iResearch
艾 瑞 咨 询

2024年中国AI基础数据服务研究报告

CONTENTS

目录

01/ AI基础数据服务行业概述

AI产业整体进展

多模态、长文本、大模型小型化成为热点研究方向

在过去几年里，大众已见识到GPT、BERT等大语言模型在自然语言理解和生成方面的卓越能力。相比单一模态的大模型，多模态大模型能够提供更自然的人机交互方式，具备更全面和准确的认知能力，并在不同情境下表现出更高的鲁棒性，从而赋能更丰富和全面的AI应用。因此，多模态技术已成为诸多大模型厂商的研发重点。此外，长文本处理能力的提升，使大模型在理解和生成复杂文档方面表现更佳，能够更好地支持多主题和多步骤的推理任务；通过知识蒸馏、模型剪枝和混合精度训练等技术，大模型得以小型化，减少了计算资源需求，提高了推理效率，使大模型在资源受限设备上高效运行，提升了响应速度和用户体验，保护了用户的数据隐私。聚焦国内AI商业化市场，大模型商业化进程加速，API市场竞争激烈，价格战频现，但同时也反映出供应商间能力同质化的问题，亟需破局；另一方面，央国企凭借较好的数字化基础、丰富的数据资源及业务场景、相对充足的科技投入预算，成为现阶段国内大模型项目建设的主力军，推动了大模型在中国AI产业的商业化落地。

全球AI产品技术进展

多模态

- 概述：多模态大模型能够同时处理和理解包括文本、音频、图像和视频在内的多种数据类型，这使得它们能够提供更自然的人机交互方式，具备更全面和准确的认知能力，并且在不同情境下表现出更高的鲁棒性，从而赋能更丰富和全面的AI应用
- 案例：2024年5月，OpenAI推出GPT-4o，可对音频、视频和文本进行实时推理；2024年5月， Google演示了多模态AI助手Astra

长文本

- 概述：长文本可支持模型理解和生成更复杂的文档、报告、小说等内容，能够更有效地进行知识管理和信息检索，提升了模型对于上下文理解的连贯性，进而更好地实现多主题、多步骤的复杂推理任务
- 案例：2024年3月，月之暗面宣布旗下大模型产品Kimi开启200万字无损上下文内测，其后阿里、百度等大模型厂商均宣布相关大模型产品的长文本能力升级规划；2024年4月，Google、Meta等机构的研究人员先后提出Infini-attention、Megalodon等无限长文本方法

大模型小型化

- 概述：通过知识蒸馏、模型剪枝、混合精度训练等方法，“大模型小型化”相关技术可减少模型参数并降低计算资源需求，提高推理效率，使大模型可在端边等资源受限的设备上高效运行，降低能耗，提升了响应速度和用户体验，还增强了数据隐私保护，未来可能催生更多的创新型智能终端
- 案例：2024年5月，微软表示Windows将附带40多个端侧AI模型，包括可用于搜索、实时翻译、图像生成和处理等任务的小语言模型Phi-Silica；2024年6月，苹果推出Apple Intelligence个人智能系统，内置3B端侧模型，可支持摘要、改写、问答等功能

来源：艾瑞咨询研究院自主研究及绘制。

中国AI商业化落地进展

API调用市场卷起价格战

为争夺大模型客户流量及背后云资源市场，24年上半年云厂商、大模型厂商等相继调整API产品定价，低价甚至免费供应

价格战的积极意义

扩大客户量及使用频次，促使大模型技术在国内更快普及，加速创新型应用的诞生；促进供应商不断优化模型及计算架构，降低模型推理成本；竞争加速产业分层，较少社会整体资源消耗

价格战的另一面为大模型产品技术壁垒的薄弱

尽管大模型相关产品技术仍在迭代，但国内大模型尤其以API方式提供标准化大模型服务的各供应商的产品能力尚未形成较大代际差异；供应商需加速技术及产品差异化建设，获取足够的利润，产业才能健康、可持续的发展

央国企引领大模型项目建设

央国企对大模型的建设投入较多，与其有较好的数字化基础、丰富的数据资源及业务场景、相对充足的科技投入预算相关

2024年上半年中国大模型相关项目中标统计

据智能超参数统计，2024年1-6月中国大模型相关项目中标数量达237个，前5个月披露的项目金额合计已过2023年；行业分布上，电信（47个）、能源（42个）位居1-6月的项目数量头两名，其次为教育、金融、政务等行业，各行业中的央国企均在积极推动大模型项目建设

来源：艾瑞咨询研究院自主研究及绘制。

©2024.7 iResearch Inc. www.iresearch.com.cn

数据、算法、算力是构建AI的三大要素

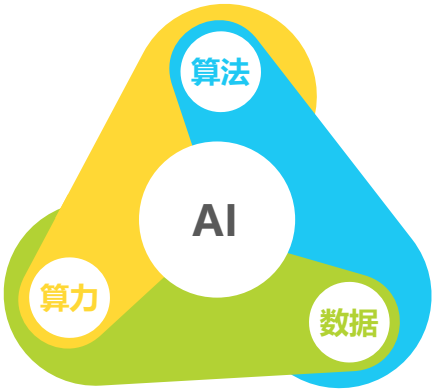
数据、算法、算力的协同促使现代AI技术实现了从理论到应用的飞跃

在人工智能领域，数据、算法和算力是构建AI系统的三大核心要素，三者的协同使现代AI技术实现了从理论到应用的飞跃。数据是AI的基础，大量高质量的数据不仅能够提高现有模型的准确率，还能促进模型的优化和创新。以ImageNet数据集为例，该数据集及相关挑战赛推动了计算机视觉算法的快速发展，2017年是挑战赛的最后一年，物体分类冠军的准确率在7年时间里从71.8%上升到97.3%。近年来，Transformer等预训练大模型在语言理解及生成等领域表现出色，大模型背后的Scaling Law（规模定律）进一步揭示了模型性能与数据量、算力之间的关系，强化了数据在提升AI表现中的关键作用。

构建AI系统的三大核心要素：数据、算法、算力

算法 是处理信息、提取特征、进行预测的逻辑框架

深度学习的兴起，CNN、Transformer等模型的迭代，极大地推动了图像识别、语义理解、文本生成等AI任务的进步



算力 支持算法处理庞大和复杂的数据集

GPU、TPU等AI芯片的发展，使得研究人员能够探索更深、更宽的网络结构，训练更强大的模型，并加速模型的推理速度。硬件的进步直接影响到AI模型的训练效率及规模化应用的可行性，从而不断拓展AI的边界

数据 是模型学习和适应不同任务的基石

高质量的数据能够帮助模型更好地理解现实世界，并做出更精准的预测；反之，即使是最先进的算法，也无法从劣质的数据中获得有效的洞察

来源：艾瑞咨询研究院自主研究及绘制。

高质量数据推动AI系统的发展进步

ImageNet数据集的成功，以及大模型的Scaling Law的发现，都证明着高质量数据对于AI发展的巨大推动

ImageNet见证CV算法在大规模数据集上的性能提升

- 2009年6月，李飞飞团队完成ImageNet初始版本，共有1500万张图片，涵盖了 2.2万个不同类别，这些图片筛选自近10亿张候选图片，并由来自167个国家的4.8万多名全球贡献者进行了标注
- 2012年，由Alex Krizhevsky、Ilya Sutskever和Geoffrey Hinton共同开发的AlexNet在挑战赛上以超过第二名10个百分点的成绩在夺冠，深度学习迎来学术探索与工业应用的热潮
- 2017年是挑战赛的最后一年，物体分类冠军的准确率在7年时间里从71.8%上升到97.3%，超越了人类的物体分类水平

Scaling Law进一步揭示数据对于提升模型性能的关键作用

- OpenAI研究团队于2020年发表的论文《Scaling laws for neural language models》中，系统地探讨了语言模型性能与模型大小、数据集大小和计算资源之间的关系。研究发现，模型的性能（如损失函数值）与这些因素之间存在稳定的幂律关系，即模型的性能会随着数据量、模型规模和计算量的增加而提升
- 现阶段，诸多大模型的研发仍在遵循Scaling Law的发展方向
  ① 今年2月，由ServiceNow、Hugging Face 和 NVIDIA联合发布的用于代码生成的StarCoder2，其数据集规模相比v1大7倍，实现了更准确的上下文感知预测
  ② 今年4月，Meta推出Llama3，其训练数据集超过15T token（是Llama2的7倍），可支持8K的上下文长度（是Llama2的2倍），在MMLU、GPQA、HumanEval等多项基准上成绩优异

