高端与行业数据标注基地是未来发展重点

▶ 樊威 燕江依 李荪

人工智能（AI）的发展离不开高质量数据的“投喂”，而数据标注工作是高质量数据诞生的基础，也是支撑人工智能技术演进和应用落地的基石。随着人工智能向垂直领域渗透，高端数据标注基地和行业数据标注基地正在成为突破数据瓶颈、释放数据潜能的关键载体。

一、数据标注的重要性日益凸显

数据标注是连接数据资源、算法模型与应用场景的关键“桥梁”，也是人工智能高质量数据集建设的基石与核心生产环节。

（一）数据标注是数据要素价值充分释放的基础

数据标注对释放数据要素价值的意义，主要体现在如下3个方面。

一是有效促进数据流通和共享。数据标注将原始数据从“信息”转化为结构清晰的“资产”，赋予数据明确的语义，使其更容易被不同用户和系统理解和使用，促进数据的流通共享。二是有效增强数据的可用性和精准度。数据标注将原始的杂乱无章的数据转化为结构化、有标签的数据，从而显著提升数据质量和可用性。三是有效提高数据驱动的决策水平。标注后的数据能够为数据分析提供更准确、更有价值的信息，帮助企业和组织机构更好理解数据背后的业务逻辑和趋势，做出更科学的决策。

（二）数据标注是人工智能技术水平提升的关键

数据标注是人工智能的基础性工作，通过给原始数据打上标签，为计算机提供学习数据特征与规律的素材，使模型获得对未标注数据的识别能力，这是模型智能的起点。而高水平的数据标注是模型能力提升的关键，贯穿于模型训练、评估、优化和应用等环节。精准的标注能帮助模型更深刻地理解数据的特征和模式，进一步提高模型的准确性和预测能力。数据标注是高质量数据集构建的核心，通过人工或智能标注对原始数据进行清洗、分类、去噪，才能形成驱动模型迭代的高质量数据集。DeepSeek-V3、GPT-4o等在训练阶段均使用了总量约15万亿token（令牌/词元）、经过严格清洗和标注的高质量数据。

（三）数据标注是人工智能赋能千行百业的支撑

数据标注支撑人工智能在垂直场景中深度应用。

沈阳国家级数据标注基地供图

在医疗领域，医疗影像中病灶标注能够显著提升数据可用性，智源研究院医疗大模型经专业医生标注的影像、病例、文献等数据训练，比通用模型疾病诊断能力提升15%。在自动驾驶领域，数据标注为自动驾驶提供精准、可操作的数据输入，百度自动驾驶大模型Apollo ADFM利用精细标注的车辆、交通标志、运动轨迹等数据，显著提升复杂场景行人识别能力。在工业质检领域，像素级标注通过精确标识缺陷在图像中的具体位置，为高精度缺陷检测模型提供详细监督信息，提升质检效能。此外，数据标注还在智能家居、智慧城市、金融服务、生物医药等多领域多场景得到有效应用。

二、数据标注产业快速发展

当前，我国数据标注产业发展驶入“快车道”，呈现规模扩张与创新实践并进的良好态势。

（一）数据标注工作成效显著

目前，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7个国家级数据标注基地数据标注总规模超过1.72万TB（太字节），约为国家图书馆数字资源总量的6倍，已形成医疗、工业、教育等行业的高质量数据集335个；赋能121个国产人工智能大模型研发；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。

（二）数据标注基地展开实践探索

各个数据标注基地积极承接数据标注任务，并主动展开实践探索。在技术创新方面，研发自动化和半自动化的标注工具，搭建一体化服务平台；在行业赋能方面，通过数据标注带动行业高质量数据集建设，推动传统产业数字化、智能化转型；在生态培育方面，加快数据标注龙头企业引育，构建数据标注产业链、价值链和生态系统；在标准应用方面，围绕数据标注技术和行业需求，引导企业积极参与标准编制和应用；在人才培养方面，通过设立实训基地、举办职业技能大赛等形式推动产教融合，培育数据标注人才；在数据安全方面，探索数据分类分级安全保护制度，构建数据安全风险防控体系，推动常态化、规范化数据安全运营。

（三）数据标注产业供需对接高效开展

目前，数据供需各方积极开展对接，在4次数据标注产业供需对接会上，7个国家级数据标注基地、全国70余个省市级数据管理部门和数百家企业参与，累计签约供需合作80余项，企业-基地签约33项，共2300余人次参会。通过现场签约、央企对接集市及共建可信数据空间等方式，释放企业数据标注需求，支撑重点行业数据要素价值化应用。

