深读

浙江启动上百个高质量数据集重点项目建设

“炼化”数据,给AI造燃料库

■ 本报记者 王逸群 王艳琼

散落的医院病历、沉默的工厂生产记录、碎片化的设备运行参数……这些藏在各行各业系统里的数据,看似普通零散,却是人工智能迭代升级的珍贵原料。经过清洗、筛选、标注等一道道“炼化”工序后,杂乱的原始素材变为规范、可用、可建模的优质数据资源。

这,就是支撑AI落地产业的高质量数据集。

当前,AI竞赛越来越激烈,算力、算法日益趋同,数据质量成为决胜关键。近日,国家数据局出台专项方案,全面部署行业高质量数据集建设,为数据赋能人工智能发展划定清晰路径。而在浙江,数据集建设早已启幕。全省已布局落地109个重点项目,并持续挖掘、推广数据赋能典型场景,不断探索数据从资源变资产、从要素变产能的转化路径。

高质量数据集究竟是何模样?如何填平AI“语义鸿沟”、为产业升级赋能?浙江又如何跑出数据要素发展加速度?近日,记者前往杭州、温州等地调研。

AI迭代升级的“营养餐”

高质量数据集长啥样?在杭州数据交易所,老板电器上架的数据产品,给出了最直观的答案——

研发数据、客服数据、设备运维数据、菜谱数据……一张张清单覆盖厨房全场景。从食材特性、热量数值,到烹饪用水量、食材损耗、减盐效果,再到营养结构分析、厨电运行规律与菜品成熟度的对应关系,整套数据完整还原了家庭烹饪全过程。

它集合文档、表格、实拍图片、专业报告等多种形态,将烹饪过程中的细微变化、真实场景、实操规律等,全方位记录下来。更关键的是,这些数据经过人工筛选、精细标注、标准化治理,剔除了杂乱无效的信息,只沉淀高密度、高价值内容,这也是它区别于普通数据的核心特质。

如今,这批高质量数据集已全面赋能企业智能体系建设,成为各类AI智能体的核心底座。“比如面向消费者的食神大模型,能够提供膳食规划、烹饪指导、智能菜谱推荐等服务。”老板电器相关负责人说。

过去,很多人觉得AI“答非所问、脱离实际”,这便是业内所说的“语义鸿沟”。由于训练数据鱼龙混杂,细分场景数据缺失,AI缺少扎实有效的数据支撑,只能凭空拼凑内容,导致话语看似通顺,实则漏洞百出。

高质量数据集,就像AI迭代升级的“营养餐”。精准、规范、贴合场景的优质数据,能够持续为大模型深度学习、能力迭代赋能。

“用户需求越来越个性化、精细化,传统产品和服务模式难以精准匹配市场诉求。”老板电器相关负责人介绍,靠着这套数据集支撑,大模型读懂烹饪逻辑、摸透用户需求,让产品更贴合市场。2025年,老板电器的AI数字厨电全年销售额23.5亿元,同比增长36%。

放眼全国,数据要素建设持续提速。截至2026年一季度,全国高质量数据集突破11.6万个,总体量超960PB(计算机存储单位拍字节)。近年来,浙江分类施策推进公共数据、企业数据、个人数据合规高效开发利用,累计向社会公众无条件开放可机读的公共数据集4.4万个。

与此同时,供需之间的缺口依然突出。通用大模型已难以满足细分行业需求,医疗、金融、制造、政务等领域,急需具备高知识密度、高应用价值的行业专识数据集。这也是全国大力推进数据集建设的重要原因之一。

抢抓数据发展机遇,浙江从顶层设计出发,将高质量数据集建设作为数据要素改革的重点任务,推动数据要素和人工智能深度融合。此前,省政府已制定出台人工智能“一揽子”政策体系,支持企业建设、利用高质量数据集,打造高价值应用场景。省数据局联动发改、经信等行业主管部门,定期遴选发布一批“以数赋智”标杆应用。如浙江同盾科技建设的金融反诈高质量数据集,训练金融风控大模型,已服务金融机构拦截涉诈资金达千万级;省交通集团建设高速公路运营数据集,训练“之江慧眼”视觉模型,有效降低高速公路二次事故率,浙江经验已推广至山东、新疆、宁夏、安徽等地……