来源：艾瑞咨询研究院自主研究及绘制。

AI基础数据服务是AI产业发展的关键支撑

加速高质量数据的获取与标注，推动AI算法的创新与持续优化

根据AI基础数据服务厂商LXT对322家有AI项目经验的美国企业的调研，训练数据的资金投入占这些企业的AI整体建设投入的15%，61%的企业认为未来2到5年对数据的需求量将会增加，62%的企业认为数据质量比数据量更为重要。LXT的调研结果揭示了企业在AI建设过程中对高质量数据的迫切需求。鉴于AI基础数据服务厂商在高效提供高质量数据集方面的专业能力，它们已成为AI研发企业的重要合作伙伴，AI基础数据服务已是推动AI产业发展的关键支撑。

AI基础数据服务厂商对AI算法研发企业的帮助

- AI基础数据服务厂商提供的标准数据集使企业能够迅速开展模型训练，而定制化数据集则助力企业针对特定应用场景优化算法性能
- 不仅缩短了AI研发周期，还显著提升了AI应用的性能和效果，激发了企业在AI领域的创新潜力

推动算法的创新与持续优化

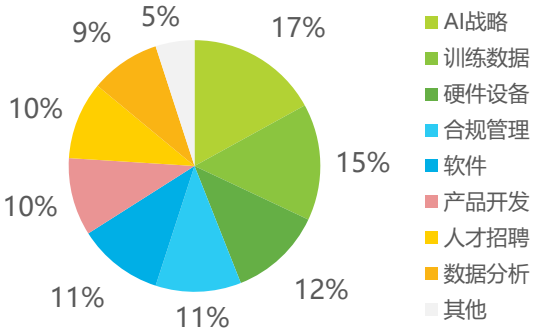
加速数据获取与标注

- AI算法的训练对数据的需求量巨大，且对数据的质量和精确度有着严格的要求
- AI基础数据服务厂商提供的专业产品与服务能够助力AI研发企业迅速获得所需的高质量标注数据

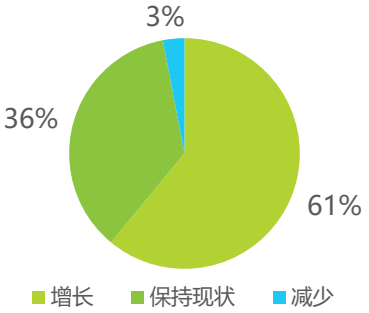
确保数据的高标准质量

- 数据质量对AI算法的性能有直接影响
- AI基础数据服务厂商依托专业的标注团队和行业领先的标注工具，确保了数据的高标准质量，为算法的精度和可靠性奠定了坚实的基础，帮助企业打造高性能的AI方案

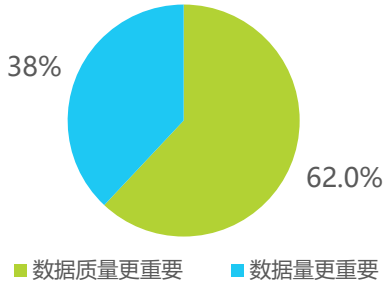
企业人工智能建设的预算分配情况



企业未来2~5年的训练数据需求情况



数据量与数据质量的重要性比较



来源：艾瑞咨询研究院自主研究及绘制。

来源：LXT-The Path to AI Maturity 2024

©2024.7 iResearch Inc. www.iresearch.com.cn

AI基础数据服务厂商及主要产品服务介绍

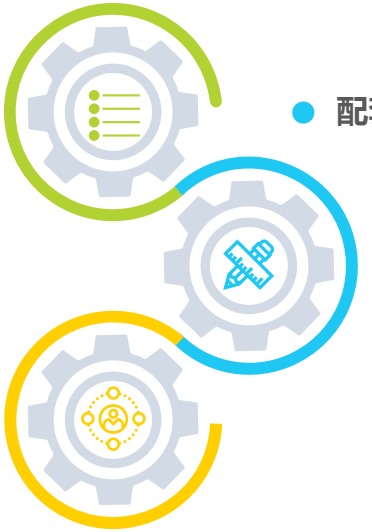
标准数据集、定制数据集、配套产品工具服务等三大产品服务

AI基础数据服务厂商是专注于为各行业的AI算法训练与调优提供基础数据产品服务的公司。这些公司通过提供标准数据集、定制数据集和配套产品工具服务，支持互联网、大模型、智能驾驶等各领域的AI技术发展。数据集按内容格式可分为文本、图像、视频、语音等类型，核心生产流程主要包括方案设计、数据采集、数据清洗、数据标注和数据质检等五个关键环节。标准数据集是由数据服务厂商研发并可多次销售的数据集；定制数据集是依据客户需求制作特定数据集，数据的知识产权归客户所有；配套产品工具服务包括标注工具、实训平台及AI模型评测等软硬件工具服务，用于满足高效标注数据、培训数据标注、评估AI能力效果等不同层次的客户需求，辅助和延展数据服务厂商的相关业务。

AI数据数据服务厂商的主要产品服务

标准数据集

由数据服务厂商研发并可多次销售的数据集



配套产品工具服务

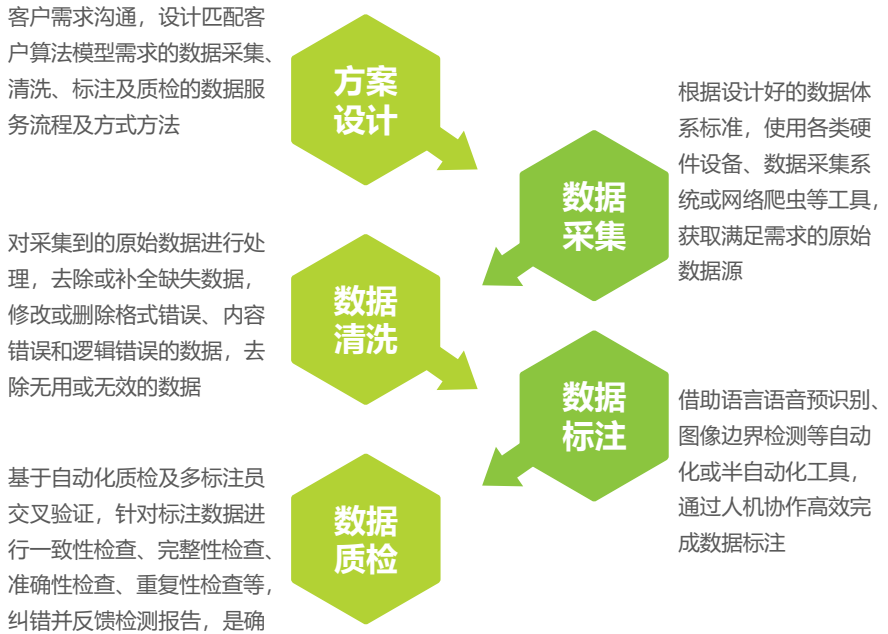
包括标注工具、实训平台及AI模型评测等软硬件工具服务，辅助和延展数据服务厂商的相关业务

定制数据集

依据客户需求制作特定数据集，数据的知识产权归客户所有

来源：艾瑞咨询研究院自主研究及绘制。

数据集的核心生产流程



来源：艾瑞咨询研究院自主研究及绘制。

©2024.7 iResearch Inc. www.iresearch.com.cn

典型服务场景——通用大模型（1/2）

数据量更大、维度更加多元，标注方式及质量评判标准也更为复杂多样

算法模型从理论到实践的应用过程依赖于大量的训练数据。训练数据越多、越完整、质量越高，模型推理的结果就越可靠。在本报告的讨论中，传统AI泛指Transformer架构出现之前的AI架构，参数量通常相对较小，大模型架构则以Transformer为代表。作为应用大模型架构的代表，ChatGPT在2022年11月上线以来，掀起了AI乃至社会经济各领域对大模型的研讨与应用的热潮。与传统AI相似，大模型依然需要大量优质数据，但其所需数据量更大，数据维度更加多元，标注方式及质量评判标准也更为复杂多样。

对比传统AI模型，大模型对数据集的需求差异

传统AI模型

- 传统AI模型由于参数量和复杂度的限制，能够吸收利用的数据量相对有限，过多的数据不仅无法有效利用，反而可能导致过拟合等问题
- 以计算机视觉的经典模型ResNet为例，其在2015年的ImageNet视觉竞赛中以3.6%的错误率夺得第一名，而其所用的ImageNet数据集有近150万张图像，总大小约150GB

