中国AIGC数据标注产业全景报告

Panoramic Report of Generative AI Data Labeling Industry in China

2023.11　　杨净　　量子位智库 QbitAI Insights

序　言

数据标注，正迎来关键时刻。作为AI认识世界的起点，数据标注本质上是将现实世界信息结构化、数字化，充分发挥数据信息的价值。

大模型时代到来，AIGC众多垂直场景落地，以及通用智能、具身智能等前沿领域探索，与高质量、专业化的场景数据密不可分，数据标注从劳动密集型加速朝着知识密集型转型，行业壁垒进一步提高。

作为底层基础服务，数据标注贯穿大模型全生命周期（训练测试、评估验证和应用迭代）。一方面，牵涉关键Know-how，更多大模型公司/AI企业选择自建标注团队和管线；另一方面，上下游合作关系将更为紧密和耦合，专业数据服务提供商更多机会将在垂直领域，帮助企业完成私有化部署。

机遇与挑战并存。合成数据作为新衍生赛道，潜在市场空间巨大。与此同时，数据标注标准难以统一、数据处理流程尚未规范，高学历多领域多专业成为标注人才的硬指标。

目 录

01 大模型时代下的数据标注

02 AIGC数据标注四大变化

03 AIGC数据标注三大影响因素

04 数据标注产业竞争格局/市场规模

05 数据标注代表玩家案例集

01

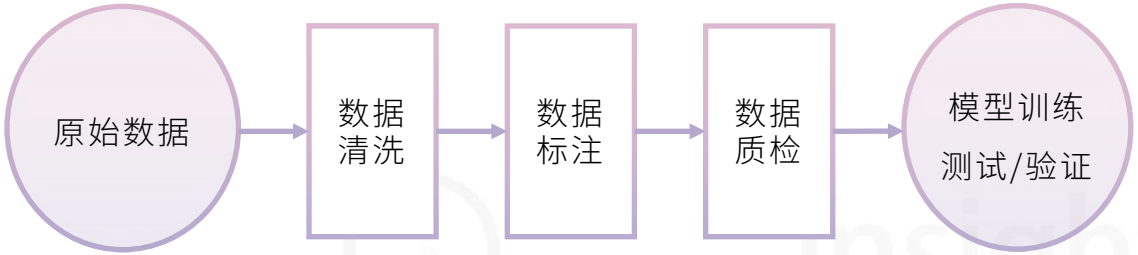
大模型时代下的数据标注

数据标注是AI认识世界的起点

数据标注是将原始数据进行加工处理，比如分类、拉框、注释、标记等操作转换成机器可识别信息的过程。

国内数据标注厂商，广义称之为**基础数据服务提供商**，通常需要完成数据集结构/流程设计、数据处理、数据质检等工作，为下游客户提供通用数据集、定制化服务、数据闭环工具链等。这也是本次AIGC数据标注全景报告的研究对象。

一般数据处理流程：



数据标注中的**二八定律**

通常在一个AI项目中，数据准备工作需要80%时长，模型训练和部署仅占20%。

根据**原始数据类型**以及**训练任务**划分：

- **文本**：
  词性标注、分类标注、情绪标注、命名实体识别、语义标注、意图标注等；
- **图像**：
  图像分类、语义分割、实例分割、拉框、OCR转写等；
- **音频**：
  语音识别、声纹识别、语音转写等；
- **视频**：
  目标跟踪、行为识别等；
- **3D点云**

大模型时代下的数据标注

上市公司股价狂飙，创业公司融资加速

海天瑞声是国内唯一一家AI数据上市公司，今年2月以来股价受ChatGPT热潮曾一度狂飙，截至11月10日股价较年初上涨59.75%。

创业代表公司融资情况

星尘数据｜22年12月5000万A轮	曼孚科技｜23年9月数千万B轮
标贝科技｜23年4月超亿元B2轮	恺望数据｜23年4月战略融资
整数智能｜23年6月数千万Pre A轮	23年9月数千万Pre A轮
柏川数据｜23年7月千万元天使轮	

大模型数据解决方案多处开花，以一站式、定制化服务为主

围绕大模型开发全生命周期（包括预训练、监督微调、RLHF、红队测试、基准测试等），专业数据服务商、大模型企业、AI公司等各方都拿出相关数据解决方案，大部分以一站式、定制化服务为主。

- 云测数据：面向垂直行业大模型数据解决方案
- 星尘数据：星尘COSMO大模型数据金字塔解决方案
- 澳鹏Appen：AI聊天反馈和基准测试两大解决方案
- 火山引擎：火山方舟（涵盖数据服务模块）
- 百度：首个大模型数据标注基地

大模型范式涌入数据标注，自动化标注门槛大幅降低

以SAM模型为代表的图像分割模型开源；GPT-4、GPT-4V为代表的大模型也被验证在文本、图像领域标注具有可行性，并衍生出专门做数据标注的大模型，大幅降低自动化标注门槛。国内不少数据服务商进行相关大模型研发，部分产品已经发布：

- 海天瑞声：数据生产垂直大模型（研发阶段）
- 曼孚科技：自动驾驶数据标注视觉大模型（已完成研发）
- 龙猫数据：自动驾驶大模型AutopilotGPT（发布）
- 商汤：明眸SenseAnnotation自动化数据标注平台（发布）
- 标贝科技：烘焙师大模型Baker-GPT（发布）

智能驾驶新感知范式，BEV+Transformer是机遇也是挑战

作为最具代表性应用场景，智能驾驶迎来新感知范式：以BEV+Transformer为代表的四维感知替代掉2D+CNN为代表的二维感知方案，给数据服务厂商带来更多机遇与挑战，包括不限于标注场景难度大、数据量产能力要求高等。目前国内部分厂商给出了数据闭环工具链和解决方案等。