今年6月,国家数据局推出六大专项行动,打通数据生产、治理提质、场景应用、流通赋能、价值释放的全链条闭环。眼下,浙江除了“拼算力、拼算法”,也加速迈进“拼数据质量、拼场景精度”的新赛道。

从“深海”里打捞数据

温州医科大学附属眼视光医院,是国内最早深耕眼科高质量数据集建设的机构之一。在这里,眼视光医院国家临床医学研究中心主任医师李明给记者还原了打造数据集的过程。

眼科检查与影像设备品类繁杂,全院设备超20种,各厂商接口标准互不统一,多数国外品牌更是完全封闭数据协议。“过去每台设备只能打印纸质报告,想要实现数据全面电子化,根本无从下手。”李明说,这意味着,一台设备就是一座信息孤岛,全院医疗数据各自割裂、互不连通。

团队做了一个“笨”决定:一台一台去攻克。能开放端口的就对接,不能开放的就想办法提取数据。整整5年,他们完成上百台设备的系统接入,打通所有系统数据壁垒,实现全域互联互通。最终建成一座跨越26年、涵盖超2200万份病历、85亿条数据、1300万份影像的大型眼科数据库。

但这只是开始。从“深海”里打上来的数据,还是“浑浊”的。“不同医生病历书写习惯不一、表述各异,零散的文本内容,无法直接用于模型训练和科研分析。”李明介绍。为此,他们针对每一个疾病建立标准字段集,逐条拆解并提取有效信息,对临床病历和检查进行标注,按照法律法规要求对数据进行脱敏和匿名化处理,将零散的文字、数字,转化为规范统一、可检索、可计算的结构化数据。最终这套国内顶尖的眼科高质量数据集才正式成型。

从中可以看出,珍贵的数据,都是从真实产业全链路中一点一滴沉淀而来的。

事实上,打造这类高精度行业数据集,难度极高:一方面,产业数据分散在各行各业的车间、设备之中,格式杂乱、标准各异,对采集技术、人力成本、专业能力都提出极高要求;另一方面,数据标注规则不统一,也导致大量数据难以复用、落地受阻。

“专业化、场景化、合规化,是高质量数据的核心特质。”一名数据要素研究专家告诉记者,标准缺失、重复建设、质量参差,是长期困扰数据行业的共性乱象。想要降低数据治理成本、破除行业发展门槛,关键在于告别“各自为战”,建立统一、通用、可落地的行业标准体系。

瞄准行业痛点,浙江在19个市县布局建设高端数据标注基地,政企协同开展智能标注技术攻关,培育数据标注龙头企业。

杭州景联文科技有限公司就是其中一家标杆企业。企业副总裁林旭峰介绍,他们依托一套成熟完备的标准化作业体系。在采集端,研发团队深入家居、工业、医疗等12大类真实场景,通过真机遥操作与人机协同采集,生成机器人专属训练数据。在数据提纯中,企业再利用自研平台,把格式混乱的数据统一对齐。最后,将杂乱原料变成干净标准的结构化数据,直接喂给AI模型学习。

当前,国内数据集建设仍面临汇聚产量不足、供给质量不均、利用效率偏低等难题。破解发展瓶颈,浙江构建起“数据供给—模型训练—场景应用—产业赋能”的产业生态。

今年3月,海纳数据枢纽创新中心(杭州)项目建设正式启动。这是之江实验室打造的一站式AI创新赋能平台,整合算力、数据、模型三大核心要素,为AI创业企业、科研团队提供全链条服务。数据素材采集、AI模型训练、智能化产品研发……未来,各类主体能在此一站式完成全流程操作。

这种“集中力量办大事”的模式,将数据采集的周期大大缩短,也为具身智能企业提供了一条快速获取海量高质量数据的新路径。

温州数据集团团队走进企业,收集调研高质量数据集。 温州数据集团供图

落脚点是赋能产业

“建设高质量数据集,落脚点是赋能产业、催生新质生产力。”省数据局相关负责人表示,唤醒沉睡的海量数据资源,最终是要实现从数据资源到可量化资产、AI技术到实体产能的转变。

据《2025年浙江省知识产权发展与保护状况》白皮书,截至2025年,浙江数据知识产权累计登记3.04万件,直接运用金额137.14亿元,浙江数据知识产权改革领跑全国。