- 传统AI模型通常需要针对目标任务场景的领域数据
- CNN主要处理图像数据，通常基于OCR、人脸识别、智能驾驶等特定任务场景的图像数据进行训练和优化；而RNN和LSTM则一般处理文本和时间序列数据

- 传统AI模型的标注维度通常比较单一
- 图像分类只需标注图片的类别，文本分类只需标注文本的主题等

- 传统AI模型对数据质量非常敏感，数据中的噪声和偏差可能会显著降低模型性能
- 传统AI模型的数据标注需要仔细审核，确保高准确度

- 传统AI模型的数据标注一般有标准答案，如图像类别、像素边界、语音文本等通常有单一答案，评判标准更客观

大模型

数据需求量更大

- 大模型通常需要更大量的数据才能训练出良好的性能，大模型原始训练数据的大小一般为TB至数百TB，但其训练首先需将文本等原始数据token化
- 今年4月开源的Llama3的训练数据集超过15T token，是Llama2的7倍

数据维度更加多元

- 大模型的数据来源非常丰富，涵盖了文本、图片、音频和视频等多种形式，含海量知识信息，涉及各类专业领域和多种语言。基于多样化的数据，大模型具备较强的通用能力和迁移能力，能够适应更广泛的任务和场景
- ChatGPT、Claude、Llama 和 Mistral 等大模型的训练数据包括文学作品、百科全书、新闻、社交媒体、学术文献等多种知识信息，且往往覆盖了图像、视频和音频等多模态数据

标注方式及评判标准更加复杂

- 大模型的标注需要考虑更加多维的信息，如新闻的标注除了包括主题之外，往往需包括时间、地点、人物等其他标签
- 为了训练大模型理解长序列数据的能力，还需要对文本进行更复杂的标注，例如对长篇小说进行按篇章结构或一定字数间隔的标注，标注每个板块的人物、事件、摘要等信息

标注维度更丰富

- 大模型能够在一定程度上从包含噪声和偏差的数据中学习
- 为了更好的模型性能，仍然需要对训练数据进行清洗和筛选，以获得更佳的模型效果

对噪声数据的利用度更高

- 大模型的标注有一定主观性，如长文本摘要、图片内容的理解、不同文风的改写、对同一问题的多个回答的打分等，评判标准更复杂，对标注者的逻辑能力、知识体系的要求更高
- 随着算法策略的调整或研发侧对数据工程理解的加深，数据标注方式及具体导向可能在项目进展中多次调整

评判标准更加复杂

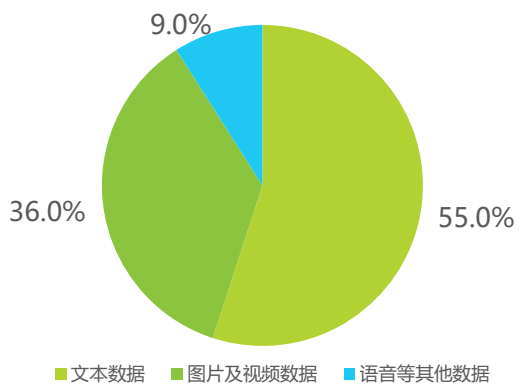
来源：艾瑞咨询研究院自主研究及绘制。

典型服务场景——通用大模型（2/2）

为提升通用能力，大模型训练数据的投入将逐步向图像、视频等多模态数据倾斜，且需要更多的采购数据支持

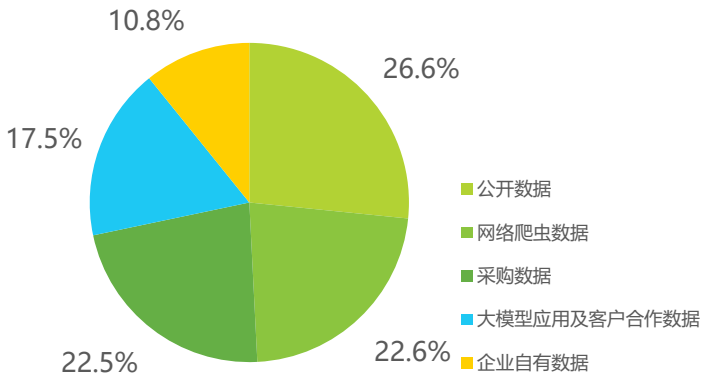
纵观业界开源及闭源大模型的能力特性，结合艾瑞对大模型研发企业的调研，虽然当下主流大模型应用仍相对侧重文本输入、文本输出的能力，但对图像、视频、语音等多模态数据的使用已越来越普遍，艾瑞预计大模型训练数据中多模态数据的占比将在未来数年持续提升。根据艾瑞对部分通用大模型及综合型AI厂商的调研，目前大模型的训练数据主要来源于公开数据、网络爬虫数据等可公开获取的数据，其次是采购数据。相比大模型初创企业，综合型AI厂商凭借现有的互联网应用和AI业务积累，具备独特的数据优势。在模型的通用能力建设方面，公开数据和爬虫数据已被广泛利用，未来这两类数据在整体上的提升空间相对有限，Epoch AI等机构的研究人员于2024年6月更新的论文中表示，大语言模型将在大约2026至2032年之间耗尽所有公开的文本数据。艾瑞预计，大模型研发厂商将通过更多的采购数据来提升模型的通用能力；而在垂直场景优化及行业客户的拓展中，公开数据和爬虫数据仍有较大的获取提升空间，大模型研发厂商也将更多地利用客户侧的合作数据，增强模型解决行业特定领域或企业特定问题的能力。

2023年大模型的各类型训练数据投入构成



来源：根据公开资料、企业调研，结合艾瑞统计模型核算。
其他说明：调研企业研发的大模型均为侧重语言能力的多模态大模型；以大模型研发企业在2023年对各类型数据的资金投入做占比计算；因调研样本的局限性，本比例可能与行业整体情况存在偏差。

2023年大模型的训练数据来源构成



来源：根据公开资料、企业调研，结合艾瑞统计模型核算。
数据类型说明：公开数据为无需借助爬虫工具，可直接下载利用的数据，如来自高校、社区的免费共享数据；大模型应用及客户合作数据，指用户在大模型C端应用中反馈的数据，以及大模型在B端行业拓展中企业客户提供的数据；外采数据包括原料数据以及数据服务公司提供的标准数据集、定制数据集等。
其他说明：调研企业研发的大模型均为侧重语言能力的多模态大模型；主要以数据token化前所需存储空间为口径做占比计算；因调研样本的局限性，本比例可能与行业整体情况存在偏差。

典型服务场景——大模型评测

公开评测基准与商业化评测服务共建大模型评测生态

随着大模型技术的快速迭代及其在众多领域的广泛应用，相关评测需求同步增长。对于模型研发企业，评测是发现模型在功能、性能、安全性和可靠性等方面优劣势的关键步骤，并可与其他企业的模型横向对比，进而针对性地优化模型，提高其表现和稳定性；对模型应用企业而言，评测是选型和项目验收的重要工具，通过专业评测服务，企业能够评估模型的实际应用适用性，确保所选模型满足需求，并保障定制类模型项目的交付质量。相较传统AI，大模型的应用空间更广，评测本身也更加复杂和多样化，市场对专业评测服务的需求潜力巨大。公开评测基准和商业化评测服务的发展，将为大模型评测提供重要支撑，促进技术与产业的健康发展。

公开基准为大模型评测提供重要参考

通过科学、客观、多场景的评测任务和指标设计，公开基准为学术研究和产业应用提供评估大模型能力的重要参考

类别	基准名称	发布机构/发布年份	评测内容
通用文本	MMLU	UC伯克利、哥大等/2020	15908个问题，覆盖基础数学、美国历史、计算机科学、法律等57个领域
	GPQA	纽约大学、Cohere、Anthropic等/2023	448个多项选择题，由生物、物理、化学等领域的专家编写
	Math	UC伯克利等/2021	12500个高中数学竞赛问题，覆盖代数、几何、概率论等学科
	HumanEval	OpenAI/2021	164个手写的编程问题，每个编程问题都由函数签名、文档字符串、函数体和几个单元测试构成
	其他典型通用文本类评测基准：MGSM、DROP、BBH等		
通用中文	SuperClue	AI评测基准社区Clue/2023	2194道多轮简答题，覆盖理科与文科两大能力，包括计算、逻辑推理、代码、知识百科等十大任务
	其他典型通用中文类评测基准：OpenCompass、CMMLU、C-EVAL等		
翻译	WMT23	国际机器翻译大会/2023	通用翻译、术语、手语、生物医学、文学等不同领域的翻译任务
语音	FlEURS	Meta、Google等/2022	包含102种语言的n路并行语音数据集，每种语言约12小时的语音监督数据
语音翻译	CoVoST2	Meta/2020	共计2900小时的语音，包含从21种语言翻译成英语，以及从英语翻译成15种语言的语料
多模态	MMMU	In.ai、滑铁卢大学等/2023	从大学考试、教科书中收集的 1.15万个多模态问题，包括图表、图示、地图、乐谱、化学结构等30种高度异构的图像类型
	MathVista	加州大学洛杉矶分校等/2023	由6141个任务组成，源自 28 个涉及数学的现有多模态数据集和 3 个新创建的数据集
	EgoShema	UC伯克利等/2023	由超过250小时的的人类自然活动的视频和超过5000个多项选择题构成，基准要求模型根据三分钟长的视频剪辑从5个选项中选出正确答案
	其他典型多模态评测基准：M3Exam、AI2D、ChartQA、DocVQA、ActivityNet等		