（图源：特斯拉）

AIGC重塑数据标注

量子位智库认为，数据标注正迎来重新洗牌的关键时刻，有四大关键趋势：

1、数据标注要求从客观到主观，很难建立统一标准

大模型的开发范式决定了大模型数据标注对自然语言要求要求很高，包括排序、改写、多轮对话、评估等操作，难以依靠客观的评价体系，比如准确率、效率等。

2、高学历多领域人才成刚需，缺口或达百万

本科以上多领域多专业开始成为标注人才的硬指标，标注角色也随着大模型全生命周期更为细分，比如AI训练师、模型精调师、指令工程师等。

3、产业链重构，大模型公司/AI企业涌入

大模型Know-how涉及到数据处理流程的设计，大模型公司/AI企业开始自建数据标注团队和数据处理管线，甚至对外输出服务，产业链重新洗牌。

4、国内百亿级市场规模，合成数据增速最高

量子位智库预计，国内AI基础数据服务市场规模将达百亿规模，约占全球市场10%份额。其中合成数据作为衍生出来的新赛道，存在巨大市场空间，增速超40%。

02

AIGC数据标注四大变化

需求变化：与行业场景强相关，高质量数据需求长期且持续

大模型时代的到来，正加速推动人工智能开发从以模型为中心朝着以数据为中心的方向转变。

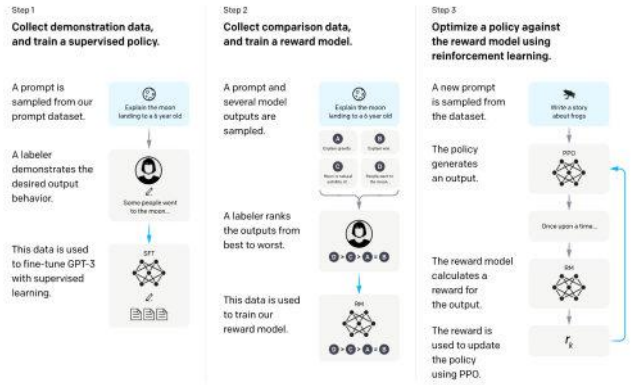
高质量数据服务需求贯穿大模型全生命周期。

目前大模型技术路径已经完整清晰，训练流程主要分为三个阶段：

*实际训练过程中，部分垂直领域大模型需用小规模语料进行二次预训练操作

数据处理流程设计涉及大模型Know-how，直接决定大模型性能好坏。

尤其后两个阶段需要专业人士**生成**数据或对数据进行**改写**或**排序**，最终形成符合人类标准（比如专业逻辑、核心价值观等）高质量数据。



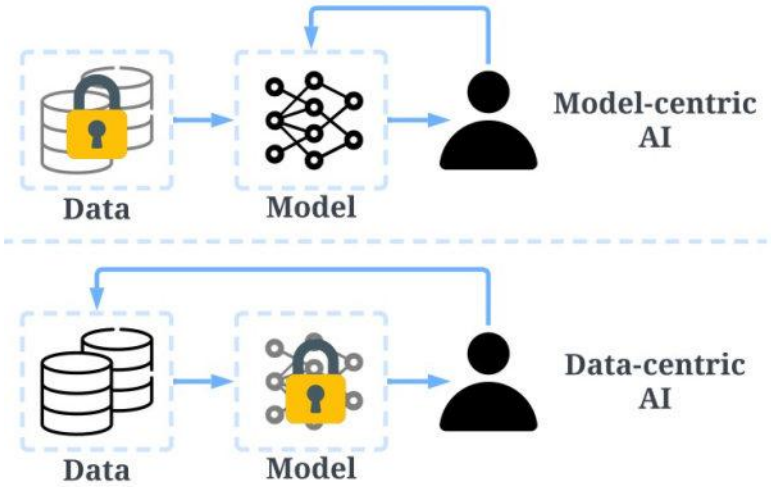
（图源：OpenAI官网）

而后随着大模型持续地实时更新迭代、朝着多垂直领域落地，尤其通用智能、具身智能等相关探索，如何快速扩展到更多真实边缘场景，高质量场景数据也将成为刚需。

除此之外，实时保障输出内容的安全合规，也远比以往更受重视。从训练、迭代到应用落地，**数据服务贯穿大模型全生命周期**。

广泛认知里，大模型是以数据为中心的产物。数据数量和质量很大程度决定着大模型能力的上限。

- 以**模型**为中心：迭代模型，数据相对固定。
- 以**数据**为中心：关注数据本身，模型成为了数据的「容器」。



（图源：Data-centric AI: Perspectives and Challenges）

企业端客户需要长期且持续的数据服务，产业链上下游供应关系远比以往更为紧密和耦合。

处理流程侧变化：标准从客观到主观，高学历多领域成人才硬指标

数据标注从劳动密集朝着知识密集型转变。

	传统数据标注	大模型数据标注
领域划分	按不同领域或任务划分	按不同阶段划分
具体实操	拉框、描点、转写等操作	排序、改写、生成等操作
标注要求	偏客观	偏主观
评价指标	准确率+效率	难以对齐标准
解决方案	工具/平台标注+人类质检	专业培训、定期开会对齐等举措
人才要求	专科为主	本科以上，多领域专业人才
标注角色	按职能划分 标注员、质检员、管理员	按阶段划分 AI训练师、模型精调师、指令工程师、红队测试军团等。
覆盖区域	主要集中在三四线城市	重新打散

例如，百度在海口专为大模型建设的数据标注基地，本科比例100%，培训专业人才已达1000人。未来五年，数据标注相关专业人才缺口将达百万量级。

业务变化：合成数据成新衍生赛道，潜在市场空间巨大

所谓合成数据，即是用AI生成数据而非真实产生，能够替代真实数据来训练、测试和验证大模型。目前主要在自动驾驶、机器人、生物医药等领域应用。英伟达Meta亚马逊等全球科技巨头均有相关布局（投资、收购等）。OpenAI CEO Sam Altman曾放言：未来所有数据都将变成合成数据。

