

# 2024年中国AI基础数据服务研究报告

---

01 AI基础数据服务行业概述

---

02 AI基础数据服务市场研究

---

03 AI基础数据服务厂商案例

---

04 AI基础数据服务行业面临的挑战与机遇

# 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，使大模型可在端侧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。

# 数据、算法、算力是构建AI的三大要素

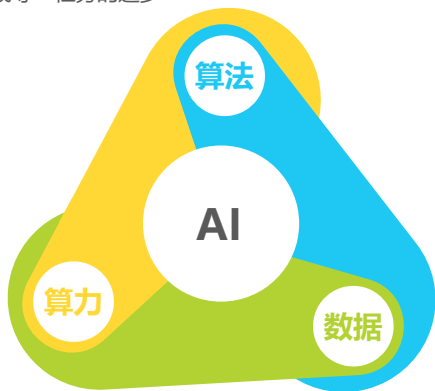
## 数据、算法、算力的协同促使现代AI技术实现了从理论到应用的飞跃

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 iResearch 艾瑞咨询

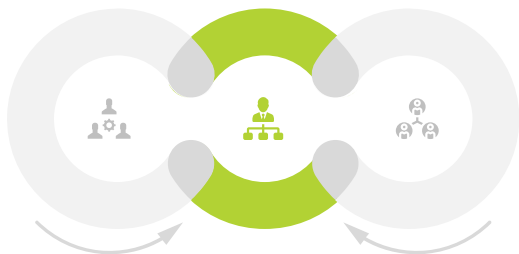
## 加速高质量数据的获取与标注，推动AI算法的创新与持续优化

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



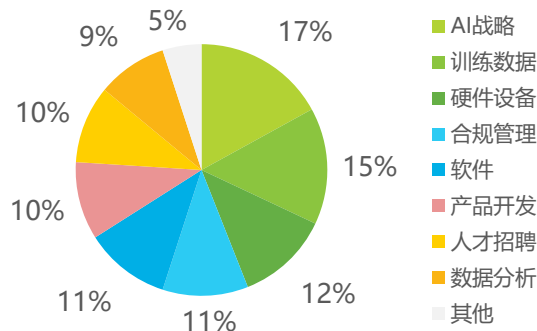
#### 加速数据获取与标注

- 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

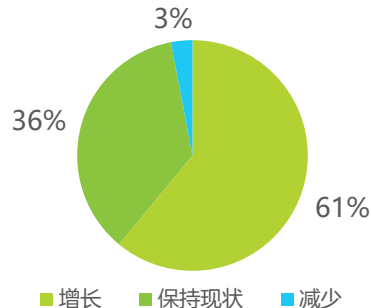
#### 确保数据的高标准质量

- 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

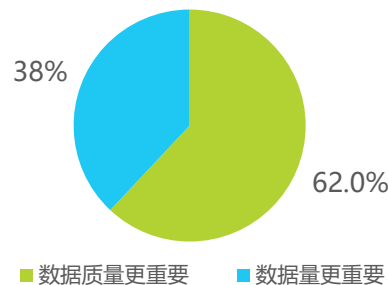
### 企业人工智能建设的预算分配情况



### 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

# AI基础数据服务厂商及主要产品服务介绍 iResearch 艾瑞咨询

## 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



#### ● 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



#### ● 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



### 数据集的核心生产流程

客户需求沟通，设计匹配客户算法模型需求的数据采集、清洗、标注及质检的数据服务流程及方式方法



根据设计好的数据体系标准，使用各类硬件设备、数据采集系统或网络爬虫等工具，获取满足需求的原始数据源



对采集到的原始数据进行处理，去除或补充缺失数据，修改或删除格式错误、内容错误和逻辑错误的数据，去除无用或无效的数据



基于自动化质检及多标注员交叉验证，针对标注数据进行一致性检查、完整性检查、准确性检查、重复性检查等，纠错并反馈检测报告，是确保数据质量的重要环节



借助语言语音预识别、图像边界检测等自动化或半自动化工具，通过人机协作高效完成数据标注



# 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
- 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
- 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

标注维度更丰富

对噪声数据的利用度更高

评判标准更加复杂

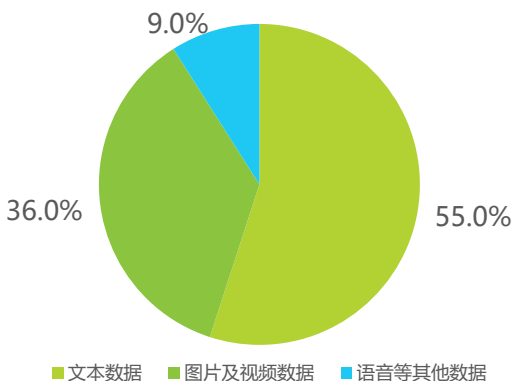


# 典型服务场景——通用大模型 (2/2)

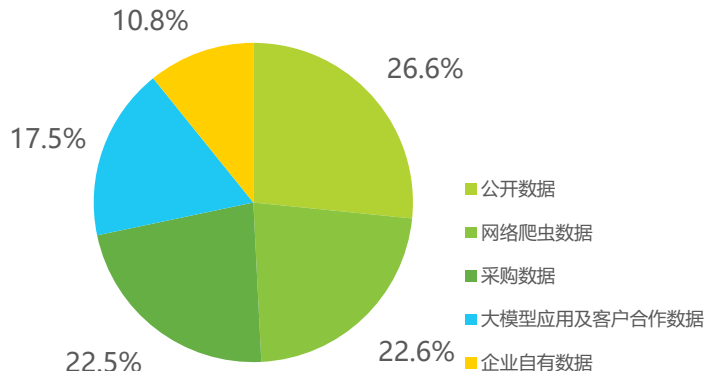
## 为提升通用能力，大模型训练数据的投入将逐步向图像、视频等多模态数据倾斜，且需要更多的采购数据支持

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年大模型的各种类型训练数据投入构成



### 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。

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。

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	FIEURS	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	<ul style="list-style-type: none"><li>二维图像中的汽车、行人、交通标志、车道线等物体</li><li>需考虑光照条件和天气影响</li></ul>	<ul style="list-style-type: none"><li>3D点云数据，需标注物体的边界、相对位置等</li><li>相对不受光照条件影响</li></ul>
标注复杂性	<ul style="list-style-type: none"><li>需综合物体的颜色、纹理和形状等进行区分</li><li>标注的主观性或不确定性相对更大</li></ul>	<ul style="list-style-type: none"><li>需理解三维空间关系</li><li>精确的距离测量</li><li>标注结果的一致性更高</li></ul>
标注量	<ul style="list-style-type: none"><li>摄像头的数据及采集到的图像数量相对更多</li><li>每张图像的标注工作量相对更小</li></ul>	<ul style="list-style-type: none"><li>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</li></ul>
标注成本	<ul style="list-style-type: none"><li>图像标注相对简单，且相关化工具相对成熟，单张标注成本更低</li></ul>	<ul style="list-style-type: none"><li>点云数据复杂，单张标注成本相对更高</li></ul>
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

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/265101201233011313>