来源：艾瑞咨询研究院自主研究及绘制。

商业化评测为客户提供体系化服务

AI基础数据服务公司及评测平台公司可通过商业化评测，为客户提供体系化解决方案，推动大模型在实际应用中的落地和发展

数据集

高质量的数据集是进行有效评测的基础，在公开评测基准的基础上，商业化评测服务可结合私有或定制数据集，为客户提供符合实际场景需求的评测数据集和指标

体系平台

商业化评测服务提供自动化、智能化的平台，支持数据管理和更新，为客户构建高效、规范且可演进的评测体系，生成详细报告，助力技术迭代及应用选型，从供需两侧加速大模型产业的发展

典型服务场景——智能驾驶

AI基础数据服务与AI算法研发相互促进，共同推动着自动驾驶的实现

在大模型和端到端技术的加持下，智能驾驶的自动化程度不断提升，相关功能已成为部分消费者购车时的重要考虑因素。除个别厂商专注于纯视觉路线外，当下高级别的智能驾驶系统中，摄像头和激光雷达是两大核心传感器。摄像头主要捕捉二维图像，具有高分辨率和丰富的色彩细节；激光雷达则通过发射和接收激光脉冲生成高精度的三维点云数据，能够精确测量物体的距离、尺寸和相对位置，受光照等环境条件影响较小。摄像头和激光雷达等各类传感器各具优势，互为补充，数据标注需对来自不同传感器的数据标签对齐和交叉验证工作。AI基础数据服务是支撑智能驾驶、大模型等AI算法研发的基石，而AI算法也大幅提升了智驾研发领域数据标注的效率和效果，为数据服务行业的发展注入了新的活力。数据与AI彼此支撑、相互促进，共同推动着自动驾驶的实现。

智驾系统核心传感器的数据标注工作对比分析

架构	摄像头	激光雷达
标注对象	• 二维图像中的汽车、行人、交通标志、车道线等物体 • 需考虑光照条件和天气影响	• 3D点云数据，需标注物体的边界、相对位置等 • 相对不受光照条件影响
标注复杂性	• 需综合物体的颜色、纹理和形状等进行区分 • 标注的主观性或不确定性相对更大	• 需理解三维空间关系 • 精确的距离测量 • 标注结果的一致性更高
标注量	• 摄像头的数据及采集到的图像数量相对更多 • 每张图像的标注工作量相对更小	• 每帧点云的数据量很大，点云数据处理和标注的工作量相对更大
标注成本	• 图像标注相对简单，且相关自动化工具相对成熟，单张标注成本更低	• 点云数据复杂，单张标注成本相对更高
集成与融合	在高级别的自动驾驶系统研发中，大多厂商通常会融合摄像头、激光雷达等多种传感器的数据，为系统提供更全面的信息。这意味着标注策略需考虑数据融合，做好来自不同传感器的数据标签对齐和交叉验证工作	

来源：综合网络公开资料，艾瑞咨询研究院整理及绘制。

自动化标注在智驾研发场景中的应用案例

特斯拉

- 2021年特斯拉人工标注团队约1000人，其后通过自动化标注系统提高了团队效率，2022年该团队裁员200余人
- 特斯拉采用“多重轨迹重建”技术自动标注车辆行驶轨迹，在集群中运行12小时即可完成10000次行驶轨迹标注，相当于节省了500万小时的人工标注时间。

小鹏汽车

- XNet的训练基于50~100万个短视频，其中动态目标的数量可能达到数亿甚至十亿量级，如果以人工标注的方式，需要1000人的团队耗时2年完成标注
- 小鹏汽车的全自动标注系统仅需16.7天即完成上述工作，且标注质量更高，信息更全面，包括3D位置、尺寸、速度、轨迹等信息

理想汽车

- 2023年之前理想汽车每年需通过人工完成约1000万帧的图片标注，每张成本6~8元，一年耗资近亿元
- 此后，理想汽车基于大模型进行自动化标注，算法可在三个小时内完成过去人工一年的工作，效率是人工的1000倍

Scale AI

- Scale AI为通用、Nuro、丰田、法雷奥等诸多智驾研发企业提供了自动化标注的工具平台或相关产品服务
- 借助Scale AI的调试训练数据集的可视化工具平台Nucleus，无人车等机器人研发企业Nuro可有效维护管理超5亿张图像数据集。Nucleus的Object Autotag功能可支持Nuro选择某一类别的未标注图像，并自动找出一组相似图像，大幅提升罕见场景的数据的准备、标注及管理效率

来源：综合华泰证券、九章智驾等公开资料，艾瑞咨询研究院整理及绘制

02/ AI基础数据服务市场现状

中国AI基础数据服务产业图谱

多源数据、人力服务、IT设施 → 数据服务 → AI算法研发厂商

AI基础数据服务产业的中游即数据标注等数据服务的供应商，包括专业厂商及云厂商两类，其中后者以支持内部算法研发及云业务客户需求为主。上游提供原料数据、人力资源支持及IT基础设施，其中人力资源服务供应商主要包括垂直做数据标注的厂商和综合IT类厂商两类，目前业界通常采用远程线上服务即云BPO的模式进行人力支持。下游为数据服务的需求方，包括大模型、智能驾驶等各行业各领域投入AI算法研发的厂商。

2024年中国AI基础数据服务产业图谱

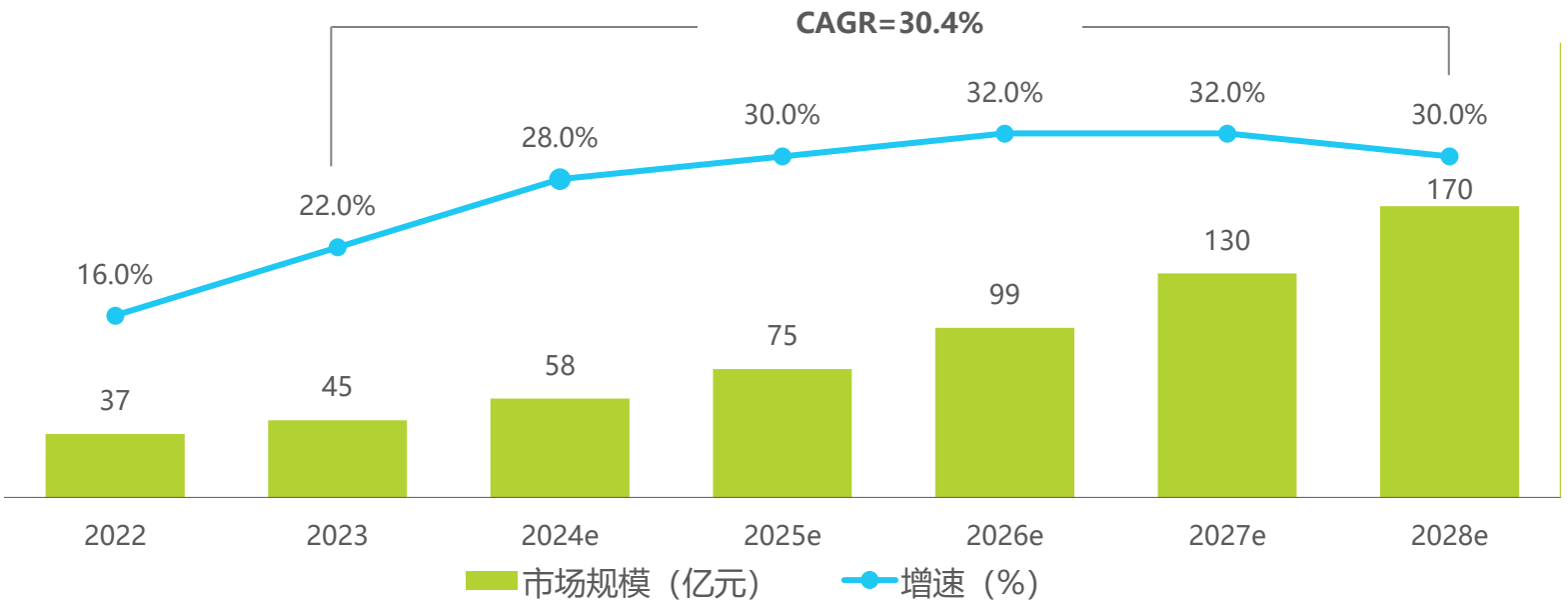
来源：艾瑞咨询研究院自主研究及绘制，图谱中所展示的公司logo顺序及大小并无实际意义。

