高质量数据集建设“提要求” 数据标注产业发展逻辑如何变？

中经记者 许璐 李晖 北京报道

数据标注行业的生产方式正在发生变化。

近日，国家数据局印发《关于推进行业高质量数据集建设行动的实施方案》（以下简称《实施方案》）。在标注环节，《实施方案》提出，发展“模型预标注+人工校准”“人工标注+模型检验”“模型预标注+模型检验”等智能化标注服务。推动形成“人机协同、专家深度参与”的多层次标注模式。梯次布局数据标注创新试验区。培育一批数据标注龙头企业、独角兽企业、瞪羚企业等。

《中国经营报》记者梳理海天瑞声（688787.SH）、世纪恒通（301428.SZ）、数据堂（831428.NQ）公开信息发现，模型预标注、辅助标注、质量检验等技术已开始进入数据生产流程，企业的业务范围也向大模型训练、自动驾驶、多模态和行业专业数据延伸。

深度科技研究院院长张孝荣在接受《中国经营报》记者采访时表示，数据标注行业的商业模式正在从“卖劳力”到“卖资产”转变，即不再按数据量“一口价”卖数据，而是转向卖API调用、卖全栈解决方案，甚至探索“Token（词元）交易”和数据订阅制。数据服务商与客户之间的关系也将由外包服务逐步转向长期协作。

从人工逐条处理转向人机协同

海天瑞声透露：“随业务场景动态浮动。完全人工标注占10%—30%，多用于无适配预训练模型的全新长尾赛道；人机协同模式占50%—70%，为行业主流。”

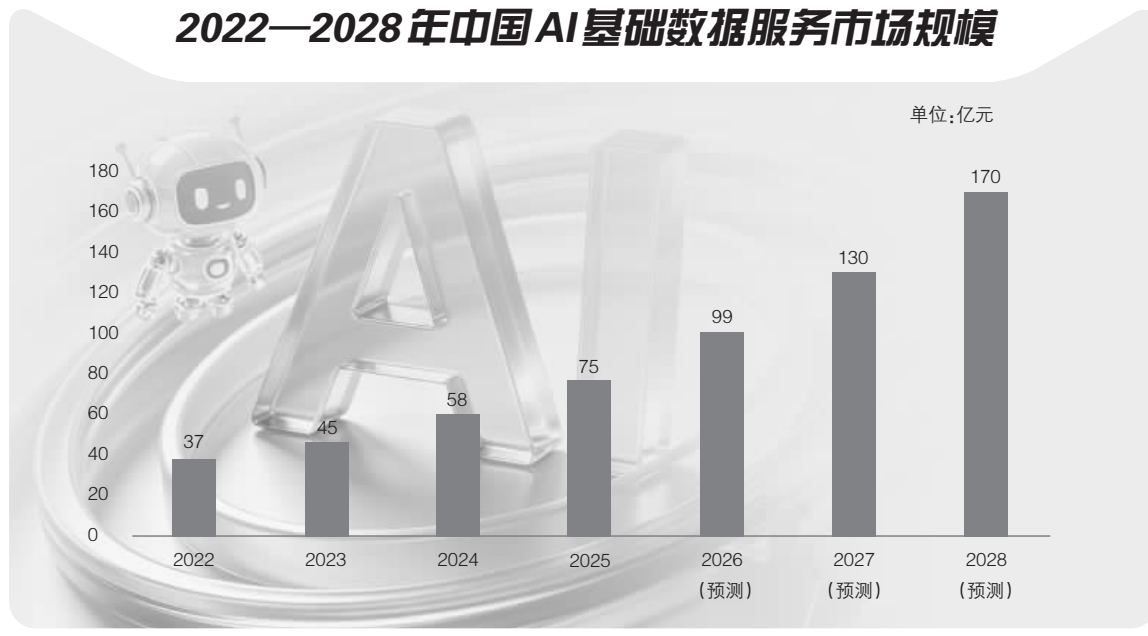
高质量数据集建设提速，进一步扩大了数据采集、清洗、标注和质量检验等环节的需求。国家数据局发布的《全国数据资源调查报告（2025年）》显示，2025年，全国高质量数据集数量超过11万个、规模超过908PB，同比分别增长61.13%和142.58%。

在数据集规模扩大、产业政策持续落地的同时，数据标注的生产方式也在发生变化。当前，数据标注正在从人工逐条处理，转向模型批量处理与人工重点复核相结合。公开信息显示，2025年全球数据标注解决方案与服务市场规模达204.1亿美元，复合年增长率达24.5%。