量子位智库预计，合成数据将成为未来增速最快赛道，年增长率可达45%。

合成数据的优势&特点

1、降本增效

降低数据获取成本，生成数据自带高质量标注，缓解“数据荒”问题。

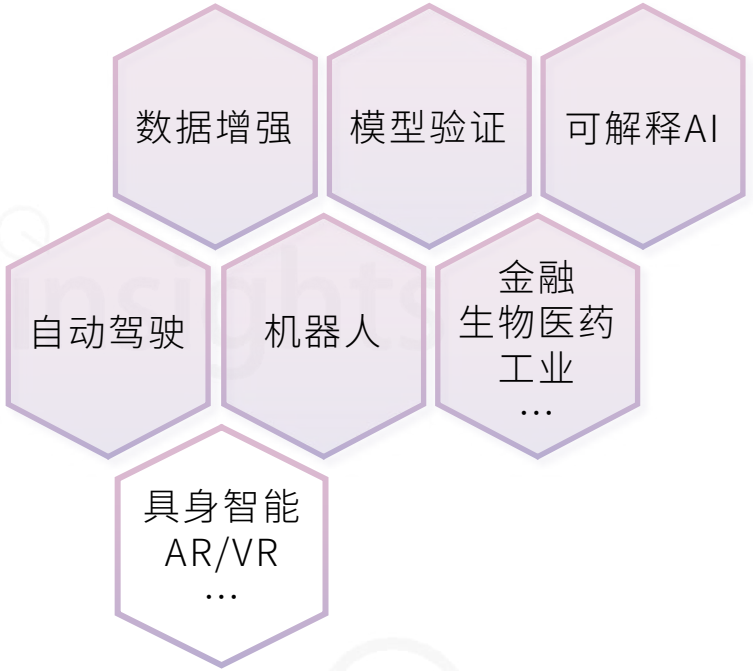
2、数据可定制

应用可扩展性强，灵活度高，可覆盖更多边缘、长尾场景。

3、隐私安全

天然规避掉数据隐私安全合规的问题。

应用场景



企业案例

（图源：官网）

群核科技Coohom Cloud（群核云）作为目前为数不多提供室内场景数据服务的代表厂商，能针对不同应用场景合成2D、3D数据集，客户覆盖全球，服务多家海内外科技巨头公司，并于英特尔在产研等开源性项目上进行深度合作。

供应链变化：重新洗牌，大模型公司/AI企业涌入

大模型公司/AI企业自建数据处理管线，对外输出大模型数据解决方案，传统产业链重新洗牌。部分厂商还具备云服务能力，同数据服务打包输出，更易建立起客户之间的口碑和信任，具备竞争优势。

硬件/云服务厂商、人力资源厂商

百度智能云	火山引擎	阿里云	华为云	腾讯云	京东云	…	综合招聘平台 …

基础数据服务提供商

专业数据服务提供商

海天瑞声	云测数据	星尘数据	曼孚科技	标贝科技	龙猫数据	群核科技	倍赛科技
数据堂	晴数智慧	37度数据	景联文科技	科乐园	整数智能	博登智能	恺望数据
澳鹏中国	卓印智能	未有科技	风云数据	朗势科技	柏川数据	冰山数据	…