中国AI基础数据服务市场规模

2023年中国AI基础数据服务市场规模45亿元，未来5年复合增长率30.4%

基于对数据服务专业厂商、云厂商、大模型研发厂商、智能驾驶研发厂商等中国AI基础数据服务市场的供需两侧企业调研，结合艾瑞对中国人工智能市场整体及AI基础数据服务市场的发展判断，艾瑞推算2023年中国AI基础数据服务市场规模为45亿元。在需求侧，随着AI算法研发从面向特定任务领域的小模型向具备更强通用泛化能力的大模型过渡，数据服务需求企业将产生大量高质量、多模态的数据需求。同时，随着大模型在通用及垂直场景中的应用拓展和智能驾驶等AI技术的规模化商业落地，良好的商业回报将进一步推动需求侧加大对基础数据的投入。在供给侧，随着数据要素等相关支持政策的持续深化，服务商将加快数据源的获取及数据集的制作。数据工程技术、数据标准规范、标注方法等日益成熟，人才生态及服务软件平台的自动化、流程化也在不断完善，供给侧的供应能力和服务质量得以加强。综合供需两侧的情况，艾瑞预计到2028年，中国AI基础数据服务市场规模将达170亿元，未来五年的复合增长率为30.4%。

2022-2028年中国AI基础数据服务市场规模



来源：根据公开资料、企业访谈，结合艾瑞统计模型核算。

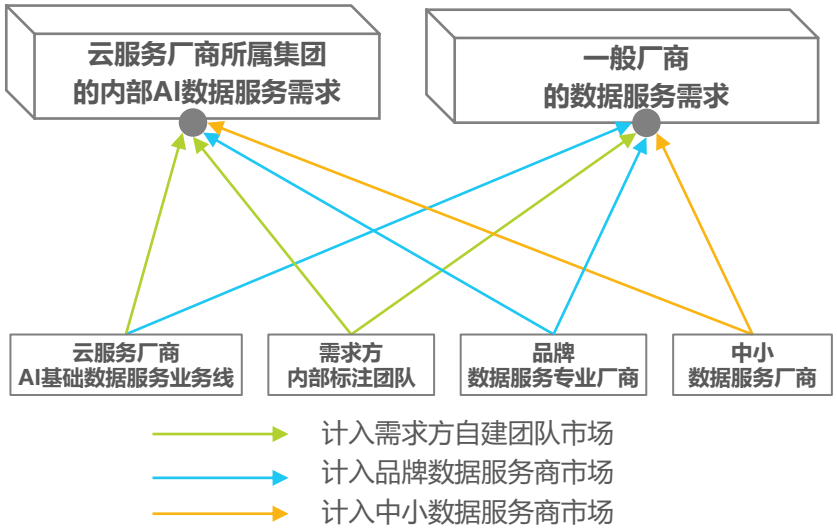
AI基础数据服务商的市场结构分析（1/2）

自建团队与品牌数据服务商主导市场，中小服务商的市场份额大幅下滑

延续艾瑞在2020年中国AI基础数据服务行业研究中的供给方划分方式，本报告将供给方分为需求方自建团队、品牌数据服务商、中小数据服务三类。其中，有AI基础数据对外服务的云厂商最为特殊，因其所属集团的内部AI算法研发所需的数据服务，可能由云服务业务线、算法研发业务线的内部标注团队，以及外部的品牌和中小数据服务商等四种团队承接。在艾瑞对供给方的市场份额统计中，云服务业务线的对内支持计入需求方自建团队的市场；因云服务厂商具备较大的市场影响力、相对完善的服务软件平台，将云业务线对外部厂商的数据服务计入品牌数据服务商的市场。

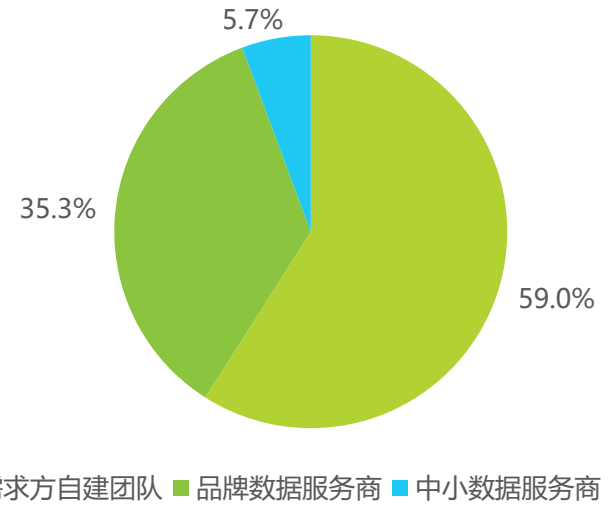
相比4年的市场份额情况，中小数据服务商的整体市场份额下滑约41%，需求方自建团队上升36%，品牌数据服务商上升5%：传统AI数据标注市场竞争激烈，而大模型、智能驾驶等新兴项目体量较大需要较强的综合服务能力，叠加疫情影响，较多中小数据服务商已退出市场；在大模型、智能驾驶等新兴AI算法及对应标注方式快速迭代时期，为追求更高的开发效率、保障信息安全，较多需求方通过自建团队满足数据服务需求；未来随着品牌数据服务商的数据版权的丰富、专业能力的提升、标注方法的成熟，品牌数据服务商将承接更多的数据服务需求。

AI基础数据服务产业的供需合作链条



来源：艾瑞咨询研究院自主研究及绘制。

2023年中国AI基础数据服务供给方的市场份额



来源：根据公开资料、企业访谈，结合艾瑞统计模型核算。

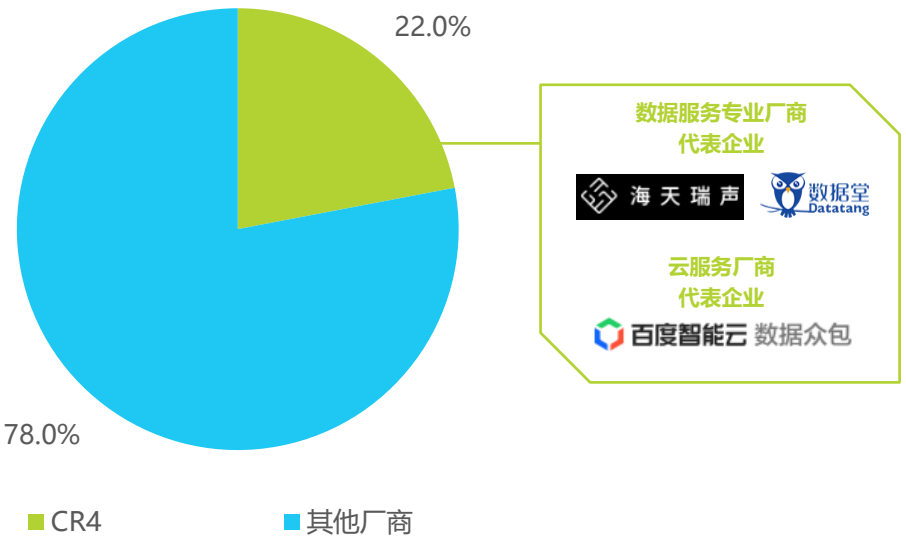
AI基础数据服务商的市场结构分析（2/2）

2023年CR4为22.0%，行业集中度相比2019年显著提升

据艾瑞调研统计，2023年中国AI基础数据服务行业的CR4（前四大企业的市场份额）为22.0%，市场仍较为分散。相比2019年14.3%的CR4，中国AI基础数据服务市场在2023年的集中度显著提升。市场份额位居前四的企业包括以海天瑞声、数据堂为代表的数据服务专业厂商以及以百度智能云为代表的云服务厂商。在传统AI标注市场的激烈竞争中，百度智能云、数据堂等AI基础数据服务企业敏锐的捕捉到了大模型标注的需求变迁，凭借强大的资源整合及项目管理能力、丰富的行业经验和专业理解，快速响应市场需求的变化，及时投入大模型相关产品和服务的研发，从而在AI基础数据服务的整体竞争中赢得了更高的市场份额，也成为了大模型标注领域的头部厂商。