数据不再是抽象概念,已成为浙江数字经济的新增长引擎。

最直观的变化,来自企业心态的反转。温州市数据集团市场运营部负责人,一年多来一直在跑企业、摸需求。他每周走访两三家企事业单位,从龙头企业到小微商户,从政务单位到医药公司,覆盖各行各业。他们发现,今年以来越来越多中小企业主动提出需求,想要整理盘活自有数据。

“传统市场增长放缓,企业急需新的突破口。”该负责人说,生产线上的设备参数、工艺数据等信息,不只是用来监控生产、保障运转,还可以作为企业资产出售。

当地一家传统包装企业,通过数据清洗、标准标注、合规脱敏后,把沉淀多年的生产、工艺、供应链数据梳理成型,销售给下游客户。不过,数据“变现”的路并不平坦。从普通数据变成优质资产,要跨过治理、合规、权属、定价等一道道门槛。“眼下,我们就遇到新的行业痛点。”温州市数安港管理服务中心相关负责人说,比如数据作为一种新的资产,评估难、定价标准不一。

“数据没有固定价,关键看需求。”林旭峰坦言,同一份数据适配不同场景、服务不同企业,价格差距最高可达数十倍。

为破解定价难题,温州依托数安港,联合上海财经大学打造全国首个数据价值智能定价体“数衡器1.0”。它不再模糊评估数据价值,而是依托数十项核心指标,精准测算数据质量、场景适配度和时效价值,把说不清、道不明的数据价值,变成可解释、可交易、可落地的公允价格。

想要让数据顺畅流动、持续变现,离不开靠谱的平台和专业的市场主体。浙江筑牢交易“主阵地”,持续做强杭州数据交易所等核心平台,目前已上架近6000个各类数据产品,与此同时,还打通了长三角数据互通互认壁垒,让浙江数据走出地域限制,在更大范围流转赋能。

应用场景也至关重要。立足产业特色,浙江还借力龙头企业,深耕细分赛道。依托阿里云、宁波航交所等行业标杆,一批金融、医疗、港航物流等垂直领域数据平台,让各行各业都有适配自身场景的专属数据服务。

2026年被定义为“数据要素价值释放年”。从实验室到生产线,从后台系统到产业一线,沉睡的数据被激活。无形的数据流,正在转化为可落地、可增值、可赋能的新质生产力,成为浙江数字经济高质量发展的新动能。

❶ 老板电器AI烹饪大模型。 受访者供图
❷ 2026数据安全发展大会现场。 受访者供图
❸ 杭州景联文科技有限公司数据标注覆盖场景。 受访者供图
❹ 中国(温州)数安港数安大厦。 共享联盟·瓯海 王斌 摄
❺ 国内首个基于开源鸿蒙底座的人形机器人专业数据训练平台在慈溪启用,40多台机器人在工作人员的引导下积累训练数据。 本报记者 贺元凯 通讯员 杨奇星 摄

专家观点

政府需发挥元治理作用

■ 谈婕

人工智能产业范式正经历从“以模型算法为中心”向“数据协同发展”的深刻跃迁。优质数据既能为大模型筑牢合规可靠的安全底座,也可为算法研发与持续优化注入效能动力。有研究预测,到2028年大语言模型或将耗尽互联网高质量文本数据。把握这一数据机遇期,正成为决定通用与垂直大模型竞争格局的核心变量。

尽管如此,高质量数据集兼具高增值、高复用与高效益特征,其显著的正外部性决定了单靠市场机制难以实现充分供给。应对市场自发配置局限,政府应扮演元治理角色,承担起高质量数据集的供给、倡导与规则制定等多重职能。

首先,政府是高质量数据集的直接供给者。大量真实行政数据、公共数据仍沉淀于体制内,政府应率先示范,打通数据壁垒、有序促进数据开放;其次,政府是高质量数据建设倡导者。政府既要“以评促建”,为筛选机制注入公共价值,引导社会力量参与,也要建立成本补偿机制,让建设者获得合理收益;最后,政府是数据治理规则的制定者。政府仍需进一步明确数据确权、隐私保护、利益分配等底层规则,推动高质量数据集建设从“盆景”转化为“风景”。

元治理并非政府包办一切。高质量数据集建设尚在起步期,不少项目仍停留在概念验证阶段,尚未形成市场造血机制。元治理者为其扶上马、送一程,最终仍应让市场机制发挥决定性作用,使应用场景与数据价值实现良性循环。

(作者系浙江大学公共管理学院特聘副研究员)

中国新闻名专栏 浙读