具身智能的下一站：数据采集

为什么机器人行业的尽头，是国家级数据基础设施？

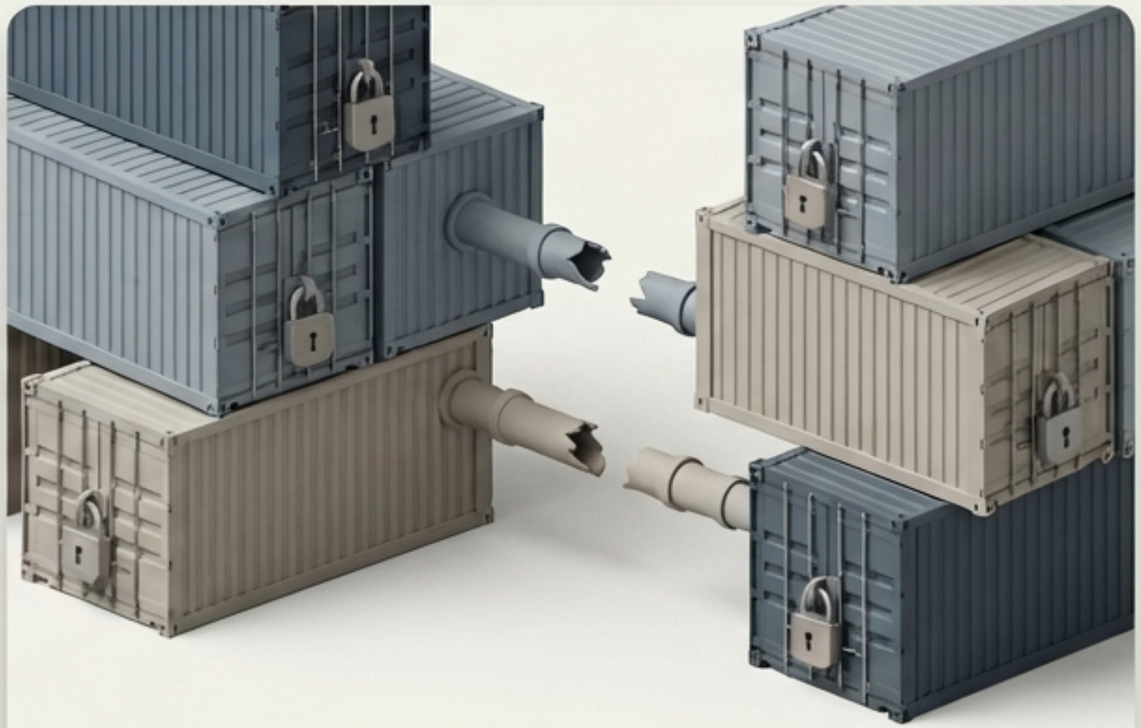
跨越鸿沟的唯一解：真机数据

真机数据必不可少

只有真机在真实世界干活采回的数据，才能跨越仿真与现实的鸿沟。

行业的困境：极度稀缺与“沉默的矿山”