展望未来，随着大模型等AI技术的发展，数据服务的需求日益庞大且复杂，这对服务企业的综合能力提出了更高的要求。没有自动化软件平台或平台能力较弱、资源整合能力有限的厂商将面临生存空间不断被挤压的困境；高质量数据版权丰富、运营管理能力强大、行业理解深刻的头部数据服务厂商有望持续提升市场份额。

2023年中国AI基础数据服务行业CR4及代表厂商



来源：根据公开资料、企业访谈，结合艾瑞统计模型核算；CR4为国内营收位居前四的企业的相关营收在中国市场的份额总和；图中所展示的公司logo顺序及大小并无实际意义。

厂商竞争要素与未来发展策略

自动化平台、深刻的行业理解、对技术与数据的前瞻性布局，将帮助优秀企业赢得市场领先

在行业集中度不断提升的过程中，基于自动化平台不断强化项目运营及资源整合能力、深刻理解行业需求，积极应用前沿算法、积累高质量数据集版权的AI基础数据服务厂商，将在激烈竞争的市场中脱颖而出，赢得市场领先地位。

AI基础数据服务厂商的竞争要素与未来发展策略

基于自动化数据服务平台的项目运营与资源整合能力

- **自动化与智能化平台**：一个高度自动化、智能化的服务平台能够高效地处理数据预处理、清洗、标注等各个环节
- **资源整合能力**：强大的资源整合能力使企业能够快速召集满足客户需求的数据服务人员，确保按时按质交付项目
- **精细化管理**：通过精细化管理标注工程师和标准质检员等项目人员，企业可以确保团队内的高效协作
- **人力资源支持**：随着平台功能的增强，更多个体可以直接通过平台为数据服务厂商提供人力资源支持，提升行业运转效率，扩大从业人员规模

深刻的行业理解与前瞻性布局

- **简化复杂需求**：大模型的标注工作复杂多样，供应商必须具备将复杂需求简化为具体标注任务的能力
- **前瞻性布局**：企业需具备前瞻性布局的能力，积极投入有前景的数据集开发，并应用前沿AI算法对平台进行自动化改造，使企业在技术变革中保持领先

原料数据的版权积累

- **定制数据集业务的局限性**：定制数据集因其个性化需求和难以重复售卖的特点，难以支撑数据服务企业的规模化增长
- **标准数据集的优势**：基于高质量原料数据制作的标准数据集则不同，这些数据集可以面向多个客户重复销售，市场需求广泛且客户接受度高，有助于企业实现规模效应，取得更高的利润水平

来源：艾瑞咨询研究院自主研究及绘制。

03/ AI基础数据服务厂商案例

海天瑞声

深耕行业近20年，向全行业提供多语言、跨领域、跨模态的人工智能数据及相关数据服务

北京海天瑞声科技股份有限公司（以下简称海天瑞声）自2005年成立以来，公司始终致力于为AI产业链上的各类机构提供算法模型开发训练所需的专业数据集。经过多年发展，公司已成为人工智能基础数据服务领域具有较强国际竞争力的国内头部企业，并实现了标准化产品、定制化服务、相关应用服务全覆盖。公司所提供的训练数据涵盖智能语音(语音识别、语音合成等)、计算机视觉、自然语言等多个核心领域，全面服务于人机交互、智能家居、智能驾驶、智慧金融、智能安防等多种创新应用场景。

海天瑞声产品服务及技术布局

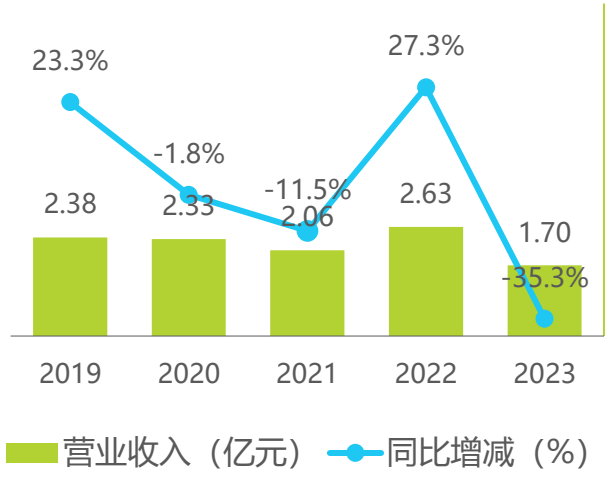
智能语音	公司通过设计、采集、加工、质检等智能语音训练数据集生产环节；或者针对客户提供的原料音频文件执行加工、质检工作，最终形成客户所需的智能语音训练数据集
计算机视觉	公司通过设计计算机视觉的训练数据集结构、采集、加工、质检；或者对客户提供的图像、视频文件执行加工、质检工作，最终形成客户所需的计算机视觉训练数据集
自然语言处理	公司通过设计自然语言处理的训练数据集结构、采集、加工、质检；或者对客户提供的自然语言文本执行加工、质检工作，最终形成客户所需的自然语言训练数据集
训练数据相关的应用服务	公司基于自身生产的训练数据提供算法模型相关的训练服务，运用训练数据研发能力助力下游客户完成其算法模型的语言拓展、特定算法模块拓展、垂直应用领域拓展等，为客户定制针对特定应用场景的专属算法模型，提高AI技术应用效果

海天瑞声客户场景及客户结构

客户场景

个人助手	语音输入	内容生成
智能家居	机器人	语音导航
智能客服	智能播报	语音翻译
移动社交	虚拟人	智能驾驶
智慧医疗	智慧教育	智慧交通
智慧城市	智慧金融	机器翻译
智能问答	信息提取	情感分析
OCR识别		

2019-2023年海天瑞声的营收情况



核心技术布局

- 通过持续的研发投入积累形成了12项核心技术，覆盖基础研究、平台工具、训练数据生产三个层次，应用于训练数据生产的设计、采集、加工、质检全流程
- 12项核心技术中，语音语言学基础研究、多语种多模态训练数据设计技术、数据同步技术、大数据驱动的高效数据处理技术、分布式高性能自动校验技术等5项具备较高技术壁垒

客户结构分析

- **Top 5**：2023年，海天瑞声Top 5客户销售额合计占比33.41%
- **境内/境外**：2023年，公司境内地区客户收入占比64.7%，境内收入额同比-25.2%；境外收入额同比-48.2%

营收变动分析

公司2023年营收有较大下滑，主要原因包括境外客户阶段性裁员、业务调整和预算释放放缓，导致境外收入大幅下滑；国内客户对研发投入持谨慎态度，预算和需求释放减缓，加上行业竞争加剧，导致境内收入下滑