作为数据标注产业链的代表性企业，海天瑞声方面在接受记者采访时表示，行业整体正从传统劳动密集型作业模式，全面转向以智能化人机协同为核心的现代化数据生产体系。

据海天瑞声方面介绍，目前公司已覆盖《实施方案》提出的三类智能化标注服务，各模式分属不同应用阶段。

其中，“模型预标注+人工校准”是现阶段规模化落地最成熟的主力生产模式。实际作业中，系统会同步调用2—3个差异化预训练模型独立输出标签并交叉融合，再通过主动学习筛选模型分歧、低置信度及识别偏差样本，由人工校准；其余无争议、高置信度样本直接免检。



数据来源：艾瑞咨询 刘洋/制图

“人工标注 + 模型检验”模式更适配医疗影像、金融文本、法律文书等高知识密度、强合规约束赛道。作业流程中先由专业专家或资深标注人员完成带有专业判断的基础标注，再由 AI 模型后置开展一致性核查，自动捕捉错标、漏标、逻辑异常等问题数据，弥补人工标注标准不统一、细节疏漏等短板，也是尚无成熟行业预训练模型的新兴专业赛道起步阶段的核心方案。

“模型预标注+模型检验”模式，更多是在一些客户侧自有数据体系或特定工具链中使用，海天瑞声的角色更多是围绕高质量数据集构建，提供更完整的人机协同生产能力。

对于海天瑞声在业务中三类模式的占比，公司方面透露：“随业务场景动态浮动。完全人工标注占10%—30%，多用于无适配预训练模型的全新长尾赛道；人机协同模式占50%—70%，为行业主流；自动化参与较深的流程，主要体现在部分标准化、结构化任务中，以模型预标注与辅助质检为主，但通常仍需人工参与关键校验与收敛。”

“场景越简单、标准化程度越高，自动化的介入程度就越深。”张孝荣认为。

针对不同类型数据的自动化程度和技术难点，海天瑞声方面称，通用图文自动化程度最高，难点集中在遮挡、小目标、多语义歧义；自动驾驶点云自动化中等，难点为极端天气噪点、微小障碍物、多传感器时空对齐；具身智能自动化程度最低，行业标准尚未统一，时序动作、空间匹配高度依赖人工与专家；多模态数据自动化中等，核心痛点是跨模态时序对齐、语义一致性校验，语义冲突样本均需人工校准。

记者注意到，除海天瑞声外，其他企业也在搭建标注平台和生产体系。世纪恒通2025年年报显示，其数据标注业务覆盖文本、图片、音频、视频、直播等数据形态，并依托太阳湖数据标注基地形成“标注师+标注平台+标注作业基地”的业务布局。

数据服务向专业化和持续化延伸

数据标注行业正在告别纯靠“堆人力”的劳动密集型模式，全面进入了人机协同的新阶段。现在的主流玩法是“机器打底，人工把关”。

数据标注生产方式变化的同时，企业的服务内容也开始向知识密集型场景延伸。

随着人工智能应用向多模态、智能体、自动驾驶和具身智能等场景拓展，数据标注开始涉及多轮对话、复杂推理、工具调用、环境感知、任务规划和运动控制。金融、医疗、法律、工业等专业数据，还需要相应的行业知识。

海天瑞声方面回复称，公司当前重点布局三类高知识密度标注场景：一是STEM、金融、医疗、法律等专业大模型文本语料；二是高阶自动驾驶、工业机械臂、具身机器人等复杂感知决策数据；三是医疗影像、车载交互、跨语言对话等多模态专家校验数据集。

财报显示，海天瑞声的训练数据生产过程主要包括四个环节：设计（训练数据集结构设计）、采集（获取原料数据）、加工（数据标注）及质检（各环节数据质量、加工质量检测）。2025年海天瑞声实现营收3.77亿元，同比增长59%。

张孝荣认为，数据标注企业的竞争能力主要体现在技术平台、行业知识和安全合规三个方面。技术平台得有自研的智能标注工具，靠“AI打底+人工把关”实现降本增效。行业知识要求企业必须懂行（如医疗、自动驾驶），能调动行业专家为数据注入专业知识。安全合规则覆盖数据采集、存储、处理和交付流程。

不同企业也在形成不同业务模式。例如，数据堂采用版权数据集授权与定制化数据服务并行的方式，其官网显示，公司拥有1500余个版权数据集，覆盖200余种语言和方言。2025年数据堂实现营业收入3.62亿元，同比增长49.20%。

