加快建设高质量数据集 推动人工智能赋能行业发展

文◆国家数据发展研究院院长　胡坚波

随着 DeepSeek R1 系列模型的发布，国内掀起新一轮人工智能的热潮，通信、互联网、汽车、能源、金融、医疗、科技等龙头企业纷纷宣布接入 DeepSeek，人工智能大模型加速向各行各业渗透。人工智能大模型的发展需要“数据粮食”，特别是高质量数据集。高质量数据集是人工智能大模型训练、推理和验证的关键基础，是按照特定标准，经过采集、清洗、归类和标注等智能化处理，具有相应更新和维护机制的数据集合。

一、建设人工智能高质量数据集的重大意义

高质量数据集建设有利于推进“人工智能 +”行动落地见效，对数字经济乃至整个经济社会高质量发展具有重大意义。

一是高质量数据集建设是人工智能发展的客观需要。人工智能大模型对数据集提出了新要求，数据集的质量影响人工智能的智商。高质量数据集是人工智能真正的“护城河”，人工智能大模型的训练和推理高度依赖高质量数据集的供给。谷歌研究发现，对于图像生成模型，当计算资源受限时，数据集大小比模型大小更加重要。纽约大学的一项研究表明，大型语言模型在使用医学数据训练过程中，即使含有 0.001% 的错误信息，也可能导致模型输出不准确的医学答案。

二是高质量数据集建设是促进我国行业数字化转型的关键举措。通过开放公共数据和促进企业数据流通，可以提升垂直领域人工智能大模型的能力，促进传统行业数字化转型升级。例如，“苏州丝绸纹样数据集”汇聚了 222 件等级文物和 7012 片近现代丝绸样本的高清数据，形成了 3 个具有较高水平的高清采集纹样数据子集，并依托市场化机制，在丝绸纺织、网络游戏、汽车、美妆、银行、贵金属等领域累计授权使用 31 批次，赋能苏州丝绸文化传承、文旅消费和数字创新。

三是高质量数据集建设是促进各地数字经济发展的重要抓手。近两

年，为促进数字经济发展，国内不同地区陆续推出各类“大模型＋数据集＋算力”一体化创新基地。例如，上海徐汇区的“模速空间”、北京石景山区的大模型“超级工厂”、济南市的“大模型创新工厂”、呼和浩特市的大模型训练基地等，为大模型训练推理提供了重要支撑，助力地方招商引资发展数字经济。

二、建设人工智能高质量数据集的目标和策略

人工智能正从以模型为中心，逐步转变为以数据为中心。高质量数据集的重要性正成为社会各界的共识，大模型发展进入多模态融合阶段，应全面打造大规模、多模态、多领域的高质量数据集，建立人工智能发展新范式。建设人工智能高质量数据集应采用如下策略。

首先，坚持场景化推动高质量数据集建设。当前，人工智能应用持续走深向实，在医疗、教育、零售、金融、制造、能源等领域实现了初步应用。建设高质量数据集不能盲目跟风、重复建设，不能仅限于将公共数据简单开放，应以终为始，从医疗、教育等重点行业入手，优先突破人工智能应用最迫切、最容易产生效果、最影响行业高质量发展的数据集建设。

其次，坚持体系化牵引高质量数据集建设。高质量数据分布在各行各业，离散性强，需要更好发挥政府作用，把行业企业、模型企业、数据企业、数字化解决方案提供商、数据交易机构等多方主体组织起来，打造数据、模型、算力等协调联动生态体系，探索新模式。

再次，坚持多元化促进大中小企业融通创新。在建设高质量数据集过程中，需要加强引导技术能力强、行业影响力高、产业链资源整合能力强的企业，依托行业领域应用，多渠道吸纳、聚合相关数据。强化中小企业产业链和生态系统意识，主动融入大企业、大项目，发挥出“船小好调头”、创新干劲足的优势，不断对数据进行深加工，形成本行业、本领域的高质量数据集。

最后，坚持安全合规为高质量数据集建设保驾护航。高质量数据集建设工程涵盖数据采集、预处理、标注、合成、质量评估、开放共享等全生命周期，不仅需要保证数据的数量、质量和多样性，更要确保数据来源的合法性、合规性和产权保护等，降低数据使用中的风险。

三、从六个方面推动人工智能高质量数据集建设

高质量数据集是决定人工智能大模型性能优劣的关键所在。为全力打造人工智能高质量数据集，推动大模型应用迈向新高度，建议从高质量数据集图谱构建、政策法规保障、建设指引制定、评测体系建设、跨域合作拓展、标杆牵引示范等方面着手，推动高质量数据集建设迈上新台阶。

一是以服务大模型应用为核心，绘制高质量数据集建设图谱，明确“建什么”。围绕应用需求牵引、典型场景切入、行业领域赋能、安全风险可控等维度，调动政、产、学、研、用各方力量，梳理高质量数据集典型场景和应用需求等，绘制高质量数据集建设图谱，实现可查询、

可下载、可应用，全面助力大规模、多模态的高质量数据集建设。

二是以保障数据集建设为目标，协同推进政策法规的制定与完善，确定“依据在哪”。在政策层面，推动各部门出台针对性政策，强化高质量数据集供给。鼓励企业积极参与高质量数据集建设，对在数据采集、清洗、标注等环节投入较大的企业给予政策支持，降低企业建设数据集的成本。在法规层面，需加快明确数据权属问题，界定数据生产者、持有者、使用者和经营者的权利与义务，保障数据在合法合规的框架内流通与使用，为高质量数据集建设营造良好的政策法规环境，促进整个行业的健康可持续发展。

三是以解决现实问题为导向，制定高质量数据集建设指引，指明“怎么建”。组织跨行业交流，分享高质量数据集建设经验及面临的问题，总结建设方法论和问题库。针对问题库，以“揭榜挂帅”方式征集解决方案。在广泛调研和总结基础上，制定发布高质量数据集建设指引，不断优化建设方案和路径。发挥人工智能技术优势，对大量文本、图像、音频等数据进行自动标注和分类，批量构建高质量数据集。

四是以推动标准建设为牵引，打造高质量数据集评测体系，指导“怎么评”。一方面，通过对高质量数据集的格式规范、类型、质量要求等方面的研究，开展系列标准的研制及细化，为各行业领域在数据采集、标注、加工治理、应用推广等提供标准化规范指引。另一方面，构建涵盖细分行业的高质量数据集质量评测方法、评测工具集。通过规范化的高质量数据集评测工具，客观地评判数据集的质量等级和价值曲线，结合应用需求不断进行迭代升级。

五是以探索跨域合作为重点，建立高质量数据集流通利用新机制，阐明“怎么流通”。依托可信数据空间、数场、数联网、数据元件等实践方案，推动医疗、交通、气象、社保等多领域高质量数据集在安全合规框架内有序流动，注重建设跨部门、跨行业、跨地区高质量数据集。运用区块链、隐私保护计算等技术实现数据集的可溯源与安全保护，促进跨域数据集交易流通，形成典型案例，催生新应用、新模式，释放数据要素乘数效应。

六是以行业标杆示范为牵引，发挥资金“风向标”作用，解决“用什么引导”。组织开展行业领域高质量数据集征集工作，鼓励各行业、各地区的企业积极参与，形成各类高质量数据集库，提高整体供给水平、供给规模。鼓励各类资金支持高质量数据集建设，持续完善建设机制，积极推广典型案例，全面助力人工智能赋能行业高质量发展。

（文章来源：国家数据局微信公众号）