大模型公司/AI企业　　中小团队

百度智能云	火山引擎	阿里云	京东	商汤科技	毫末智行	…	…

数据需求方

（AI企业、传统企业、政企机构、科研机构等）

03

AIGC数据标注三大影响因素

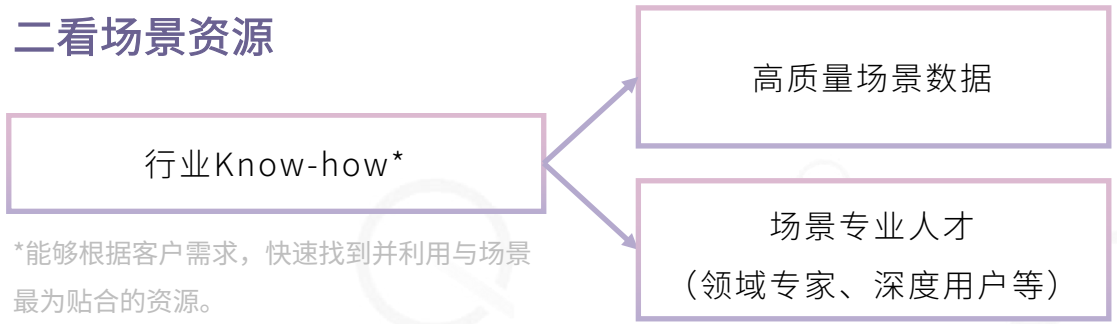
三大影响因素：以技术+场景聚合的飞轮效应

一看技术能力

数据标注作为AI底层服务，最本质是为客户降本增效。持续迭代技术能力的企业将有机会脱颖而出，包括不限于以下几点：

- 数据闭环工具链的智能化水平
- 对大模型/算法Know-how的理解
- 数据工程化能力、数据基础设施建设
- ……

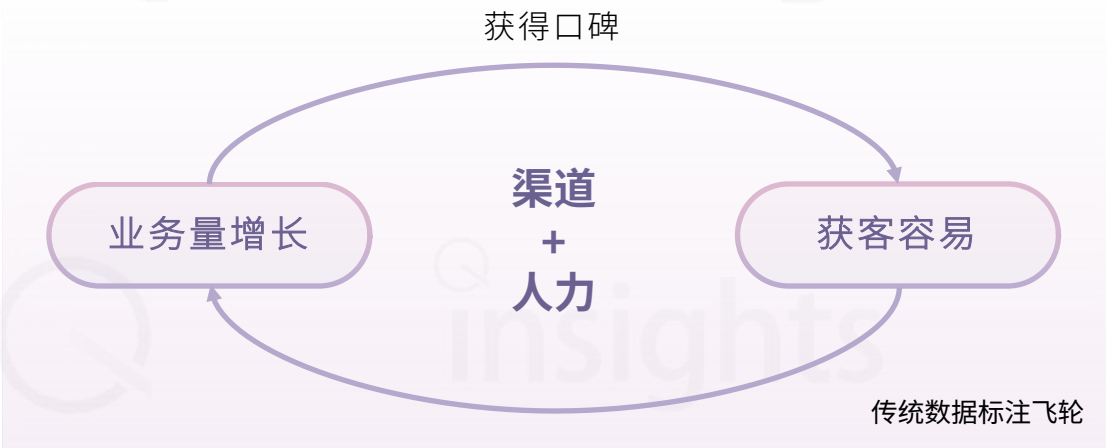
二看场景资源



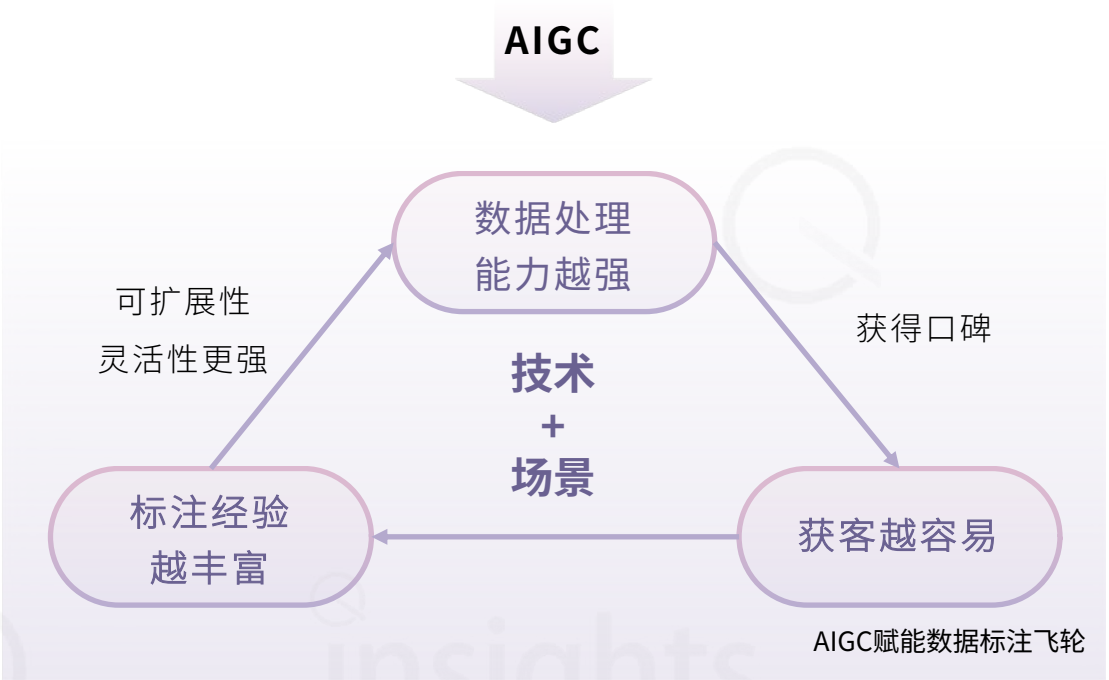
*能够根据客户需求，快速找到并利用与场景最为贴合的资源。

三看飞轮效应

- 数据标注仍具备**飞轮效应**；
- **新创业公司**入局门槛进一步提高；
- 专业数据服务商更多机会将在**垂类场景**，帮助企业完成私有化部署；
- 对外输出数据服务的大模型公司/AI企业也存在竞争优势。



传统数据标注飞轮



AIGC赋能数据标注飞轮

04

产业竞争格局/市场规模

市场竞争格局

数据标注行业传统依靠渠道、人力等形成的低成本竞争优势将被重塑，数据需求方将更看重数据质量、场景多样性和可扩展性。基于以上原因，量子位智库将从数据基础设施、场景资源两个方面来分析目前的业内玩家分布及现状。

第一象限：有技术有场景的明星公司

该象限存在两种情况：第一种是模型层公司本身有大模型技术范式以及场景落地经验积累，可快速输出数据解决方案，与云服务打包输出建立信任；第二种则是主要以技术驱动的明星企业，大部分拥有数据闭环工具链，再结合几年来行业经验，在大模型浪潮下易受到企业用户青睐。

第二象限：有强技术支撑的创业新势力

该象限主要聚焦在近两年创立的创业公司，主要以自动驾驶场景作为切入点，再覆盖到AIGC及其他领域。他们饱受资本市场认可，以恺望数据为例，一年半时间就是完成了三轮融资。

第四象限：场景壁垒更为深厚的行业玩家

该象限着更为深厚的行业数据壁垒，可为下游用户提供高质量数据集或拥有大模型数据标注团队，以海天瑞声为例，不仅是LIama2的唯一中国伙伴，还发布超大规模中文多轮对话数据集DOTS-NLP-216，合作企业超810家，覆盖全球近200个主要语种及方言，有近20年行业深耕。