三、加快建设高端与行业数据标注基地

随着数据标注产业快速发展，数据标注基地建设呈现清晰的发展路径：一方面是向“高精尖”迈进的高端数据标注基地；另一方面是深耕垂直场景的行业数据标注基地。高端数据标注基地和行业数据标注基地是在7个国家级数据标注基地建设经验总结的基础上，向专业化纵深和区域化广泛覆盖两个方向的演进与补充。国家级数据标注基地将与高端和行业数据标注基地共同构成一个功能衔接、层次分明、协同发展的体系。

（一）加快建设高端数据标注基地

高端数据标注基地是高质量数据供给的关键，具有“高技术含量、高人才素质、高质量把控、高行业价值”的特征，其核心目标是通过人机协同标注、合成数据标注、大模型智能标注等前沿技术，结合多学科知识，实现数据标注的专业化、标准化与高质量输出。

具体而言，高端数据标注基地是以高技术、高水平的数据标注能力强化高质量数据供给，以产教融合新模式培养多元化数据标注人才，以权威的高质量数据集质量评估和模型验证能力体系提升数据质量和模型能力，以数据生态服务矩阵繁荣数据要素市场、促进产业迭代升级。

对此，国家层面应通过政策引导和建设指引，明确高端数据标注基地的建设内容，推动关键技术突破和标准体系完善，与区域数据资源联动，带动数字经济发展。地方政府应激励骨干企业、科研院所等积极参与基地共建，加强技术研发，建立合作网络，形成从需求提出到成果应用闭环。

（二）加快建设行业数据标注基地

行业数据标注基地是人工智能深度应用的重要支撑，具有强行业属性、强场景导向和强专业需求的特征，旨在围绕医疗健康、智慧交通、智能制造、能源电力、金融服务等重点行业场景，提供专业的定制化标注服务，结合行业标准和业务流程，将分散异构的原始数据转化为符合行业应用需求的高质量数据集。

行业数据标注基地的建设，重点面向行业主管部门、龙头企业和产业联盟，特别是对行业数据安全、准确性和专业性要求高的领域。通过推动专业化标注体系建立和行业规范落地，提升行业数据的结构化与可用性水平，形成可复制推广的标注标准，降低企业自行标注成本；同时提升模型在特定任务上的训练效果，推动模型精准解决行业痛点问题。

对此，建议通过政策引导，鼓励龙头央企承担行业数据标注基地建设任务，加强行业数据的合规采集、分级管理与安全流通，打造一批行业标注标准和示范应用典型。鼓励龙头央企牵头搭建行业标注平台，带动上下游企业协同参与，推动跨企业、跨行业的数据共享与标准统一。

四、高端与行业数据标注基地建设需要素保障

高端数据标注基地和行业数据标注基地的建设，除顶层设计外，还需依托完善的要素条件。为此特提出4点思考建议。

一是强化人才保障。数据标注需要既懂人工智能又熟悉行业场景的复合型人才。应加快建设数据标注人才培养体系，支持高校开设相关课程和实践平台，鼓励基地与科研院所、企业等联合开展人才培养。二是建立多元化资金投入机制。标注基地建设周期长、投入大，需建立中央财政引导、地方专项资金配套、社会资本参与的多元化投入机制，提供长期稳定的资金保障。三是加强智能化工具研发应用。传统人工标注成本高、效率低，应加快自动化、半自动化标注工具研发，推动自然语言处理、计算机视觉、生成式人工智能等技术与标注工具深度融合，推动建设一体化的智能标注平台。四是促进产业转型升级，加强示范引领。应引导数据标注企业和平台向高端、智能方向转型，鼓励龙头企业打造分领域特色标注平台。支持有能力的基地先行先试，在技术、标准、安全等方面形成可复制推广的典型经验，促进技术交流和成果转化。

加快建设高端数据标注基地和行业数据标注基地，是推动数据标注产业向深向实发展、释放数据要素价值、支撑人工智能赋能经济社会发展的关键。未来，需推动产学研用协同，共建繁荣产业生态，以高质高效的数据标注，为我国人工智能产业的高水平自立自强筑牢根基。

（作者单位：中国信息通信研究院人工智能研究所。作者：樊威，高级工程师；燕江依，工程师；李荪，高级工程师）