世纪恒通将重心从基础数据服务向前沿AI产品延伸，重点投入AIGC文创Agent及AI-Hub等产品的研发与推广。根据公司2025年年报，数据标注已成为商务流程服务的核心增长方向，当年商务流程服务实现营收2.65亿元，同比增长10.67%，占总营收比重为25.06%。

在张孝荣看来，数据标注行业正在告别纯靠“堆人力”的劳动密集型模式，全面进入了人机协同的新阶段。现在的主流玩法是“机器打底，人工把关”。这种模式不仅让效率翻了倍，也让标注质量有了保障，整个行业正在向智能化和平台工程化转型。

针对行业智能化转型存在的痛点，海天瑞声方面指出，技术工具上，模型跨场景泛化能力不足，多模态工具链割裂；质量标准上，缺少全国统一、跨企业互认的量化测评体系；专业人才上，兼具行业知识与AI标注能力的复合型人才稀缺，专家留存成本高；数据安全上，金融、医疗、车企敏感数据全流程合规管控成本持续走高；商业回报上，基础标注低价竞争，智能化研发投入周期长，数据资产化交易模式尚未普及。

根据艾瑞咨询的数据，2024年中国人工智能基础数据服务市场规模为58亿元，2028年规模将达到170亿元，年复合增长率为30.84%。

记者注意到，《实施方案》同时提出，发展专家型数据标注服务，建立行业专家认证机制，推动专家深度参与指令微调、强化学习等阶段的专业知识标注。

针对专家参与机制，海天瑞声方面回复称，公司已搭建覆盖30余个行业、规模超过5万人的全球专家协同网络，并依托DOTS平台实现智能派单、线上评审与全流程溯源，形成常态化专家协同体系。此外，公司对专家实施分级认证和动态考核，根据项目难度匹配不同层级专家，并在项目初期专家参与制定标注规则和疑难样本判断标准。平台筛选出的高专业风险样本自动流转专家仲裁，修正后的样本再用于垂直领域模型训练。

点咖啡送Token券 词元经济成多地招商引资新引擎

中经记者 石健 北京报道

今年以来，多地将词元经济作为招商引资新引擎，通过“政策＋词元”新模式，吸引外地企业落户本地投资兴业。在多位地方政府招商部门负责人看来，中西部地区地方政府通过算力、词元资源以及政策支持，将会承接京津冀、长三角、粤港澳大湾区数字产业的外溢产能，创新招商引资新模式。

《中国经营报》记者注意到，目前，已经有武汉、贵阳、庆阳等地推出优惠政策，吸引人工智能大模型企业进驻本地。

武汉市经信局人工智能产业处有关负责人对记者表示：“词元贯穿AI‘能源—芯片—基础设施—模型—应用’五层蛋糕，是AI时代的‘石油’和‘电力’。在AI产业的飞速发展期，所有人都是站在同一条起跑线上，谁早、谁快、谁主动，谁就最有可能抓住新的机遇。”

抢滩词元经济

“点一杯咖啡，就送Token券。”在武汉经开区南太子湖创新谷，一位智能体行业创业者已经拿到了20万Token（词元）的免费算力。

前不久，国家数据局正式将词元定义为“智能时代的价值锚点、连接技术供给与商业需求的结算单位”。与此同时，词元经济正成为武汉招商引资的新引擎。

记者注意到，前不久，武汉江岸区召开Token（词元）经济大会，参会方除了本地企业外，还吸引了大批来自京津冀、长三角、粤港澳大湾区的大模型头部企业。

在大湾区科技创新服务中心相关负责人看来，江岸区聚合全国资源要素，率先提出打造词元经济推动转型发展，这是江岸区抓住武汉全力打造“五个中心”、全面建设现代化大武汉的历史机遇，围绕AI不断塑造发展新动能新优势的重要举措。

在摩尔线程（688795.SH）相关负责人看来：“城市产业的智能化转型已是必然趋势。武汉在产业配套、应用场景开放、政策扶持力度等方面已经具备先行优势。未来，公司将参与江岸区词元经济建设，在算力供给和赋能上贡献力量。”

今年以来，江岸区政府发布了《江岸区大力支持Token经济发展的若干措施》，计划每年拿出共计5000万元专项资金，用于支持词元经济发展，包含每年1000万元词元券、每年1000万元算力券、每年500万元模型券、每年500万元数据券、每年1000万元场景资金、每年500万元资金支持线上词元服务平台建设和每年500万元资金支持线下OPC生态社区建设。