我国数据标注行业企业竞争格局

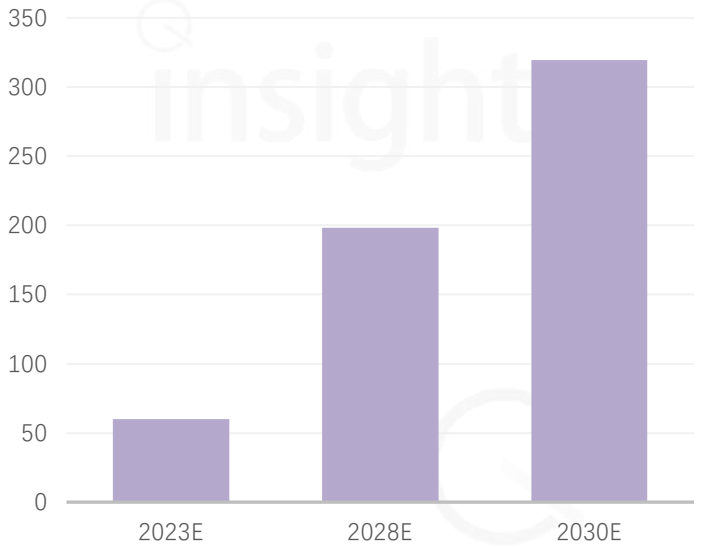
国内基础数据服务百亿市场规模

人力标注（2017年前）	平台/工具标注（2017-2022年）	重新洗牌（2023-2025年）	知识密集（2025年后）
以训练任务、算法模型为导向； 简单图像标注为主。	以自动驾驶为代表的场景爆发； 人力标注无法满足数据需求，自动化标注兴起； 大量AI数据初创公司开始涌现。	数据质量驱动； 产业链重新洗牌，更多企业参与数据标注，供应合作关系紧密； 创业门槛提高。	人机协同关系进一步耦合，人力更多承担关键决策角色； 市场竞争格局趋于稳定。

- 【人力标注】关键节点：2007年，李飞飞团队启动ImageNet，借助亚马逊众包平台完成图像分类和标注来训练机器学习算法。数据标注从此拉开序幕。
- 【平台/工具标注】关键节点：2017年，以数据驱动的深度学习成为行业共识，自动驾驶大爆发，国内外初创公司涌现，数据标注迎来庞大的市场需求。
- 【重新洗牌】关键节点：2023年，以ChatGPT为代表的大模型涌现，更高质量、专业化的数据标注成为刚需。
- 【知识密集】关键节点：垂直大模型落地加速，数据处理范式、标准基本确定。未来机器将满足大部分标注需求，人力将承担关键决策任务。

国内AI基础数据服务市场规模

单位：亿元



需求推算：作为AI底层基础服务，始终依托于人工智能的发展，约占人工智能市场份额10%左右。目前大模型垂直领域落地仍处于探索阶段。

典型样本：海天瑞声市占率达12.9%，上半年营收比去年同期增长翻番。

05

数据标注代表玩家案例集

百度智能云 百度智能云数据众包

百度智能云数据众包，依托百度10余年AI数据经验、产品技术能力和国内产值规模领先的单体数据标注基地，具备数据“采、标、存、管、训”一体化的服务能力，根据特定领域、特定场景的客户需求与委托，可提供数据采集、标注、加工等处理服务，为客户交付标准化、结构化的服务成果。

当前，百度智能云升级大模型数据服务能力，**在海口市建设全国首个专业大模型数据标注基地，专业大模型数据标注师达数百人，人员本科率达100%。**

大模型标注服务：人员、工具、质控、研发多管齐下，保证高质高效

指令数据标注服务
交付：输入提示和输出的高质量监督数据

人类反馈标注服务
交付：代表人类偏好的打分排序数据

大模型数据标注生产线
大模型数据生产Copilot赋能

数据接入 → 资源调度 → 数据分发 → 数据标注 → 质量审核 → 数据交付

规则增强学习 | 自动分类 | 智能标注 | 自动质检

标注资源
各领域众包专家+专职基地人力

+

运营能力
专业化数据咨询+安全标注方案

大模型评估服务：全面评价应用表现，洞察短板，牵引优化

洞察与优化

可视报表与案例分析

优化提案与服务支持

大模型能力评估体系

应用能力	通用能力	学习能力
问答	指令约束满足	SFT
创作	上下文记忆	In-Context-Learning
对话	跨语言处理	
代码		
基础语言处理		

评估流程与工具

专业	公正	高效
人员定向募集与准入	盲评、拟合多轮审验	Copilot辅助评估

群核科技 群核科技 MANYCORE

Coohom Cloud（群核云）是群核科技（酷家乐）推出的，面向室内智能体认知和图形智能的AI训练合成数据平台。基于真实三维场景数据资源以及AIGC技术的驱动，提供丰富的2D/3D数据集，针对智能机器人、人工智能、元宇宙、智能房产、自动驾驶等领域，为AI模型以及仿真器研究提供丰富的训练资源，让智能体更智能。