2025年前全行业开源真机数据极度匮乏，几乎处于真空状态。



各家格式不同、标注各异、互不相通。采了，也用不了。

国家队入场：修路与定规矩

从私域流量池，走向公共数据基础设施。

定交规

开放原子开源基金会联合信通院、百度、乐聚等，制定标准、治理框架与合规路径。

铺高速

搭建具身智能开源数据集社区，打造不属于任何单体厂商的“公共仓库”。

数据规模化元年：从手工坊到大工厂

14

个具身训练场

（全国总数，乐聚占9个）

2500万+

条真机数据/年

（乐聚年产能）

60000+

分钟开源数据集

（已覆盖工业/商业/家庭）

20000+

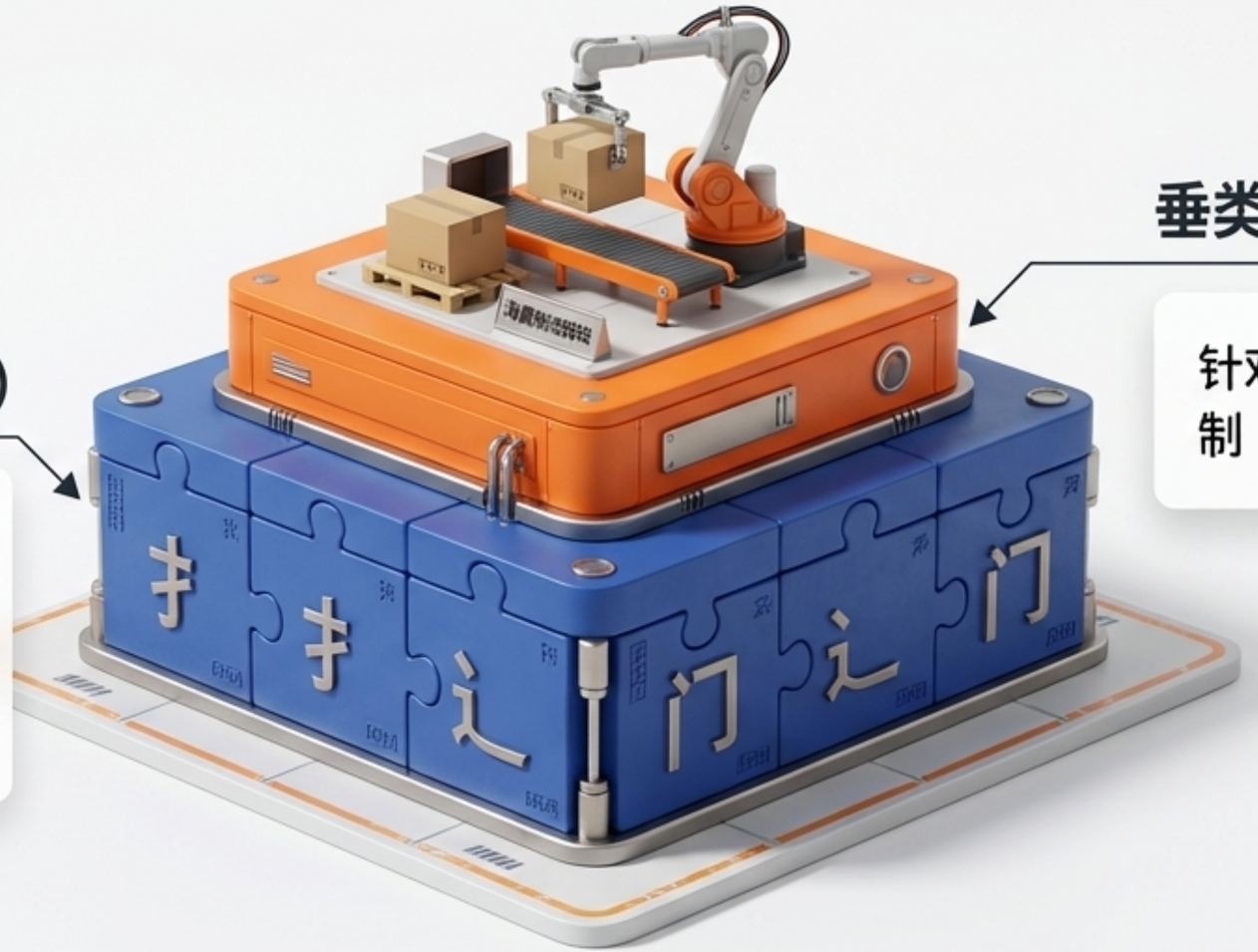
小时交付量

（已向付费客户商业交付）

真正跑通了“采集 — 开源 — 交易”的全链路。

数据的双层商业逻辑：通用做生态，垂类做收入

原子技能如“汉字部首”，拼装出无限的商业化垂直场景。



通用数据（生态底层）

包含开门、抓取等117种原子技能。
全行业公共底座，适合开源。

垂类数据（直接变现）