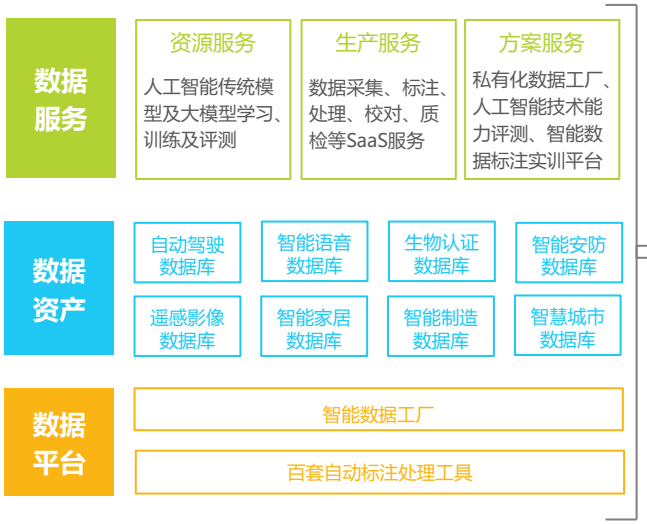
来源：综合企业财报、官网等公开信息，艾瑞咨询研究院整理及绘制。

数据堂

凭借高质量数据服务，数据堂已帮助全球上千家企业提升AI模型性能

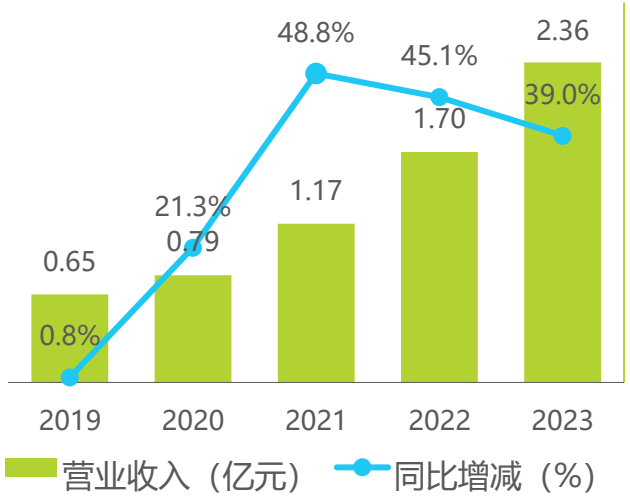
数据堂（北京）科技股份有限公司（以下简称数据堂）成立于2010年，是一家面向支撑人工智能产业发展，专业从事人工智能基础数据服务的企业。经过十余年积累，数据堂形成了数据多模态采集、自动处理、质量评测、安全计算的全链条核心技术体系及服务平台。数据堂专注于为国内外人工智能技术和应用客户提供一站式基础数据资源服务、基础数据生产服务以及基础数据处理解决方案服务，主要覆盖大模型、智能语音、自动驾驶、生物认证、智能安防、智能家居、智能娱乐、智慧城市、智能制造、智能医疗等领域。

数据堂产品服务及技术布局



数据堂的客户场景及客户结构

2019-2023年数据堂的营收情况



重点研发项目

- 数加加平台：旨在为项目提供自助化、自动化的高效处理流程的柔性生产系统。最大限度地提升供应商项目执行的效率和质量，并通过数智化和自动化的方式实现更好的业务运营和管理效果
- 数加价Pro：专为客户打造的一套数据标注生产线系统，旨在提供快速搭建数据标注生产线的解决方案

客户结构分析

- **Top 5**：2023年，数据堂Top 5客户销售额合计占比39.08%
- **境内/境外**：2023年，数据堂境内地区客户收入占比73.1%，境内收入额同比提升55.7%；境外收入额同比增长7.61%

营收变动分析

数据堂近几年收入大幅增长，主要原因是全球人工智能产业规模快速增长，AI技术的发展和迭代，导致对人工智能数据产品及解决方案的需求快速增长，国内收入的增长同时受益于国家层面对数据生产要素发展的重视

来源：综合企业财报、官网等公开信息，艾瑞咨询研究院整理及绘制。

活树科技

专注于多语言数据服务，为AI公司和科研机构提供高质量数据解决方案

活树科技（Lifewood）成立于2004年，是一家面向全球的多语言数据服务企业。活树科技专注于文本、图像、音频和视频数据的采集和标注，提供50+种语言的数据服务，助力AI算法的训练和优化。凭借二十年的行业经验，活树科技为AI公司、互联网公司及科研机构提供高质量、大规模、结构化的训练数据。活树科技的数据解决方案覆盖个人助手、语音输入、智能客服、智慧医疗、智慧教育、智慧交通、智慧城市、智慧金融、智能问答、信息提取、情感分析、OCR识别等多种应用场景。活树科技致力于推动AI技术的实践应用及商业化落地，赋能AI技术与实体经济深度融合。

活树科技全球人力资源布局--16国22交付中心

全球资源

助力企业出海

活树科技数据解决方案

业务布局及项目积累

- **16个** 国家
- **22个** 交付中心
- **53个** 语种数据
- **3,000+个**项目经验
  LLM项目落地全球**16个**国家

来源：综合官网等公开信息，艾瑞咨询研究院整理及绘制。

澳鹏

AI生命周期数据的创新和实践者

澳鹏（Appen）成立于1996年，公司总部位于澳大利亚，公司通过在美国、中国等国家的九个办事处和营业部为全球客户提供可靠的AI训练数据服务。澳鹏是AI生命周期数据的创新和实践者。凭借在数据获取、数据标注和模型评估方面超过25年的经验，澳鹏使组织能够推出具有创新性的人工智能数据系统。澳鹏的专业知识包括遍布全球170个国家/地区的70000多个地点的超过100万名精通290+种语言和方言，以及业界先进的人工智能辅助数据标注平台。澳鹏的产品和服务让技术、汽车、金融服务、零售和医疗保健领域的领导者有信心启动优秀的AI项目。

澳鹏产品服务及技术布局

数据集

- **成品数据集**：鹏提供600+个成品数据集，其中包括27600多小时的音频、490000多幅图像和超过一亿字/词的文本数据集，涵盖80种语言和多种方言
- **数据集应用场景**：安全驾驶/自动驾驶、互联网虚拟人/智能客服、智慧金融、智能家居、智能终端、智能安防

数据服务

- **数据采集**：拥有全球范围250+语言资源及100万众包团队，澳鹏提供全面的数据定制采集服务，为客户的AI部署提供高质量的数据支持
- **数据标注**：为客户提供多应用场景和行业的定制数据标注服务，为客户的AI应用提供全面数据

MatrixGo数据标注平台

高精度数据标注平台，使用专业多样的工具集创建高质量、精细化的数据，满足复杂的标注需求，基于自研AI算法大幅度提升标注效率

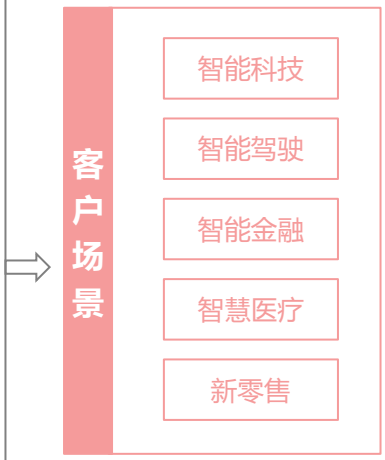
智能LLM开发平台

集大模型数据准备、训练、推理、部署应用于一体，提供数据集管理、数据标注、计算资源调度、模型评估、模型微调等全栈管理产品，助力企业轻松拥抱大模型。

核心技术布局

- 澳鹏力求通过技术和创新方案简化和自动化流程，从而能够大规模交付AI训练数据
- 澳鹏的工程、隐私和网络安全团队致力于确保数据可用性目标的实现，并确保数据的保护和安全
- 2023年投资0.35亿用于技术和系统建设，包括对ADAP的增强，以支持LLM产品，并更好地支持众包和客户

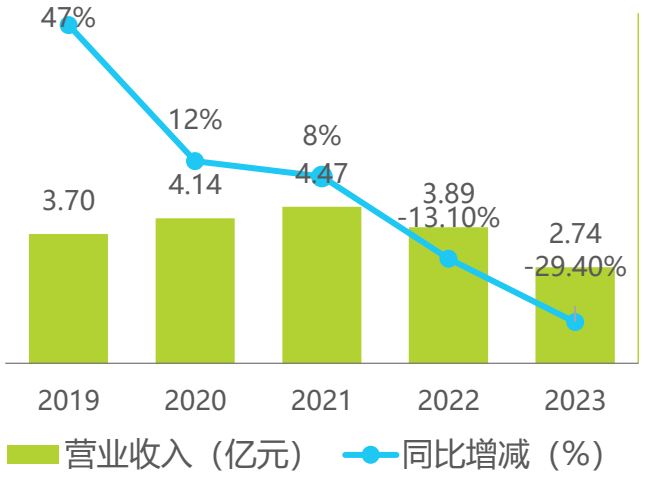
澳鹏的客户场景及客户结构



客户结构分析

- **Top 5**：2023年，奥鹏Top 5客户销售额合计占比74.8%
- **地区分布**：2023年，公司澳大利亚客户收入占比0.6%；美国客户收入占比80.5%，收入同比下滑35.3%；其他国家地区收入占比19.4%，收入同比增长6.6%

2019-2023年澳鹏的营收情况



营收变动分析

2023年，澳鹏营收同比下滑29.4%，主要受全球经济环境下行影响，客户支出缩减，导致澳鹏全球服务业务收入下降36.1%。
尽管如此，澳鹏全球服务业务的所有客户均已完成或正在进行生成式AI项目；同时，得益于中国市场、Quadrant和政府业务的贡献，澳鹏的新市场业务增长2.2%