行业 智能机器人 人工智能 元宇宙 智能房产 自动驾驶

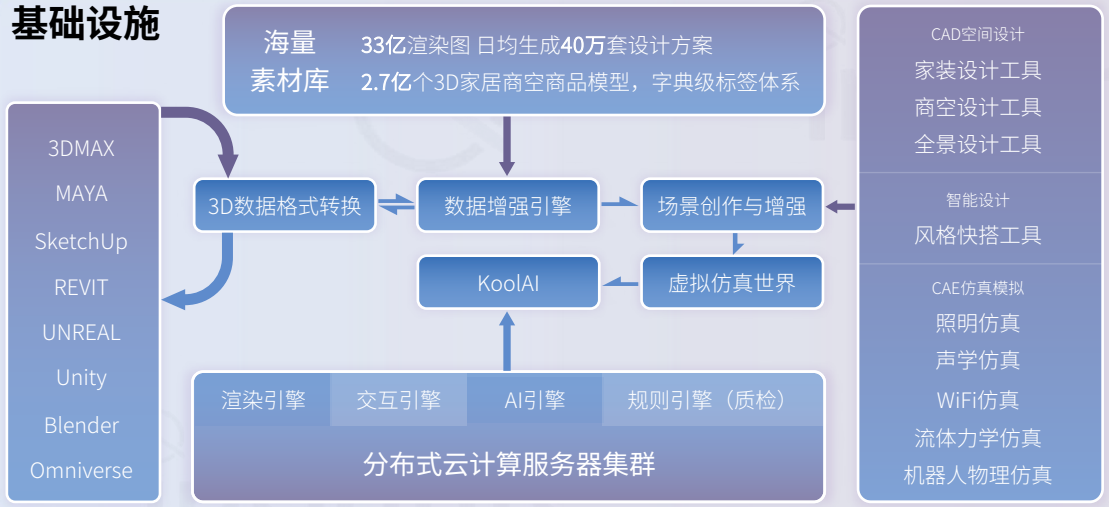
技术应用 数字孪生 视觉感知 三维重建 内容生成 SLAM 决策与控制

解决方案 提供以虚拟仿真合成数据集为中心的一站式服务

	成本	优劣对比	数字化
真实数据	人力为主成本高	采集耗时久/标注错误多 成本高昂/侵犯隐私	无：一次性数据集+项目制
仿真数据	**算力为主成本低**	复杂场景标注成本低/支持高难度采集 完美实验/格式统一/多样性丰富	**有：高复用性数据集+基于任务的灵活修改**

数据集作为AI训练的核心要素，其规模和质量与算法效果，效率密切相关

基础设施



应用产品


室外无人机


导航机器人


工厂机器人


厨房机械臂


清洁机器人

生态兼容

仿生软件：Isaac Sim、UE、Gazebo、Unity 等

数据格式：USD/UE/SDF/OBJ/HM3D/PCD/COCO/VOC/NYU40标签/自定义

超高性价比

成本降低10倍
场景确定后，数据集规模越大，单图成本越低

效率提升10倍
GPU集群并发渲染，可合成20w组数据/日

体验提升10倍
可视化交互工具，实现所见即所得

质量提升10倍
像素级精准标注

合作成功案例与伙伴

论文
- InteriorNet BMVC 2018
- Structured3D ECCV 2020
- MINERVAS CGF 2022

高校&企业
- Imperial College London
- SVIP LAB
- 英特尔 科沃斯 追觅 美的 等

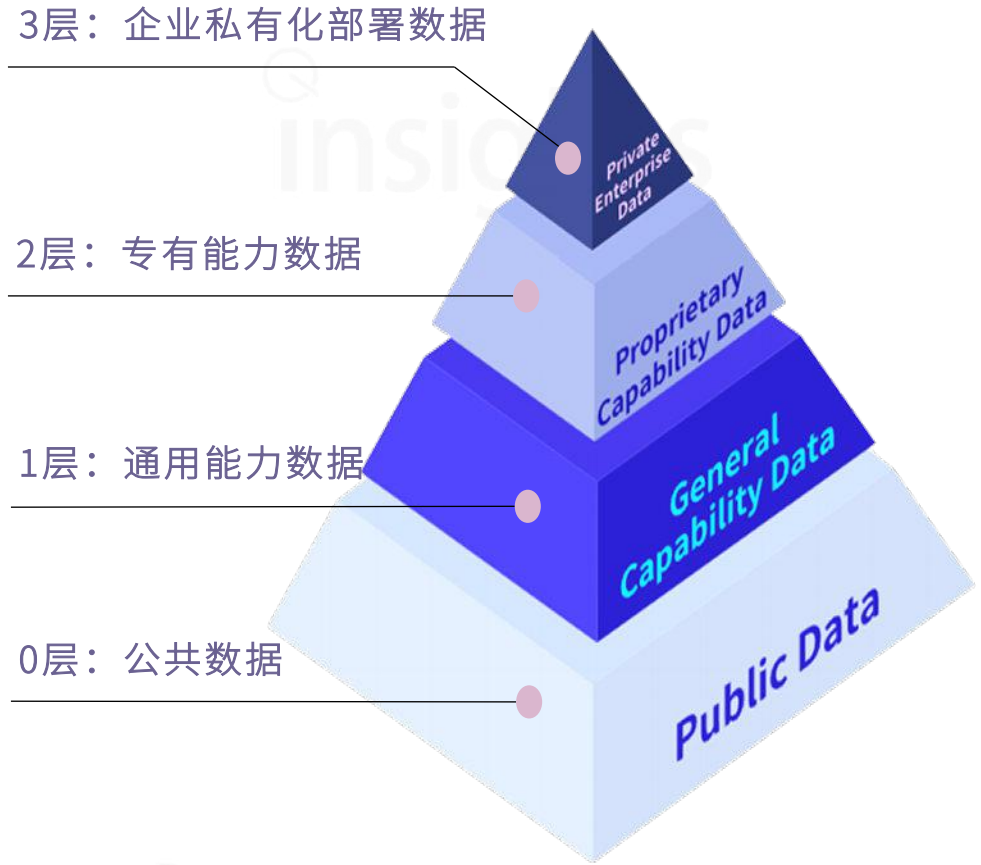
星尘数据 STARDUST 星 尘 数 据

星尘数据成立于2017年5月，2023年1月宣布完成5000万A轮融资。通过自动化标注技术、数据策略专家服务和数据闭环系统，服务自动驾驶（50+头部客户）、大模型、智能家居、智慧城市、智能机器人、智慧医疗、智慧教育、智能零售、智能遥感、智慧金融等众多数据场景。

核心产品：Rosetta平台3.0

可支持**几万人**以上同时在线标注，数据年处理量**过亿**，可提供先进的AI算法辅助标注工具和项目管理工具，可支持图像、点云、文本、语音、多模态等各类型**100+种**主流采集和标注场景，目前平台自动化水平达到60%以上，数据质量达到**99.9%**。

星尘COSMO大模型数据金字塔解决方案



四层数据结构，加速大语言模型构建

云测数据

云测数据是Testin云测旗下AI训练数据服务品牌，以高质量、场景化的AI训练数据服务为基础，持续为智能驾驶、智慧城市、智能家居、智慧金融等众多领域提供通用数据集、数据标注平台&数据管理工具、数据采集/数据标注等服务。

