的开发利用力度，激发企业释放更多的数

据标注需求，共同建设高质量的行业数据集，为人工智能技术在多领域的应用赋能。此外，数据标注企业应与各行业开展深度合作，推动标注数据在金融风险评估、智能制造等具体场景中的应用，助力企业优化业务流程、增强市场竞争力，加速实现智能化转型。

### （三）积极完善数据标注生态体系

加速构建数据标注生态，通过实施“龙头引领+中小微孵化”双轮驱动策略，有利于加速构建完善的产业链、价值链和生态系统。一方面，集中资源培育和引进数据标注龙头企业，发挥其在技术、资金和市场方面的优势，引领产业方向，制定行业标准，推动数据标注技术的创新与应用。另一方面，通过税收优惠、资金扶持和创业空间等为中小微企业提供良好的孵化环境，激发中小企业的创新活力，形成产业链上下游的协同发展。此外，支持龙头企业与中小企业建立紧密的合作关系，促进资源共享与优势互补，共同开展项目研发和业务合作，实现互利共赢。

### （四）大力推动数据标注标准编制和应用

积极推动数据标注标准编制和应用，鼓励数据标注头部企业积极参与数据标准产业标准的制定，构建涵盖技术、质量、流程等多维度的标准框架体系，加快制定国家标准与行业标准，为数据标注提供明确规范。同时，推动标准在实际标注过程中的广泛应用，通过实践不断检验和完善标准体系，促进数据标注产业的规范化与高

质量发展。此外，建立健全标准实施与监督机制，强化对数据标注企业和项目的监督检查，确保标准有效执行。

### （五）着重强化数据标注人才培养力度

加强数据标注人才培育力度。通过设立实训基地、举办职业技能大赛等多种形式，推动产教融合发展，培育高端标注人才队伍，形成对就业的带动效应。此外，支持高校和职业院校开设数据标注相关专业和课程，结合产业需求更新教学内容，培养适应数据标注产业发展的专业人才。鼓励行业联盟、高校、科研院所与企业建立长期合作机制，共同开展科研项目和人才培养，实现资源共享、优势互补，推动数据标注技术的创新和应用。

### （六）切实保障数据安全可靠

持续强化数据安全保障措施，搭建数据标注安全溯源机制、推动数据标注安全生产环境建设、开展数据合规认证、建立完善的数据安全管理体系，加强数据在采集、传输、存储、处理等全生命周期的安全防护，采用加密、权限管理等技术手段，防止数据泄露、篡改和滥用。此外，加强员工的数据安全培训，提高安全意识，定期开展安全审计和风险评估，及时发现和整改安全隐患，确保数据标注过程的安全可靠。

# 附 录


来源：中国信通院

附图 1 人工智能数据标注产业图谱（2024 年）

附表 1 国家层面关于数据标注相关政策文件

| 发布时间 | 政策名称 | 发布机构 | 相关内容 |
| --- | --- | --- | --- |
| 2025.01 | 《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。 |
| 2024.12 | 《关于促进数据产业高质量发展的指导意见》 | 国家发展改革委、国家数据局等六部门 | 从规划布局、技术创新、数据资源开发利用、数据流通交易、基础设施支撑、安全保障和产业发展环境等八个方面提出了具体措施，旨在优化产业结构、培育多元经营主体、提升技术创新能力、增强数据资源供给、促进数据合规流通交易、强化基础设施互联互通、提高数据安全保障能力以及完善产业发展环境，从而推动数据产业高质量发展，形成区域协同的发展格局。 |
| 2024.01 | “数据要素×”三年行动计划（2024—2026 年） | 国家数据局 | 聚焦重点行业和领域，挖掘典型数据要素应用场景，培育数据商，繁荣数据产业生态，激励各类主体积极参与数据要素开发利用。加强试点工作，探索多样化、可持续的数据要素价值释放路径。推动在数据资源丰富、带动性强、前景广阔的领域率先突破，发挥引领作用。充分发挥市场机制作用，强化企业主体地位，推动数据资源有效配置。更好发挥政府作用，扩大公共数据资源供给，维护公平正义，营造良好发展环境。推动数字经济领域高水平对外开放，加强国际交流互鉴，促进数据有序跨境流动。坚持把安全贯穿数据要素价值创造和实现全过程，严守数据安全底线。 |