来源：综合企业财报、官网等公开信息，艾瑞咨询研究院整理及绘制。

Scale AI

结合尖端技术与卓越运营，为客户提供机器学习全生命周期的端到端方案

Scale AI成立于2016年，总部位于美国。Scale AI的公司使命是加速人工智能应用的发展。Scale AI提供管理整个机器学习生命周期的端到端解决方案，将尖端技术与卓越运营相结合，帮助客户利用更好的数据更快地实现人工智能投资的价值。Scale AI通过结合机器学习驱动的预标注和主动工具，辅以不同程度和类型的人工审核，将原始数据转换成高质量的训练数据。截止目前，Scale AI已完成130亿次的标注，为超过8700万的2D及3D场景打上了标签。2024年5月21日，Scale AI宣布完成一笔10亿美元的融资，估值为138亿美元。

Scale AI产品服务及技术布局

构建AI

Scale 数据引擎

收集、整理、标注数据

- **生成式AI数据引擎**：快速创建经过审核的由领域专家编制的定制高质量数据集，以训练世界上最先进的模型
- **测试与评估**：持续测试和评估大型语言模型，识别风险，认证AI应用的安全性
- **公共部门数据引擎**：为国防、情报、民间机构的AI提供数据支持
- **智驾数据引擎**：支持L2-L5的自动驾驶系统开发

应用AI

Scale GenAI平台

支持客户构建、测试和优化可释放数据价值的生成式 AI 应用程序，借助Scale AI先进的RAG、测试评估平台以及ML专业知识，针对特定领域优化LLM的性能

Scale Donovan

服务于国家安全的人工智能数字参谋

核心技术布局

- 结合了机器学习驱动的预标注和多层次人工审核的数据引擎，可将原始数据高效的转换为高质量的训练数据；数据引擎可智能管理和优化数据集，识别高价值数据进行标注，最大化标注投资回报率；数据测试、模型评估和比较等工具，帮助客户充分利用数据资产
- 生成式AI、基于人类反馈的强化学习（RLHF）

服务的行业及客户案例

行业及客户

美国政府

美国陆军、美国空军
国防创新单元
首席数字和人工智能办公室

生成式AI

Open AI、Cohere
Adept

其他企业

Microsoft、Meta、GM
Nvidia、GAFG、Chegg
Howard Hughes、BCG

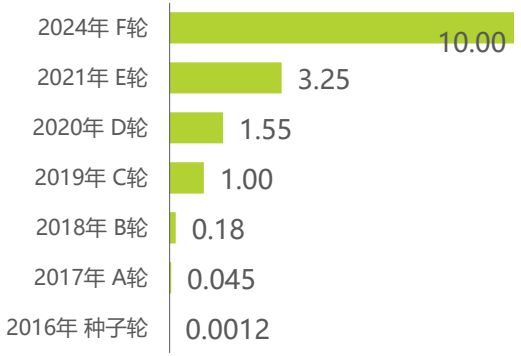
典型客户及行业应用案例

- **智能驾驶**：Scale AI的自动驾驶数据引擎推动了L4级自动驾驶的突破
- **国防**：Scale AI的公共部门数据引擎推动了美国国防部的许多重大AI项目
- **OpenAI**：Scale AI与OpenAI在GPT-2上合作进行了首批RLHF实验，并将这些技术扩展到InstructGPT等更多模型上

Scale AI的融资情况

16亿美元 累计融资额

130亿美元 估值



来源：综合企业官网等公开信息，艾瑞咨询研究院整理及绘制。

04/ 行业面对的挑战与机遇

AI基础数据服务行业面对的挑战与机遇

由于需求量大且需求复杂，行业面对人力短缺、项目难管理等挑战

由于大模型对数据集的要求更加复杂、高质量数据需求的增加，以及需求方对数据安全及保护核心技术的重视，AI基础数据服务行业面临诸多挑战，包括数据标注工程师的门槛提升、项目管理复杂性增加、项目规模大、高质量数据获取困难、信息安全问题等。尽管面对挑战，行业也迎来了新的机遇。大模型等AI技术的快速发展带来了高涨的数据需求，推动了AI基础数据服务市场的增长，高质量数据集成为供应商的核心竞争力，此外，多模态数据集的需求也将增加。凭借精细的流水分工和日益精准的AI算法，数据服务软件平台在行业中的价值不断提升，平台可帮助服务方更好的满足需求方的高质量数据需求，应对好人力及项目管理方面的挑战。

AI基础数据服务行业面对的挑战

数据标注工程师的从业门槛提升

大模型对数据集的评判标准更加复杂，对标注者的逻辑能力、知识体系的要求更高，对从业者的专业背景或学历水平提出了更高的要求，部分项目面对人力短缺

项目管理的复杂性增加

标注的方式方法欠缺统一客观标准，标注方式或评估标准的细节在项目过程中多变，需要服务方在项目进展中与需求方持续沟通、对标注人员持续培训拉齐

项目规模大

大模型对数据量有更高要求，数据服务厂商单项目要处理的数据体量大幅增加，进一步凸显人力短缺及项目运营管理上的挑战

信息安全问题

出于数据安全、保护核心技术等考虑，部分需求方选择通过自建标注团队的方式满足大模型、智能驾驶等前沿AI技术研发所需的数据集需求，这种方式一定程度上限制了数据服务公司乃至行业整体的的专业化和规模化发展

高质量数据获取困难

目前大模型训练已利用较多公开数据，为进一步提升模型的通用化及垂直领域能力，将需要更多专业领域的高质量数据，包括多语种的专业书籍、文献期刊、深度媒体报道，各类优质的影像及音视频作品等，但受版权政策或授权模式不明朗的限制，相关数据的获取较为困难

来源：艾瑞咨询研究院自主研究及绘制。

AI基础数据服务行业的发展机遇

蓬勃的数据需求

通用及垂直大模型、智能驾驶及各行业场景的AI技术研发与应用，伴随着高涨的AI数据服务需求

高质量数据集

独有的高质量数据是数据服务厂商的核心竞争力之一，竞争优势与相关数据集的种类和数据量正相关，且标准数据集可多次售卖，提升数据服务的毛利。数据服务厂商需在政策及相关数据共享平台的支持下，努力拓展更新数据集资源

多模态数据集

目前大模型的能力构建主要基于文本数据，伴随需求方对于大模型多模态能力的强化，图片、视频、音频等多模态数据的需求与之提升

数据服务软件平台

凭借精细的流水分工，数据服务平台可以高效响应大规模数据标注需求，且结合日益精准的AI算法，不断提升人工与平台整体对数据的清洗预处理、标注及质检审核的效率。在满足需求方对高质量数据要求的同时，提升了数据服务方在应对人力及项目管理等方面的挑战的能力

来源：艾瑞咨询研究院自主研究及绘制。

BUSINESS
COOPERATION

业务合作

联系我们

400 - 026 - 2099

ask@iresearch.com.cn

www.idigital.com.cn　　www.iresearch.com.cn

官　网

微信公众号

新浪微博

企业微信

LEGAL STATEMENT

法律声明

版权声明

本报告为艾瑞数智旗下品牌艾瑞咨询制作，其版权归属艾瑞咨询，没有经过艾瑞咨询的书面许可，任何组织和个人不得以任何形式复制、传播或输出中华人民共和国境外。任何未经授权使用本报告的相关商业行为都将违反《中华人民共和国著作权法》和其他法律法规以及有关国际公约的规定。

免责条款

本报告中行业数据及相关市场预测主要为公司研究员采用桌面研究、行业访谈、市场调查及其他研究方法，部分文字和数据采集于公开信息，并且结合艾瑞监测产品数据，通过艾瑞统计预测模型估算获得；企业数据主要为访谈获得，艾瑞咨询对该等信息的准确性、完整性或可靠性作尽最大努力的追求，但不作任何保证。在任何情况下，本报告中的信息或所表述的观点均不构成任何建议。

本报告中发布的调研数据采用样本调研方法，其数据结果受到样本的影响。由于调研方法及样本的限制，调查资料收集范围的限制，该数据仅代表调研时间和人群的基本状况，仅服务于当前的调研目的，为市场和客户提供基本参考。受研究方法和数据获取资源的限制，本报告只提供给用户作为市场参考资料，本公司对该报告的数据和观点不承担法律责任。

合作说明

该报告案例章节包含部分企业的商业展示，旨在体现行业发展状况，供各界参考。

iResearch
艾瑞咨询

THANKS

艾瑞咨询为商业决策赋能