面向垂直行业大模型AI数据解决方案

持续预训练

垂直行业知识

定向垂直场景的数据服务能力

- 场景化数据采集能力
- 持续订阅服务能力
- 基于数据要求清洗分类能力

下游任务微调

人机协作优化

基于下游任务微调的人机耦合标注能力

- 多轮对话
- 4D叠帧
- 音视频转写
- 文章判断
- 文生图排序
- OCR预识别
- Prompt编写
- …

灰度发布联调

基准评测

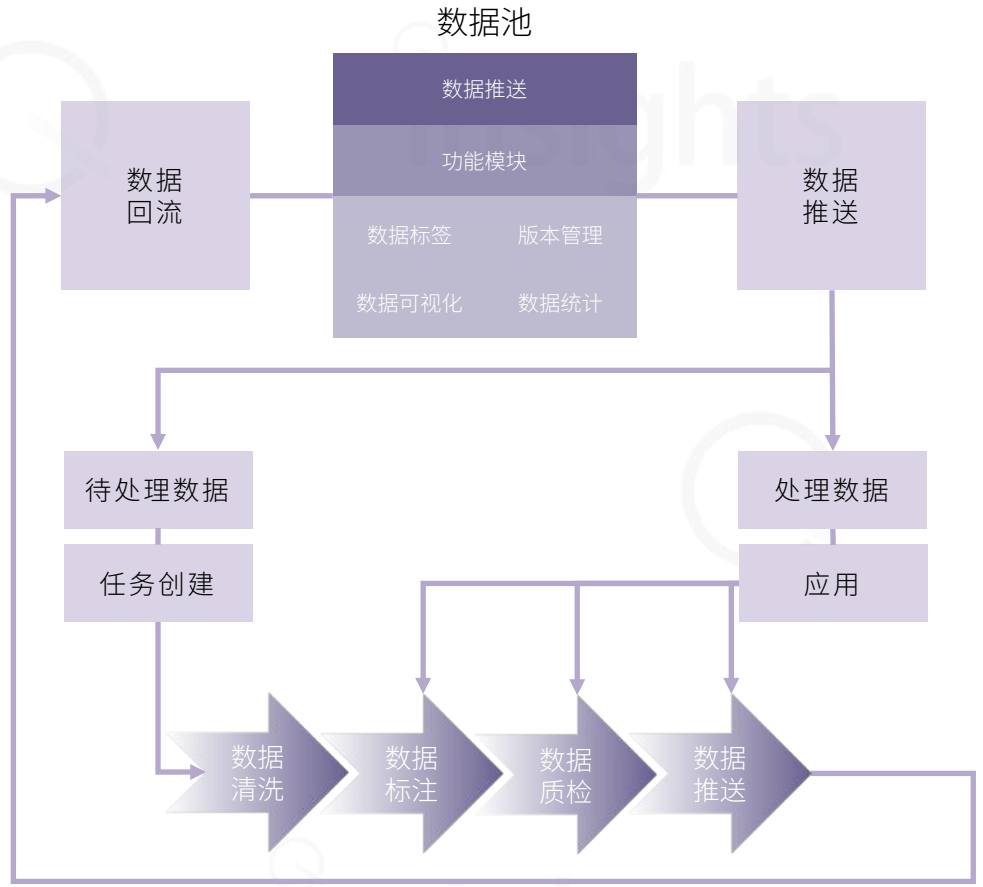
基于定向垂直领域人员测试

- 特定领域专家池
- 场景化服务能力

系统集成支持特定数据回流处理

适用于新一代AI工程化数据处理平台

适用于新一代AI工程化数据处理工作台



*通过标准API接口与其他业务集成

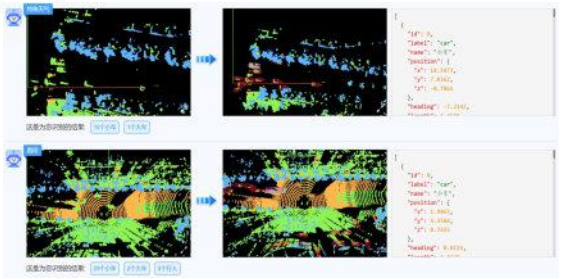
龙猫数据

龙猫数据成立于2014年，专业提供自动驾驶、计算机视觉、智能语音、自然语言理解数据采集标注服务，具备数据标注、数据采集、内容审核等能力。针对AIGC类业务，龙猫数据2016年推出标注平台1.0版本，目前已执行1000+项目，标注人力2000+。

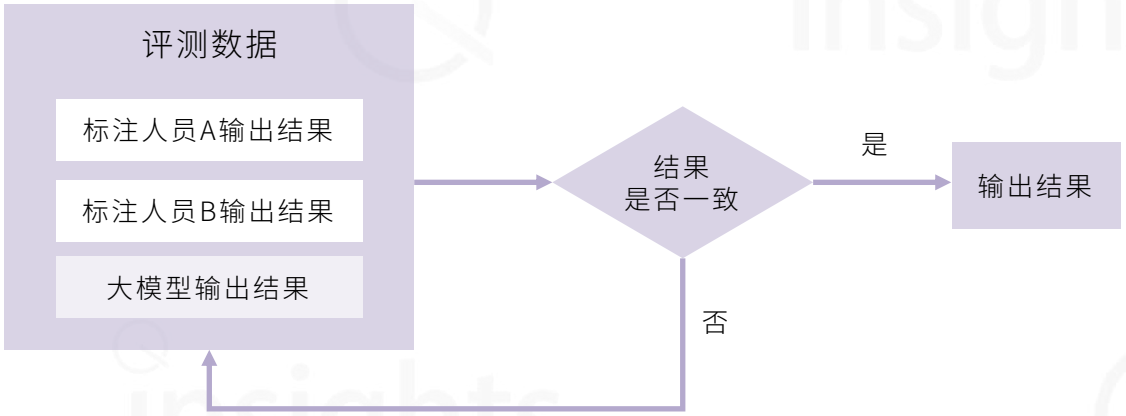
自动驾驶大模型AutopilotGPT

AutopilotGPT是基于Transformer的百亿参数模型，可识别图片、点云类型。支持多传感器数据类型，可进行目标检测、目标追踪、目标分割、行驶区域识别。只需上传图片（通用格式均可）、点云pcd格式，就可自动识别结果。