在多位地方政府招商部门负责人看来，如果说“东数西算”掀起东西部数算协同的合作新范式，那么随着词元经济发展，“绿电—储能—算力—词元”必将形成闭环协同模式。中西部地区谁抢抓承接京津冀、长三角、粤港澳大湾区数字产业产能外溢机遇，谁就能在招商引资实现新突破。

从产业到生态

今年4月，中央政治局会议强调，全面实施“人工智能+”行动，发展智能经济新形态。国家数据局随即将“词元经济”纳入工作体系。从“东数西算”工程到全国一体化算力网建设，从高质量数据集建设到数据要素市场化改革，顶层设计层层递进，环环相扣。

与此同时，词元经济版图正在形成。京津冀依托科研资源成为技术创新高地，长三角凭借制造业基础成为工业词元示范区，成渝地区利用绿电成本优势成为算力承载地。

记者注意到，地方政府依托词元经济创新招商引资形式，逐步探索三条创新路径：

一是打造能源算力型城市，通过“词元工厂”产生虹吸效应。如甘肃庆阳、内蒙古包头等地，这些城市依托风光电资源形成低成本的算力，打造“词元工厂”，精准招商算力、人工智能大模型企业。二是打造词元生态型城市，依托“政策+空间”双轮驱动激活词元生态。除了前述提到的武汉外，广州也以“政策红利+优质空间”为抓手，构建词元产业生态。三是打造数据场景型城市，依靠多年对数据要素市场深耕，以场景搭建词元经济。如贵州贵阳、广东汕头等城市，坐拥数据资源聚焦词元数据服务、出海应用，打造特色产业集群。

庆阳数据局相关负责人介绍称：“庆阳构建一体化招商机制，累计对接数字经济企业8126户，签约1760户，落地571户，吸引凌穹瞬联（庆阳）数据科技有限公司、首都在线（300846.SZ）等龙头集聚，成为Kimi大模型核心算力供给地。”

今年以来，随着“包你满意，包你放心”营商口号喊出，内蒙古包头市吸引了一批外地企业前来考察。

4月20日，“国内首个园区级词元基地项目”在包头市固阳县实现签约。在运营与产品端，由内蒙古数通国联科技有限公司负责园区词元基地整体运营，浪潮通信信息系统有限公司负责词元产品线的建设与运维。在算力与调度端，中国算力网西部运营中心（四川浮点运算科技有限公司）及其浮点算力圈平台，将接入中国算力网调度平台，通过算力调度与纳管，构建多元异构一体化算力底座，链接模型训练及推理环节。在能源与服务端，北方合创作为绿色能源交易平台，将赋能包头金山开发区，打造集园区级算电协同、Token（词元）生产、调度交付于一体的一站式服务。

前不久，贵州贵安词元经济发展推进大会提出，自4月1日起，面向全国发放1.4亿元算力券，符合条件的企业最高可享30%用算补贴。从算力、数据、模型、业态、场景五大维度协同发力，推动已投用智算中心扩容提质，加大智算中心招引力度。推动交通、矿产、文旅、气象、中医药等领域建设5个以上行业高质量数据集，依托贵阳大数据交易所、贵阳城市可信数据空间完善词元交易体系。

贵阳市大数据发展管理局提出招商引资新思路，梯次培育词元经济经营主体，既要招引培育龙头企业、领军企业，也要引导创新创业，孵化培育OPC公司。

值得注意的是，采访中，一些从事数据要素行业人士提示，完善地方词元经济招商引资配套举措的同时，国家层面应该加强顶层设计。目前，国家层面尚未出台统一的词元计量和定价标准，若两地缺乏协同，将导致市场分割、交易成本高，阻碍词元要素跨区域自由流动，不利于全国一体化算力网络建设与跨区域数字产业协同发展。因此，建议国家层面探索建立词元统一计量和定价标准，完善词元经济治理体系。

部分地方词元经济扶持政策汇总

地区	核心政策文件 / 行动	补贴 / 扶持品类
武汉江岸区	《江岸区大力支持 Token 经济发展的若干措施》	词元券、算力券、模型券、数据券、场景资金、线上平台建设、线下 OPC 生态社区
贵州贵安新区	词元经济发展推进大会政策	算力使用补贴
内蒙古包头固阳县	落地国内首个园区级词元基地项目	园区全链条配套扶持
甘肃庆阳	一体化数字产业招商机制	算力产业落地配套服务
广州市	“政策红利 + 优质产业空间” 双轮驱动	产业空间、落地配套政策

刘洋/制图