针对特定产线/工位定制（如海晨物流拆垛）。

谁要数据？一平台供给，多边变现

终端制造方

中兴/海晨/一汽

需求

垂直场景技能数据包

目的

立刻能在我的场景里干活

算法/模型公司

百度/蚂蚁灵波

需求

海量、多模态训练语料

目的

训练多模态大模型与策略学习

高校/科研机构

哈工大

需求

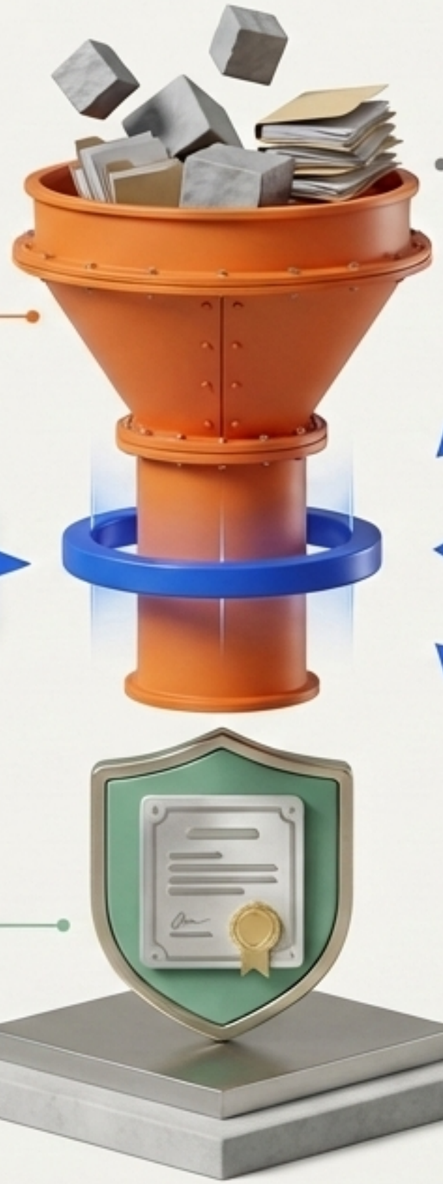
可复现、有标注的基准数据

目的

跑 Benchmark，发表学术论文

同一批训练场，同一个开源社区，服务三大核心客群。

终极形态：从“沉没成本”到“金融资产”



沉没成本：权属不清

原始数据素材

数据确权：明确所有权

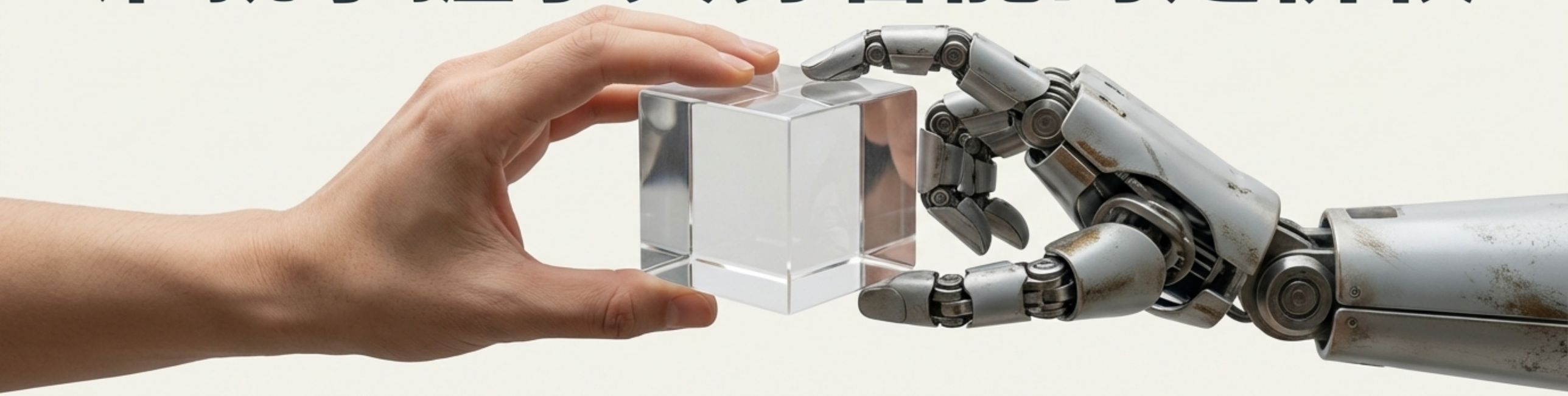
标准制定：建立规则

资产化：定价与保险

生态闭环：吸引资本投入

金融资产：合规与保障

谁掌握了真机数据，谁就掌握了具身智能的定价权



具身智能的背后，是一座正在被确权、被规模化开采、并且能持续变现的“金矿”。

关注开放原子开源数据集，共同建设机器人时代的公共基础设施。