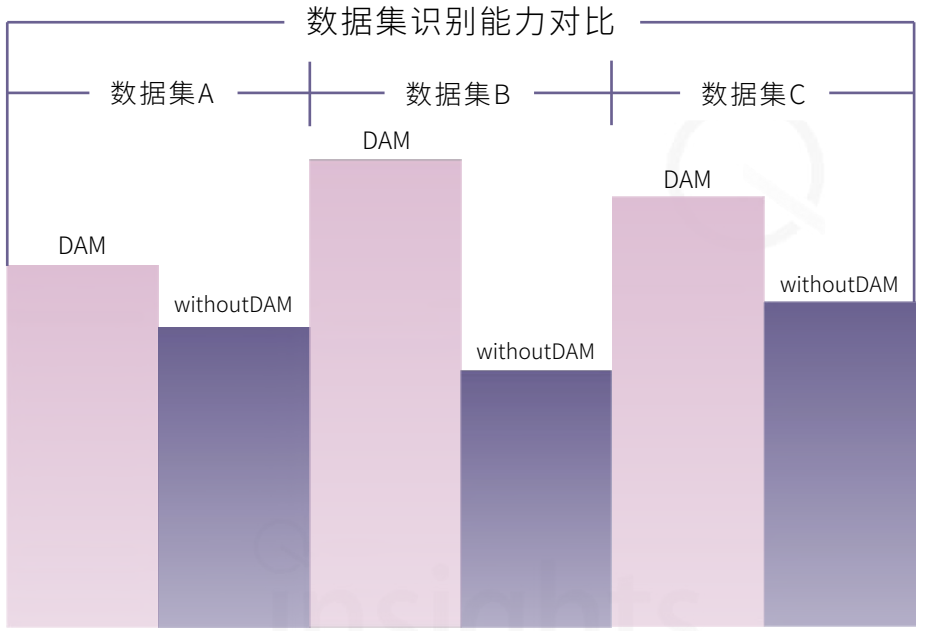
点云示例：



AIGC数据标注流程「质量保障」：引入大模型，交叉验证人工对齐评测结果。



AutopilotGPT示意图



恺望数据

恺望数据成立于2022年2月，团队成员来自字节跳动、阿里巴巴、Uber、Momenta、奔驰等头部企业。公司致力于打造AI数据自动化平台，并为车企、自动驾驶公司以及人工智能等跨产业企业提供一站式AI数据解决方案，目前客户数已超百余家。

核心能力	提供合规数据、高质量数据、高效率的稳定大规模数据。
创新技术与平台模式	自动化AI数据产线 / 高效率运营

“3456”数据服务工具包

- **“3D辅助标注”工具平台：**可在2D中标记后反投影到3D中找到标注物。
- **“4D-BEV数据拼接与标注”工具链：**可支持大数据流并行作业、可同时支持200万人同时标注， 目前已在车企应用。
- **“5KW大点云”工具平台：**可在8G内存电脑上运行的5千万点云数据。
- **“6大数据生态闭环解决方案”：**供应商生态、行业生态、知识库生态、工具生态、前沿技术生态、专家科研生态。

融资历程

- 2022年9月，千万级天使轮战略融资，投资方包括辰韬资本、三一集团和溪山天使汇，用于加速建设数据快充站以及团队完善，持续为汽车产业的智能化，提供数字化、一站式的数据解决方案。
- 2023年4月，新一轮战略融资，投资方为Plug and Play、辰韬资本，探索出海路径，并继续投入到产品迭代升级当中。
- 2023年9月，数千万元Pre-A轮融资，由亚盛投资领投，清智资本跟投。本轮融资资金将用于自动化产线和工具链的持续研发和迭代。

恺望数据学院

通过高校合作储备及培训有大批高校学生标注员，通过共建产融实训基地的形式为行业迅速提供大量稳定且优质的数据标注服务，同时运用AI工具辅助管理、基地化管理、专业化高级人才培养等方式，获得最优人力和最优人效的平衡，降本增效表现领先行业。

目前恺望数据学院已培训**50所学校**，培养**超过1500名学生**为恺望提供数据标注服务，计划至今年年底将**超过2000人**规模。

我国值得关注的数据标注行业代表机构TOP20

基于数据基础设施建设、大模型/AI技术理解以及行业深耕和其他因素，量子位智库评选我国值得关注的20家数据标注机构。

百度智能云	海天瑞声	云测数据	星尘数据	龙猫数据
群核科技	倍赛科技	标贝科技	曼孚科技	晴数智慧
恺望数据	整数智能	博登智能	火山引擎	商汤科技
数据堂	37度数据	未有科技	景联文科技	澳鹏中国

*排名不分先后

量子位 量子位智库 insights

关于量子位智库：

量子位旗下科技创新产业链接平台。致力于提供前沿科技和技术创新领域产学研体系化研究。

面向前沿AI&计算机，生物计算，量子技术及健康医疗等领域最新技术创新进展，提供系统化报告和认知。

通过媒体、社群和线下活动，基于专题技术报道及报告、专项交流会等形式，帮助决策者更早掌握创新风向。

关于量子位：

量子位（QbitAI），专注人工智能领域及前沿科技领域的产业服务平台。

全网订阅超过500万用户，在今日头条、知乎、百家号及各大科技信息平台量子位排名均为科技领域TOP10，内容每天可覆盖数百万人工智能、科技领域从业者。

微信号：Qbitbot020
量子位智库小助手