（续表）附表 1 国家层面关于数据标注相关政策文件

| 发布时间 | 政策名称 | 发布机构 | 相关内容 |
| --- | --- | --- | --- |
| 2023.07 | 生成式人工智能服务管理暂行办法 | 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，提升尊法守法意识，监督指导标注人员规范开展标注工作。 |
| 2023.01 | 关于促进数据安全产业发展的指导意见 | 工信部等十六部门 | 面向数据安全合规需求，发展合规风险把控、数据资产管理、安全体系设计等方面的规划咨询服务。深度分析工业、电信、交通、金融、卫生健康、知识产权等领域数据安全需求，梳理典型应用场景，分类制定数据安全技术产品应用指南，促进数据处理各环节深度应用。 |
| 2022.12 | 关于构建数据基础制度更好发挥数据要素作用的意见 | 国务院 | 加快推进数据采集和接口标准化，促进数据整合互通和互操作，在数据采集汇聚、加工处理、流通交易、共享利用等各环节，推动企业依法依规承担相应责任。承认和保护依照法律规定或合同约定获取的数据加工使用权，尊重数据采集、加工等数据处理者的劳动和其他要素贡献，充分保障数据处理者使用数据和获得收益的权利。 |
| 2022.06 | 国务院关于加强数字政府建设的指导意见 | 国务院 | 建立健全数据治理制度和标准体系，加强数据汇聚融合、共享开放和开发利用，促进数据依法有序流动，充分发挥数据的基础资源作用和创新引擎作用，提高政府决策科学化水平和管理服务效率，催生经济社会发展新动能。 |
| 2022.03 | 中共中央国务院关于加快建设全国统一大市场的意见 | 国务院 | 加快培育数据要素市场，建立健全数据安全、权利保护、跨境传输管理、交易流通、开放共享、安全认证等基础制度和标准规范，深入开展数据资源调查，推动数据资源开发利用。 |
| 2022.01 | “十四五”数字经济发展规划 | 国务院 | 支持市场主体依法合规开展数据采集，聚焦数据的标注、清洗、脱敏、脱密、聚合、分析等环节，提升数据资源处理能力，培育壮大数据服务产业。推动数据资源标准体系建设，提升数据管理水平和数据质量，探索面向业务应用的共享、交换、协作和开放。 |

（续表）附表 1 国家层面关于数据标注相关政策文件

| 发布时间 | 政策名称 | 发布机构 | 相关内容 |
| --- | --- | --- | --- |
| 2021.11 | 工业和信息化部关于印发“十四五”大数据产业发展规划的通知 | 工业和信息化部 | 加快数据要素化，开展要素市场化配置改革试点示范，发挥数据要素在联接创新、激活资金、培育人才等的倍增作用，培育数据驱动的产融合作、协同创新等新模式。推动要素数据化，引导各类主体提升数据驱动的生产要素配置能力，促进劳动力、资金、技术等要素在行业间、产业间、区域间的合理配置，提升全要素生产率。 |
| 2021.05 | 全国一体化大数据中心协同创新体系算力枢纽实施方案 | 国家发展改革委 | 建设数据共享、数据开放、政企数据融合应用等数据流通共性设施平台，建立健全数据流通管理体制机制。试验多方安全计算、区块链、隐私计算、数据沙箱等技术模式，构建数据可信流通环境，提高数据流通效率。探索数据资源分级分类，研究制定相关规范标准。 |
| 2021.01 | 《要素市场化配置综合改革试点总体方案》 | 国务院 | 建立健全高效的公共数据共享协调机制，支持打造公共数据基础支撑平台，推进公共数据归集整合、有序流通和共享。探索完善公共数据共享、开放、运营服务、安全保障的管理体制。优先推进企业登记监管、卫生健康、交通运输、气象等高价值数据集向社会开放。探索开展政府数据授权运营。 |

