

大力发展数据标注产业 推动我国人工智能创新发展

文 ◆ 清华大学公共管理学院教授 孟庆国

人工智能是新一轮科技革命和产业变革的重要驱动力量。数据标注是对原始数据进行采集、清洗、分类、标记、质量检验等专业数据治理活动，能有效提升数据供给质量，是人工智能发展的关键环节。大力发展数据标注产业，对推动人工智能创新发展、带动大学生就业和促进数字经济高质量发展具有重大意义。

一、发展数据标注产业的重大意义

（一）数据标注是提升数据供给质量的关键环节，是提升人工智能产业竞争力的先手棋

大模型已成为中美科技竞争的制高点，数据供给质量决定了模型的基础能力。当前，全球主流基础大模型，中文语料仅占全部语料的 1%，高质量中文数据成为制约我国基础大模型能力的瓶颈。训练一个领先的大模型，需要数百万甚至数千万条标注数

据。数据标注发展水平成为决定中美人工智能胜负的关键因素。美国在数据标注领域投入巨大，培育了一批数据标注独角兽企业，产业迅速发展。例如，OpenAI 在训练 GPT 系列模型时，投入数千人力和数亿资金进行数据标注，高质量语料集炼就了全球最好性能基础大模型。相比之下，我国虽是全球第二大数据资源国，但数据质量不高，开发利用比例低。全面提升中文语料质量，成为破解我国大模型发展和提升人工智能产业竞争力的关键环节。

（二）数据标注是充分释放数据要素价值的前提条件，是数据要素价值化的必选项

我国是全球第一个将数据资源作为核心生产要素的国家，加快数据要素市场化配置，发挥数据的倍增、叠加、放大效应，是实现数据要素价值赋能经济创新发展的关键。我国是数据资源大国，但不是数据资源强国。2023 年，我国数据资源 32ZB（泽字节），保存数据仅有 2.9%，数据留存率远低于发达国家水平。数据质量低，大量数据被浪费，成为制约我国数据资源价值发挥的瓶颈。数据标注能为机器提供高质量的数据，赋能机器学习、深度学习等人工智能算法的训练，从而实现数据价值转化。以医疗领域为例，通过对大量医疗影像数据进行标注，可以训练出精准的疾病诊断模型，提高医疗诊断的准确性和效率，为患者提供更好的医疗服务。目前，标注数据已广泛应用于生成式人工智能、自动驾驶、电子商务、精准医疗、智能制造等智能场景应用，是数据要素化的必选项。

（三）数据标注产业逐步向智能化、知识化和专业化迈进，是解决大学生就业的有效路径

随着人工智能从通用领域向专业领域演进，数据标注对专业性、技

术性的人才需求日益增长，数据标注产业也从劳动密集型，逐步向知识密集型转变，为大学生创造了数百万就业岗位。比如，美国从事数据标注的岗位超过 30 万，德国最大的数据标注公司 Clickworker 全球雇佣了 120 万标注员，印度到 2030 年数据标注人才的需求也将到 100 万。我国各地发展的数据标注基地，也对拉动大学生就业产生了显著效果。比如，日照数据标注基地，2024 年从业人员突破 1.1 万人。国内大型互联网公司在全国数据标注基地的人员都超过了数万人。通过发展数据标注产业，能够有效缓解大学生就业压力，促进人才培养与产业发展的良性互动。

（四）发达国家纷纷布局数据标注产业，使之成为数字经济发展的基础性、战略性产业

美国、德国、英国、印度等国家和地区纷纷将数据标注产业作为数字经济发展的基础性、战略性产业，支持数据标注产业发展。美国是数据标注产业发展的领航者，通过政府战略、企业创新、产业领先等方面，全方位推动了数据标注服务的高质量发展。特别是前沿技术的深度融合应用上。欧盟则通过构建“欧洲共同数据空间”和推动数据相关法案，整合了工业、绿色协议、移动、卫生、金融、能源、农业、公共行政、技能等多个战略性行业和领域的数据，试图构建欧洲统一数据市场并大力发展数据标注产业。印度凭借庞大的低成本数字人才优势，吸引全球科技巨头投资，大力发展数据标注产业，成为全球重要的数据标注基地。

二、推动我国数据标注产业发展的思路与重点

一是加快发展一批数据标注基地，培育壮大市场主体。今年 5 月，国家数据局公布首批全国数据标注基地建设城市名单，通过推动数据标注基地建设，以点带面大力发展数据标注产业，激发行业发展活力。以国家数据标注基地试点为抓手，围绕技术创新、行业赋能、生态培育、标注应用、人才就业和安全发展等六个方面做好试点，探索推进数据标注产业发展的最优路径。另外，通过加大财税政策支持力度、完善数据标注基础设施建设、制定行业标准规范、鼓励技术创新与应用、强化人才培养与引进等措施的实施，培育和壮大市场主体。

二是鼓励数据标注核心技术研发，提升数据标注水平。通过国家重

点研发专项，加大对人工智能技术在标注技术研发投入。加快数据标注技术研发费用纳入研发加计扣除，鼓励企业加快科技创新投入。重点布局跨语言、跨领域、跨模态语义对齐、4D 标注、大模型标注、数据合成等标注核心技术攻关项目。支持研发专家标注、多模态标注、众包标注、标注审查、质量评估等智能化高端化工具，提升数据标注科技水平。

三是围绕数据标注关键环节，制定数据标注标准。构建数据标注标准框架体系，加快制定数据标注的国家标准和行业标准，明确标注质量、流程、技术等方面的要求。建立健全数据标注标准实施与监督机制，促进协同创新。加强对数据标注企业和项目的监督和检查。

四是加大数据标注人才培养，提升人才专业化素养。加强职业教育和培训体系建设，支持高校和职业院校开设数据标注相关专业和课程。推动企业与教育机构合作，共建实训基地和培训中心，培养具备专业技能和职业素养的数据标注人才。制定吸引和留住高层次人才的政策措施，促进数据标注领域的国际交流与人才合作。^[5]

（文章来源：国家数据局公众号）