

# 具身智能产业发展,数据困局卡在哪里?

► 本报记者 孙庆阳

9月2日,北京人形机器人创新中心宣布,其开源数据集RoboMIND全球下载量突破2000万次,近一个月实现翻倍增长。该数据集自2025年12月开源以来,已累计释放30余万条双臂操作轨迹,覆盖700余项实操任务,面向全球开发者全量开放。此前,国家发展改革委政策研究室副主任、新闻发言人李超在新闻发布会上表示,将以具身智能实训场和应用中试基地为抓手,构建高质量真机数据采集系统,破解具身智能训练“数据饥渴”问题。

2026年被业界称为具身智能数据元年,高质量具身智能数据已成为机器人行业竞逐的关键资源。

## 数据成“硬通货”

“互联网已经发展二三十年,数据是现成的。当前,大模型可以依靠互联网上现成的文本、图片、视频训练。但是,机器人需要的物理世界数据,比如,机器人如何抓取、走路、躲避障碍,这些数据现实中很少。”创世伙伴创投创始主管合伙人周炜认为,宇树科技上市市值一度冲到2000多亿元,不少公司只有Demo(演示/示范)已经估值几十亿元。在具身智能机器人产业加速发展的过程中,谁能率先掌握高质量的数据,谁就握住了下一阶段入场券。

周炜举例说,灵巧手公司能让机器人手的灵活性做到20-30个维度,配合触觉、电子皮肤的功能,机器人手在运作中采集的数据量可能不亚于一台行走的机器人。具身智能数据的采集方式远比想象中多元,机器人数据竞赛也越来越激烈。

研究机构Interact Analysis的



在四川成都天府绛溪实验室机器人训练技术创新中心,工程师进行精细动作采集。 新华社记者 王曦/摄

报告显示,目前,中国已有至少90个人形机器人数据采集和训练中心处于使用或规划建设中。周炜表示,一个数采中心的建设需要场地、设备、操作人员、技术人员等要素,投入动辄千万元级别。数据采集和训练中心的井喷,意味着具身智能的“ChatGPT时刻”还未到来,数据竞赛已开启。可见,数据在具身智能发展中至关重要。

普渡机器人联合创始人兼CTO郭璵表示:“目前,普渡机器人在全球部署了约13万台设备,每年积累了3650万小时导航数据和1580万小时操作数据。这些数据的价值,可能比他们卖出去的机器人本身还要高。”

## 数据从哪里来

真机采集数据,最真实也最贵。“这是让机器人在真实环境

中工作,以采集数据。”一位业内人士透露,在这一过程中,采集的有效数据可能有一半被直接使用,剩下的数据仍然需要清洗、标注、优化。例如,智元开源包含100万条轨迹的数据集AgiBot World,英伟达GROOT N1模型80%的真机数据就来自数据集。但一个数采工厂需要几百人的团队、大量设备、长期的运营成本,代价是高昂的。

仿真生成数据,精度不够但便宜。谋先飞创始人兼CEO崔汉青认为,真机数据太慢太少,“我们能不能通过烧算力和电就把数据产出来”?谋先飞自研物理引擎,在仿真环境里让机器人“上学”,成本低、效率高、场景可以无限扩展。但崔汉青也承认:“仿真训练是机器人的‘虚拟学校’,我们永远做不到百分之百和真实数据一样,能做到90%、

95%,就已经很好了。”

崔汉青观察到市场共识正在转变:去年,业内有人说仿真生成数据没用,今年机器人企业在仿真环境里训练出的机器人已取得了出色的效果,英伟达等企业也在持续加码仿真生成数据。崔汉青提醒,数据不仅要有规模,分布也很重要,真实数据必须多样性,才能给模型带来本质的信息输入。

Echo数据被称为第一视角数据。崔汉青把这类数据比作金矿,这是人类拿东西、走路、工作的真实经验的集合。

“但问题在于,人类的手和机器人的手不一样,怎么让仿真生成派上用场,把人的动作翻译成机器人能用的数据?我们通过动作重定向,把人类手的数据映射到机器人手上。仿真和Echo的结合才是天作之合。”崔汉青说。

## 期待标准的统一

酷哇科技联合创始人兼COO李柯宏举了个例子:“特斯拉FSD(全自动驾驶)被业内看作是二维智能,几百万辆车在路上跑了多年,目前还未达到L4级别,数据采集仍是核心瓶颈之一。具身智能机器人被看作三维智能,它的数据采集难度更高。”