来源：政府官方文件

附表 2 地方层面数据标注相关产业发展政策

| 省市 | 发布时间 | 政策名称 | 主要内容 |
| --- | --- | --- | --- |
| 河北 | 2025.01 | 《河北省数字技术赋能制造业高质量发展实施方案》 | 通过龙头企业“数字领航”带动工程、中小企业数字化普及应用工程等十大工程，推动制造业数字化转型。数据标注作为数字技术的重要环节，将在这些工程中发挥重要作用，助力制造业的智能化升级。 |
| 山东 | 2024.12 | 《关于组织开展2025年全省数据要素市场化配置改革揭榜挂帅工作的通知》 | 组织开展2025年全省数据要素市场化配置改革揭榜挂帅工作。旨在推进数据要素市场化配置改革，加速培育数据市场，释放数据要素价值。 |
| 河南 | 2024.03 | 《河南省加快制造业“六新”突破实施方案》 | 加快发展人工智能。依托战略支援部队信息工程大学、郑州大学、省科学院、嵩山实验室等高校和科研机构，开展重大基础研究和前沿科学探索。支持郑州数据交易中心探索建设公共数据专栏、社会数据和城市专区，引导国家超算郑州中心、中原人工智能计算中心联合科研机构和龙头企业共建公共数据训练集。加快建设数据标注行业标准体系，支持商丘、安阳市打造高水平数据标注产业集群。到2025年，突破一批关键算法，初步建成较为完善的算法转化与应用生态。 |
| 上海 | 2022.09 | 《上海市促进人工智能产业发展条例》 | 推动人工智能领域高质量数据集建设。支持相关主体将数据与行业知识深度融合，开发数据产品，服务算法设计、模型训练、产品验证、场景应用等需求。鼓励相关主体开展大数据与人工智能技术协同研发，研制数据标注的专业工具和系列标准，建设面向人工智能训练的大数据实验室，构建大规模人工智能数据资源库。 |

（续表）附表 2 地方层面数据标注相关产业发展政策

| 省市 | 发布时间 | 政策名称 | 主要内容 |
| --- | --- | --- | --- |
| 湖北 | 2023.12 | 《湖北省推进人工智能产业发展三年行动方案（2023—2025 年）》 | 方案明确实施“333”发展路径，即以武汉、襄阳、宜昌三大科创中心为核心支撑，以“光谷”“车谷”“网谷”三大区域载体为引领，聚焦产业底座、融合应用、行业服务三大核心领域。目标包括建设 1-2 家全国重点实验室，打造 5 家以上省级创新平台，培育 30 家以上有影响力的人工智能高新技术企业，100 家以上专精特新“小巨人”企业，打造 5 个以上行业大模型和 500 个以上应用示范场景。 |
| 湖北 | 2022.09 | 《关于印发湖北数字经济强省三年行动计划(2022-2024 年)的通知》 | 围绕数字产业化、产业数字化、数据价值化、治理数字化、数字新基建和生态构建等领域,实施六大行动,加快关键要素协同联动、加快进行全省数字经济发展布局,努力打造全国数字经济发展高地。 |
| 湖北 | 2022.03 | 《湖北省人工智能产业“十四五”发展规划》 | 完善数据资源开放共享政策，建立数据资源开放共享机制。引导公共服务机构开放数据，搭建综合性基础数据资源库和共享服务平台，推进公共服务数据资源统一汇聚和集中向社会开放。引导人工智能行业龙头企业或行业协会，建设各类行业数据平台。建立数据共享交换监管制度，强化数据安全与隐私保护，在数据安全的前提下实现数据共享交换。 |
| 黑龙江 | 2022.03 | 《黑龙江省“十四五数字经济发展规划》 | 大力发展数据服务产业，鼓励企业开展数据清洗、脱敏、建模、分析挖掘、应用服务等大数据分析和技术服务，发展数据标注和数据分析企业，拓展采集、交易等专业化数据服务新业态。 |
| 江苏 | 2021.09 | 《江苏省“十四五”数字经济发展规划》 | 突出数据的战略资源和核心要素地位，加大数据资源共享开放，深化数据应用创新，探索数据资源流通交易，加强数据和个人信息安全保护，加速数据资源化、资产化、资本化进程，释放数据要素价值，为数字经济发展提供动力。 |

（续表）附表 2 地方层面数据标注相关产业发展政策

| 省市 | 发布时间 | 政策名称 | 主要内容 |
| --- | --- | --- | --- |
| 重庆 | 2021.12 | 《重庆市数据治理“十四五”规划（2021—2025 年）》 | 积极培育大数据、区块链、人工智能、云计算等数字产业，聚焦数据要素市场产业链，在数据采集、存储、加工、分析、流通等领域落地落户一批技术型企业，培育发展数据抽取、清洗、加工、编目、转换等基础性专业服务。 |
| 北京 | 2024.09 | 《北京市“数据要素×”实施方案（2024—2026 年）》 | 支持北京市各区结合资源禀赋，开发数据应用，打造标杆示范案例。同时，鼓励数据要素型企业通过“揭榜挂帅”等方式参与应用开发，并组织开展相关竞赛，激励社会各界挖掘市场需求。 |
| 北京 | 2024.07 | 《北京市推动“人工智能+”行动计划（2024—2025 年）》 | 北京市在推动人工智能产业发展中，高度重视数据资源的整合、应用和安全保障，通过多领域数据融合创新，助力人工智能技术的落地和应用。 |
| 山西 | 2019.08 | 《山西省加快数据标注产业发展的实施意见》 | 坚持以数据资源为核心。按照“龙头+集聚”的推进路径,聚焦专业领域数据标准化和数据资源价值延伸,积极探索数据服务模式创新,培育构建基础数据服务体系,着力打造国家级数据标注产业基地和数据资源集散地,推动人工智能产业快速发展。 |
| 天津 | 2023.05 | 《关于推动生产性服务领域平台经济健康发展的实施意见》 | 围绕数据计算、存储、交易、清洗、标注、分析、可视化等需求，加快推动数据服务企业向专业化、工程化、平台化发展。重点吸引数据资源服务、在线数据服务等平台企业在津落户，开展数据标注、数据分析、数据咨询等业务，加速服务模式和商业模式创新及产业价值链体系重构。到 2027 年，力争引育超过 10 家数据服务领域平台企业。 |