“企业热衷于数据采集的同时,面临着不容忽视的问题:目前,市场上有双足人形机器人、轮式机器人、四足机器人、固定底座机械臂,每种机器人的数据格式、传感器配置、动作逻辑都不同,数据采集中心采集的数据,另外型号的机器人可能就用不了。”银河通用创始人王鹤说。

机器人构型各种各样、硬件接口互不兼容、运动控制协议各异、数据采集与存储格式各行其道,几乎是当下人形机器人行业的普遍现状。数据难以在不同机器人之间复用,一份数据难以发挥十份的效能,也是具身智能数据短缺的一个重要因素。

王鹤认为,当数据无法跨本体流动,“数据规模”只是账面资产,而非真正的能力。从这个角度看,RoboMIND将30余万条双臂操作轨迹向全球全量开放,意义在于为行业提供一套可对照、可检验的“标准答案”。

京东工业工业自动化事业部负责人冯天霖判断,当前,机器人及其零部件的技术路线复杂多样,谁能帮行业把标准树立起来,谁就能占据产业链主导地位。当不同本体、不同场景的数据都能在统一标准下翻译、复用,机器人产业才算真正从“教会一台机器人干活”走向“教会整个产业干活”。

# 海光信息首发 AI 算力创新架构

本报讯(记者 张伟)“当前,智能体(Agent)正成为人工智能(AI)落地真实业务的核心抓手,而Agent对传统CPU(中央处理器)有着更高、更新的需求。”近日,在2026中国国际大数据产业博览会上,海光信息技术股份有限公司(以下简称“海光信息”)首发“Agent to Token开放计算架构”,依托CPU与DCU(数据处理单元)双芯协同及开放互连,为词元(Token)经济规模化发展提供从芯片到应用的全新算力框架。

“Agent的全面普及,直接改写了数据中心的算力配比规则。”海光信息产品部总经理张攀勇介绍,

词元作为连接业务需求、算力投入与服务产出的关键载体,其生产效率、单位成本与供给稳定性,直接决定Agent应用深度与词元经济商业化落地进程。针对这一趋势,海光信息首发“Agent to Token开放计算架构”,旨在面向词元经济完整价值周期,打通“业务需求—智能体调度—词元生产—结果交付”闭环,精准匹配Agentic AI负载特征。

据介绍,该架构通过“CPU+DCU”双芯及开放互连体系,将Agent调度与词元生成纳入统一算力架构,为词元规模生产、成本核算和行业应用提供系统级

计算支撑。其中,CPU承担Agent业务流的系统枢纽职能,覆盖数据预处理、多智能体沙箱隔离、任务编排决策、记忆体分级存储、向量数据库检索五大核心模块;DCU聚焦词元生成侧的并行计算加速,承担批量解码、KV缓存优化、预处理加速、长序列支撑等任务。双芯协同可根据Agent并发规模、任务复杂度与词元生成量级动态调配资源,将算力精准转化为词元生产力。

此外,随着词元经济加速渗透实体产业,异构资源协同、模型高效迁移、安全可信与生态接入等

问题,成为产业规模化落地的核心诉求。“Agent to Token开放计算架构”在算力、互连、安全、软件栈4个维度全面开放,为多样化算力提供从Agent任务调度,到词元生产的完整组织方式。

算力开放层面,该架构支持海光信息双芯原生组合、异构芯片交叉适配等多元部署模式,可灵活适配不同行业、不同规模的词元生产需求;互连开放依托自研HSL高速互连协议,突破传统芯片交互带宽瓶颈,大幅提升芯片间数据传输与词元全链路流转效率;安全开放依托海光信息芯片级内生

全栈安全能力,从硬件底层实现数据隔离、权限管控、加密防护,为Agent业务运行、词元全流程流转提供可信安全屏障;软件栈开放配套全栈开发工具链,适配百余种主流AI框架,大幅降低大模型迁移、二次开发与商业化部署门槛。

据介绍,下一步,海光信息将进一步联合整机、软件及行业伙伴推进产品适配,推动国产CPU、DCU和不同加速卡在统一计算体系下调配应用。行业用户可沿用部分现有设备,并根据业务增长分阶段扩容,推动算力资源更高效落地实体业务。