来源：政府官方文件

附表 3 七个数据标注基地相关产业发展政策

| 城市 | 发布时间 | 政策名称 | 主要内容 |
| --- | --- | --- | --- |
| 大同 | 2024.12 | 《大同市国家级数据标注基地建设实施方案》 | 方案旨在高水平建设国家级数据标注基地，发挥数据要素乘数效应，赋能经济社会发展。方案遵循国家数据局任务书要求，通过“以链带面，面状思维”发展数据产业，聚焦“算力”和“数据”两条建设路径，并围绕“能源、文旅、农业、科技”四个赛道，深入挖掘人工智能技术，开放应用场景，建设高质量数据集。方案还提出打造研发一体化智能标注平台，推动数据标注赋能千行百业，如交通运输、医疗健康、教育教学等，并形成以龙头企业为引领，中小企业蓬勃发展的数据标注产业新生态。 |
| 大同 | 2022.02 | 《大同市加快推进数字经济发展的实施方案和若干政策的通知》 | 政策强调，加快发展大数据产业。以数据标注等产业为切入口，构建集数据采集、清洗、标注、交易、应用为一体的基础数据服务体系。面向行业应用需求，组织开展大数据关键技术攻关，形成垂直领域大数据解决方案。以大同市经济技术开发区为依托，加快推进数字类产业园区或基地建设，引入专业运营服务机构，推进构建智慧园区管理体系，培育大数据产业集群。加大对重点数字企业（项目）的支持力度，发挥示范效益，集聚呼叫产业，建成国内知名、特色鲜明的大数据服务外包及呼叫产业示范基地。 |
| 保定 | 2021.11 | 《保定市支持数据服务产业发展若干措施》 | 保定市在支持数据服务产业发展的政策中，高度重视数据标注领域的发展。为推动数据标注产业生态建设，保定市积极集聚数据服务企业，形成了涵盖数据清洗、标注、治理、加工、评估等细分领域的完整产业链。在政策支持方面，数据标注企业可享受包括房租补贴、税收减免、基础设施配套等在内的多项优惠政策，助力企业降低成本、快速发展。同时，保定市将数据标注职业纳入政府补贴性职业技能培训项目指导目录，支持企业开展岗位技能培训，为数据标注产业提供充足的人才保障。此外，保定市还计划建设智能标注实训平台和一体化智能标注平台，提升数据标注的技术水平和效率，推动数据标注与智能网联汽车、电力高端装备等行业的深度融合，构建高质量数据集，赋能传统人工智能应用场景。 |

（续表）附表 3 七个数据标注基地相关产业发展政策

| 城市 | 发布时间 | 政策名称 | 主要内容 |
| --- | --- | --- | --- |
| 成都 | 2024.12 | 《成都市深化数据要素市场化配置改革工作方案》 | 旨在加快完善促进数字经济发展体制机制，推进数据要素市场化改革，探索建立有利于数据安全保护、有效利用、合规流通、价值释放的数据制度和市场环境，提升数据要素服务经济社会发展能力。主要内容包括构建数据制度规范体系、绘制数据资源全域图谱、赋能人工智能产业及建设国家数据标注基地等方面，旨在通过一系列创新性举措，促进成都数据产业生态的持续健康发展。 |
| 成都 | 2024.06 | 《2024 年成都市数字经济发展工作要点》 | 主要明确了成都市在 2024 年将加快推进数字产业化和产业数字化，旨在打造人工智能产业创新高地，构建现代化大数据产业体系，并聚焦提升公共服务数字化智能化水平，在智慧教育、智慧医疗、智慧养老等场景上拓展应用。同时，工作要点还围绕数字基础设施、数据基础制度、数字技术创新突破等方面部署了多项重点任务，以实现数字经济核心产业增加值占地区生产总值 15%以上的目标，推动成都市数字化综合发展水平迈入全国第一梯队。 |
| 成都 | 2024.06 | 《成都市数据条例》 | 旨在加强数据保护、管理和开发利用，发挥数据要素作用，推动智慧蓉城建设，促进高质量发展。明确了数据收集与治理、流通与交易、应用与促进、安全与保护等方面的规定，强调数据资源利用与权益保障并重、数据价值转化与安全保护并重、数据制度创新与依法监管并重的原则。 |
| 沈阳 | 2025.01 | 《沈阳国家级数据标注基地建设实施方案》 | 主要围绕推进沈阳国家级数据标注基地建设展开，明确了产业锚定方向，提出以技术指导助力产业创新升级，以产业扶持推动产业发展壮大。方案还涉及组建数据标注工作专班和产业联盟，开展数据标注专项招引活动，以及推动数据标注技术在工业制造、数字教育等领域的应用等内容，旨在为企业提供全链条支持，全力营造数据标注产业蓬勃发展的良好政策生态。 |

（续表）附表 3 七个数据标注基地相关产业发展政策

| 城市 | 发布时间 | 政策名称 | 主要内容 |
| --- | --- | --- | --- |
| 沈阳 | 2024.11 | 《沈阳市公共数据授权运营管理办法（试行）》 | 规范公共数据授权运营行为，培育数据要素市场，推动数字沈阳发展建设。明确了公共数据的定义、授权运营的原则及流程，涉及数据汇聚、处理、授权、加工、经营、安全、监管等方面。办法还强调了网络安全和数据安全的重要性，要求运营单位建立严格的安全管理制度，并接受相关部门的监督。此外，还提出了公共数据授权运营平台的集约化建设要求，以及数据授权运营的具体操作流程和期限规定。 |
| 合肥 | 2025.03 | 《合肥数据标注产业发展规划（2025-2027 年）》 | 《合肥数据标注产业发展规划（2025-2027 年）》提出，到 2027 年底，合肥市多语种标注和语音标注能力达到国际领先水平，标注数据规模达 3000TB，构建 11 个以上行业高质量数据集，拉动标注产业规模达 30 亿元，支撑相关产业规模超千亿，打造国际上有特色、有影响力的数据标注基地。产业布局上，形成“一核引领、两区支撑、多园协同、区域联动”的格局，合肥高新区为核心引领区，蜀山与新站高新区为支撑区，各县（市）区建设特色产业园区，同时加强区域联动发展。重点任务包括攻关数据标注关键共性技术、建设支撑项目、培育市场主体、赋能行业发展、营造发展生态和创新运营机制，通过这些举措推动数据标注产业高质量发展，为经济社会高质量发展注入强大动力。 |
| 合肥 | 2020.02 | 《合肥市数字经济发展规划（2020-2025 年）》 | 规划中指出，要支持市场主体依法合规开展数据采集，聚焦数据的标注、清洗、脱敏、脱密、聚合、分析等环节，提升数据资源处理能力，培育壮大数据服务产业。 |
| 海口 | 2024.07 | 《海口市加快数字经济高质量发展若干政策（征求意见稿）》 | 文件提出，大力发展数据标注产业，对于规模在 200 席及以上的数据标注企业，提供坐席补贴，补贴总额为 10000 元/席/年，且补贴时间不超过两年，同时给予专项房租补贴，补贴总额为 2000 元/席/年，该房租补贴持续三年。为了促进人才培育，政策鼓励企业与院校共建大数据相关专业和科研机构，对向数据标注企业输送实习生并达到一定留用率的院校给予实习就业奖励。 |

来源：政府官方文件

# 编制说明

本研究报告在国家数据局数字科技和基础设施建设司的指导下，自 2024 年 6 月启动编制，分为前期研究、框架设计、文稿起草、征求意见和修改完善五个阶段。本报告整体分析了数据标注产业总体发展现状，全面总结了数据标注产业发展的六大核心要素，系统梳理了数据标注产业发展的问题与趋势，并进一步提出了发展建议，为政策制定者、行业从业者及企业投资者等提供全面的行业洞察、策略建议与决策依据。本报告由中国信通院人工智能研究所联合多家单位撰写。报告先后征求并采纳清华大学、北京理工大学、航天二院、赛迪网安所等多位专家意见，以及国家数据局综合司、政策司、资源司、数经司、国合专班意见，形成相关研究成果。

参编单位: 中国电信集团数据发展中心、中电信人工智能科技（北京）有限公司、沈阳市数据局、合肥市数据局、成都市发改革（成都市数据局）、东莞市政数局、呼和浩特市新城区政数局、雅安市发改委（数据局）、航天网信、中国信通院河北科技创新研究院、沈阳数字经济协会、新津区人民政府、中国电信四川公司、东莞市万江街道办事处、招商局、中国联通、中国建筑、中国物流、国家呼吸医学中心、中车研究院、中国东方航空、苏高新国控、广州实验室、海天瑞声、砺英数智、腾讯集团、软通动力、抖音集团、标贝科技、希尔贝壳、柏川数据、景联文科技、云测数据、索贝运维、飞数信息、幸福泉国际人工智能公司、中博伦、识因智能、猎户星空、枫清科技、中关村数智人工智能产业联盟、兰州新区商投集团数投公司。

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn