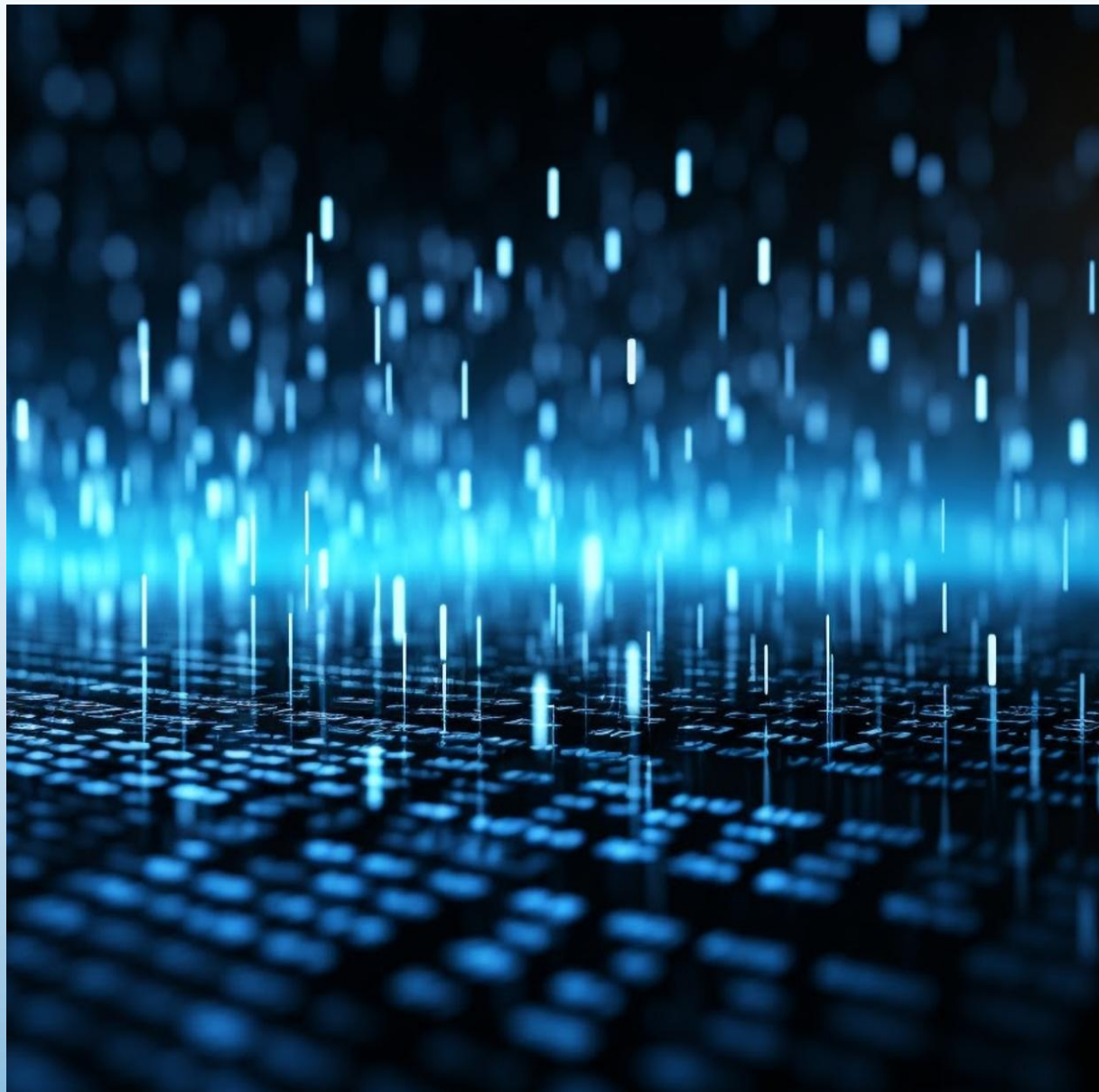




亿欧智库 <https://www.iyiou.com/research>

Copyright reserved to EO Intelligence, July 2025

2025中国语料市场发展及榜单报告

**亿欧
智库** EO Intelligence

研
究
报
告

序言

- ◆ 人工智能的浪潮正以前所未有的速度重构全球技术版图，而语料——这一滋养智能系统的“数据血脉”，已成为各国竞逐AI话语权的战略基石。中国语料市场在政策、技术与需求的三重共振下，正经历从规模扩张向质量跃迁的关键转型。2023年我国AI语料市场规模达68.7亿元，预计2025年将历史性突破百亿大关，达到109亿元，年复合增长率超25%。这一爆发性增长背后，是深度学习与多模态技术的迭代催生了更精细的语料需求，也是智能客服、自动驾驶、医疗诊断等60余个应用场景对高质量数据的迫切渴求。
- ◆ 然而，繁荣之下暗流涌动。中文高质量语料的结构性短缺，仍是悬于产业头顶的“达摩克利斯之剑”。国际主流大模型的训练语料中，中文占比不足千分之一，远低于英文的90%以上份额，迫使国内研发者长期依赖外文数据集或网络爬取数据。更深层的挑战在于：典籍文献的数字化滞后（国家图书馆仅4%古籍完成深度数字化）、语料标注标准的分歧、数据孤岛与版权壁垒，以及生成式AI“幻觉”对语料纯净度的反噬，共同构成了产业跃升的梗阻。
- ◆ 面对困局，一场自上而下的变革已然启动。政策端，国家发改委联合多部门发布《促进数据产业高质量发展的指导意见》，明确支持高质量数据集开发；教育部牵头建设“中华文脉新型语料库”，推动跨机构语料资源共享。市场端，上海人工智能实验室发起全国首个大模型语料联盟，深圳、上海数据交易所开辟语料交易专区，加速构建开放生态。企业亦积极破壁，中文在线深耕文学与教育语料库，华策影视激活海量视频素材，视觉中国以5亿图文对夯实视觉语料基座；库帕思语料平台则聚焦跨行业多场景语料融合，通过AI驱动的动态清洗与标注技术构建高精度工业级语料池——从文本、图像到视频，多元主体正合力打通“收集-处理-共享”的全链路闭环。
- ◆ 展望未来，中国语料市场的竞争维度正从“数量”转向“价值”。多模态融合将推动语料形态向图文视一体化演进；合成数据技术有望缓解稀缺领域语料困境；而全球化语料布局更将成为中国企业出海的关键筹码。但这一切的前提，是建立“安全与创新并重”的治理范式——唯有在数据主权、隐私保护与标准化建设间找到平衡点，中文语料才能真正成为人工智能文明的“源头活水”，而非受制于人的技术附庸。本报告梳理了语料基础概念与发展概况，展示了“2025语料风云榜”，并介绍了相关优秀企业案例，以期为广大从业者和各方帮助人士提供有益帮助，促进语料从收集到应用的大发展。

目录

CONTENTS

1 语料市场概况

- 1.1 人工智能语料发展历程
- 1.2 语料发展中外对比
- 1.3 中国人工智能语料主要阻碍

2 语料市场生态发展

- 2.1 语料生态三方分析
- 2.2 语料加工解决方案
- 2.3 语料平台建设路径

3 语料风云榜榜单

- 3.1 评选方式与规则
- 3.2 语料风云榜转型典范Top 10
- 3.3 语料风云榜新锐势力Top 10

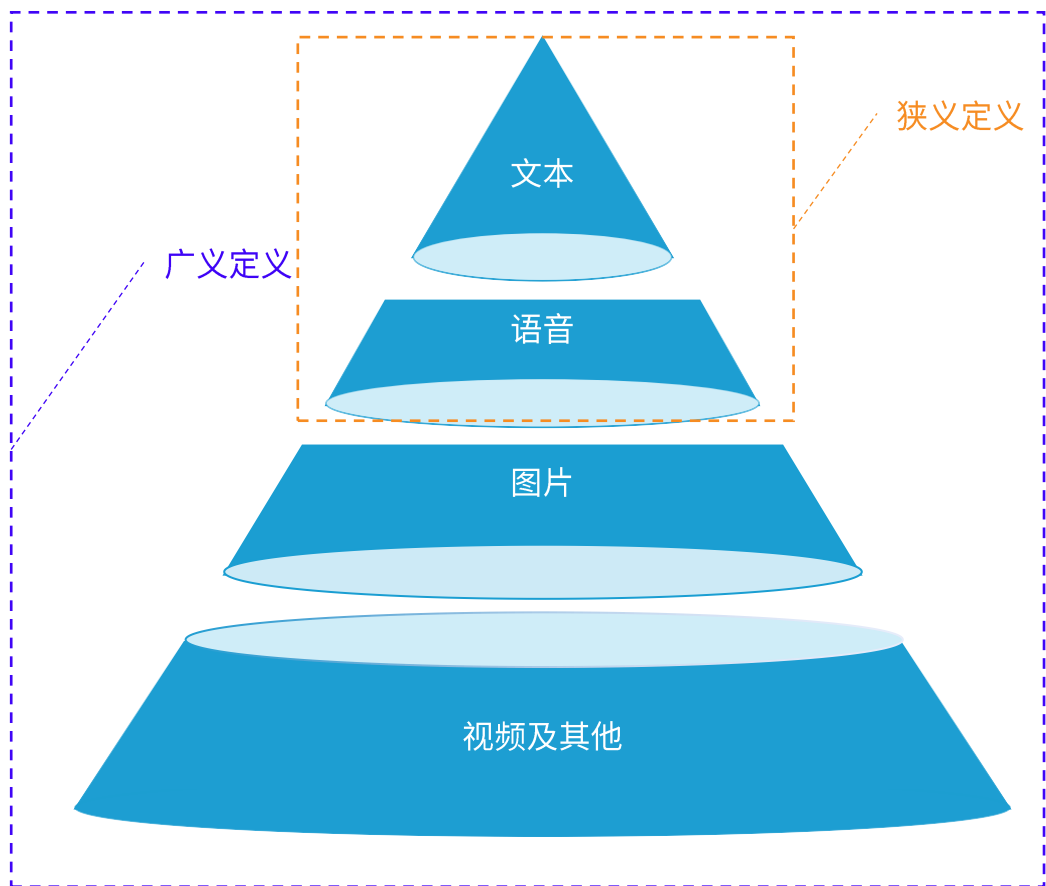
4 语料优秀案例分析

- 4.1 中文在线
- 4.2 中启易联
- 4.3 OpenDataLab
- 4.4 Scale AI
- 4.5 CloudFactory

1.1 语料数据的定义概念

- ◆ 语料数据定义：指用于开发和训练人工智能系统的文本或语音数据。然而，在广义的人工智能和自然语言处理领域，图片以及其他形式的数据（如视频等）也可以被视为一种语料。
- ◆ 语料处理方式：对于文本语料进行分词、构建词汇表，并通过自监督学习任务（如掩码语言模型）进行预训练，音频语料则需要特征提取（如梅尔频谱图），通过类似自监督任务（如掩码音频模型）进行预训练，图片语料的处理包括图像增强、标注（如对象检测、分类），然后使用卷积神经网络（CNN）等架构进行训练，与文本和音频不同，图片语料处理侧重于像素级特征提取和空间信息的学习。
- ◆ 高质量的语料是构建大模型的核心能力：高质量语料在提升模型能力方面具有关键作用，它能够准确模拟真实世界，使模型的预测更贴近实际数据分布，借助优化算法减少训练中的损失函数，从而提高模型的精确性和稳定性。另外，由于高质量数据包含丰富且可靠的信息，经过清洗后能消除噪音和错误，增强了训练过程的稳定性。此外，高质量语料数据的多样性有助于降低人工智能模型对特定数据集的依赖性，增强其鲁棒性和泛化能力，减少偏差和错误。因此，一个高质量的语料需要具备多样性、准确性、大规模、干净和一致性等特征。

亿欧智库：语料数据定义



数据来源：亿欧智库《2024中国AIGC商业潜力报告》

1.1 人工智能语料发展历程

人工智能语料作为自然语言处理与大模型训练的核心资源，其发展历程与技术进步密切相关。从20世纪60年代的布朗语料库到如今多模态TB级数据集，语料规模、类型和应用场景不断扩展，驱动了AI技术的飞跃。

人工智能语料的发展与计算语言学及AI技术的演进同步，经历了从手工收集到自动化、多模态的几个关键阶段：

亿欧智库：大模型语料四大阶段

20世纪60-80年代：早期语料库诞生

人工智能概念于1956年提出，但早期语料研究主要集中在计算语言学领域。1967年，美国布朗大学创建了布朗语料库（Brown Corpus），包含约100万词的英语文本，标志着系统化语料库的开端。这一时期，语料库规模较小，主要用于语言学研究，数据多为手工收集和标注。

20世纪末-21世纪初：语料库规模化与多样化

随着互联网的兴起，语料来源从书籍、报刊扩展到网页、论坛等。1990年代，WordNet、Penn Treebank等语料库推动了统计语言模型的发展。中文语料库如北京大学CCL语料库、Lancaster中文语料库开始出现，但规模和标注质量有限。这一阶段，语料库开始服务于机器翻译、语音识别等AI应用。

2010年代：深度学习驱动的语料增长

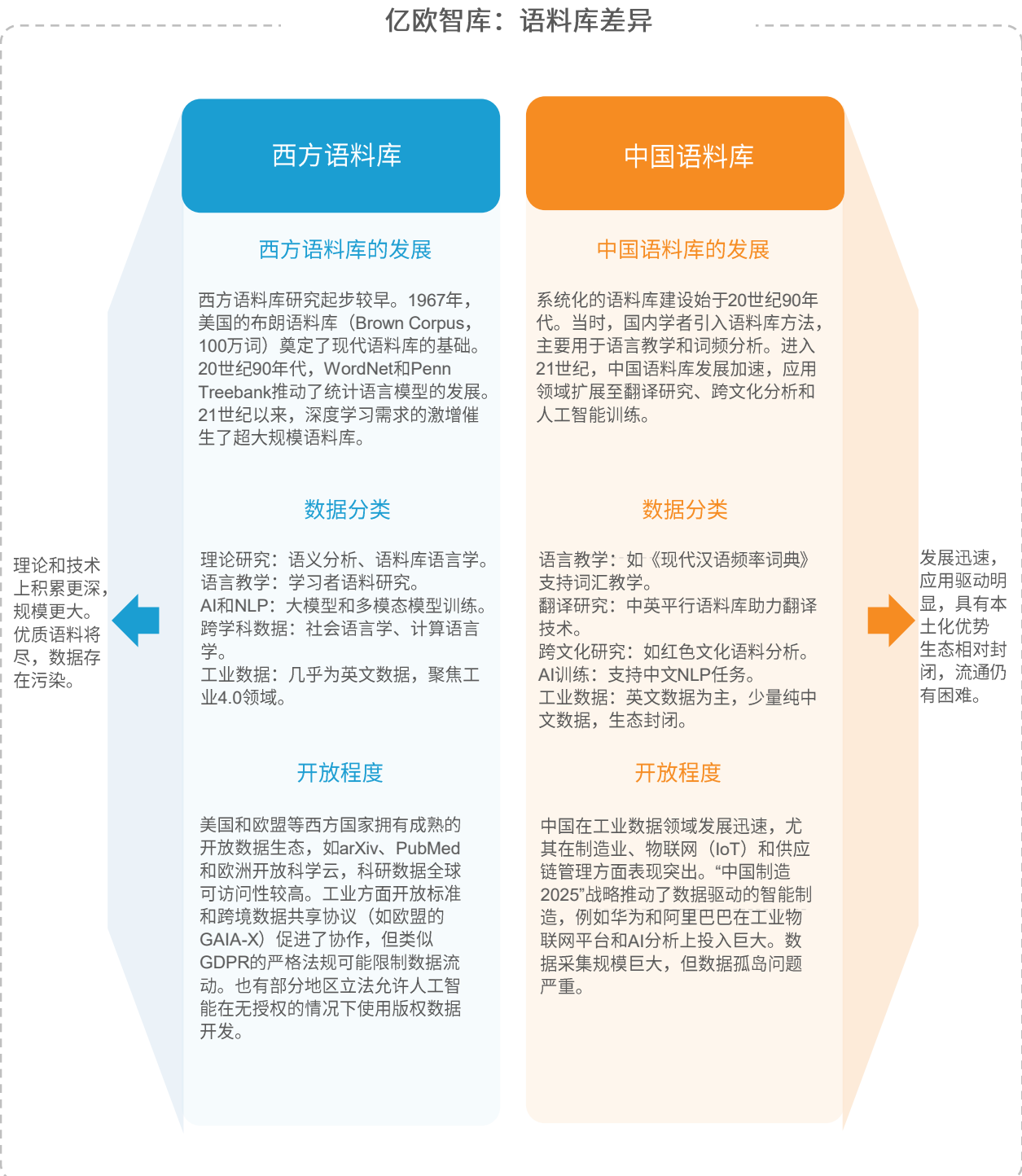
深度学习技术的突破对语料需求激增。2013年，Google发布Word2Vec模型，依赖大规模无标注语料训练词向量。开源语料库如Wikipedia、Common Crawl成为NLP研究的核心资源。中文语料方面，百度、腾讯等企业开始构建大规模网络语料库，CLUE等开源数据集推动了中文NLP发展。

2020年代至今：优质大模型语料走向枯竭

ChatGPT等大语言模型的成功将语料需求推向新高度。2020年后，语料从单一文本扩展到多模态（文本、图像、音频、视频）。然而互联网可公开访问的高质量文本（如新闻、论坛）增长速度放缓，难以满足大模型需求。网络爬取数据中广告、垃圾内容占比高，清洗成本激增。优质语料可能在2026-2028年间耗尽。

1.2 语料发展中外对比

中国和西方语料库发展各有优势与短板。西方语料库起步早，规模庞大，技术先进，开源生态成熟，应用广泛且理论基础深厚；而中国语料库起步较晚，规模较小，以应用驱动为主，注重本土化需求，但开源生态和数据质量有待提升。西方面临版权与隐私挑战，中国则需解决数据开放生态问题。未来，双方需依托技术创新与国际合作推动语料库发展。



1.3 中国人工智能语料主要阻碍

◆ 在大型人工智能模型的开发过程中，语料数据贯穿始终，是推动模型从概念到实际应用的核心要素。在前期设计阶段，语料帮助开发者明确目标和方向；在训练阶段，大量标注的语料数据是模型学习的基石；在评测阶段，代表性的测试集确保模型的实用性和可靠性；最后在迭代阶段，数据用于细化和优化模型，应对现实世界的复杂需求。因此，高质量和大量的语料数据不仅支持模型的开发，更是确保其成功部署和长期有效性的关键。



目录

CONTENTS

1 语料市场概况

- 1.1 人工智能语料发展历程
- 1.2 语料发展中外对比
- 1.3 中国人工智能语料主要阻碍

2 语料市场生态发展

- 2.1 语料生态三方分析
- 2.2 语料加工解决方案
- 2.3 语料平台建设路径

3 语料风云榜榜单

- 3.1 评选方式与规则
- 3.2 语料风云榜转型典范Top 10
- 3.3 语料风云榜新锐势力Top 10

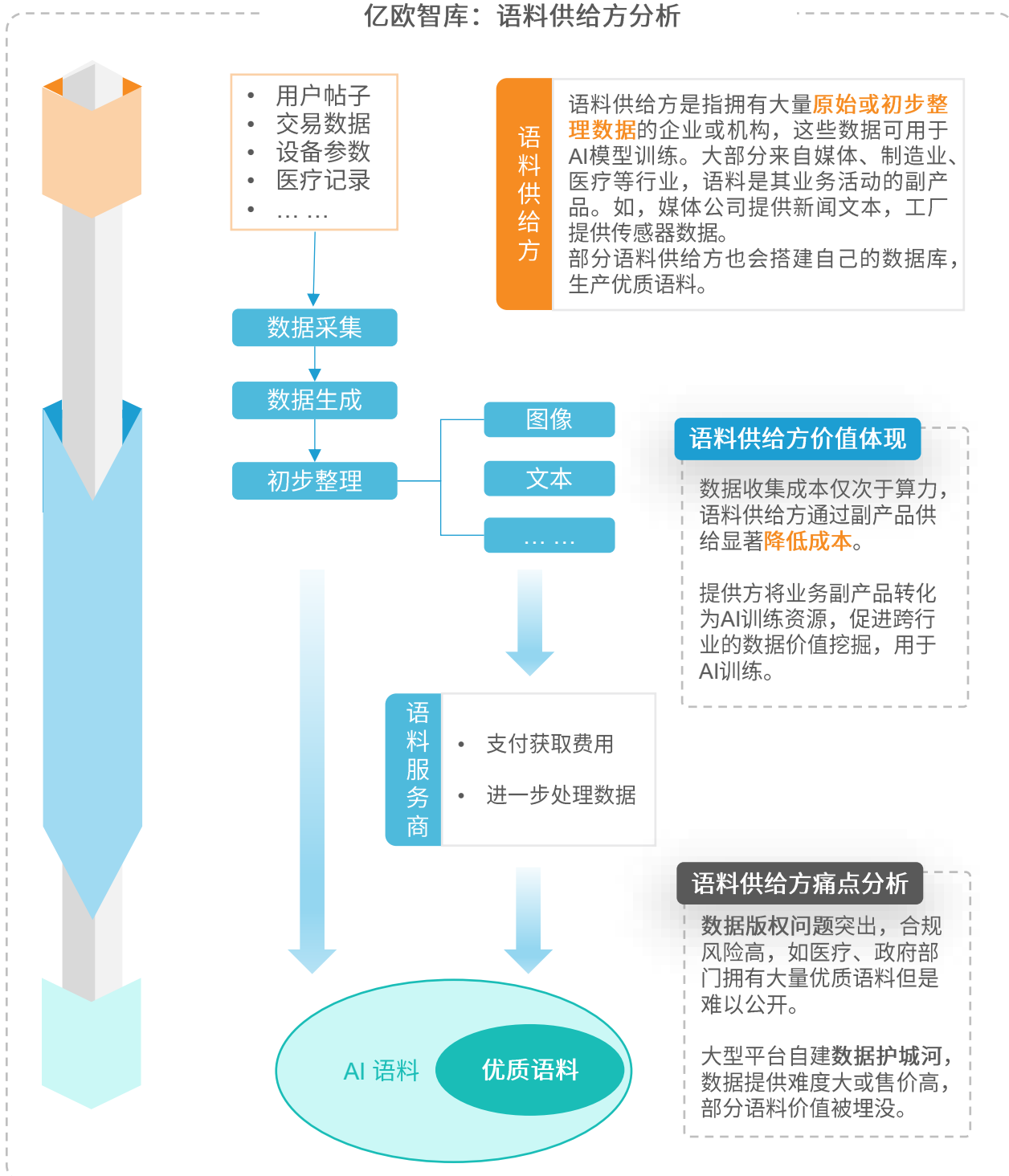
4 语料优秀案例分析

- 4.1 中文在线
- 4.2 中启易联
- 4.3 OpenDataLab
- 4.4 Scale AI
- 4.5 CloudFactory

2.1 语料供给方

- ◆ 语料供给方是AI模型训练的数据源头，通常来自媒体、工厂等行业，语料为其业务副产品。它们负责采集、生成和初步整理结构化、半结构化及非结构化数据，提供的内容包括通用数据集、行业特定数据及数据访问接口。其价值在于为AI研发提供多样化数据、降低成本、促进行业应用。

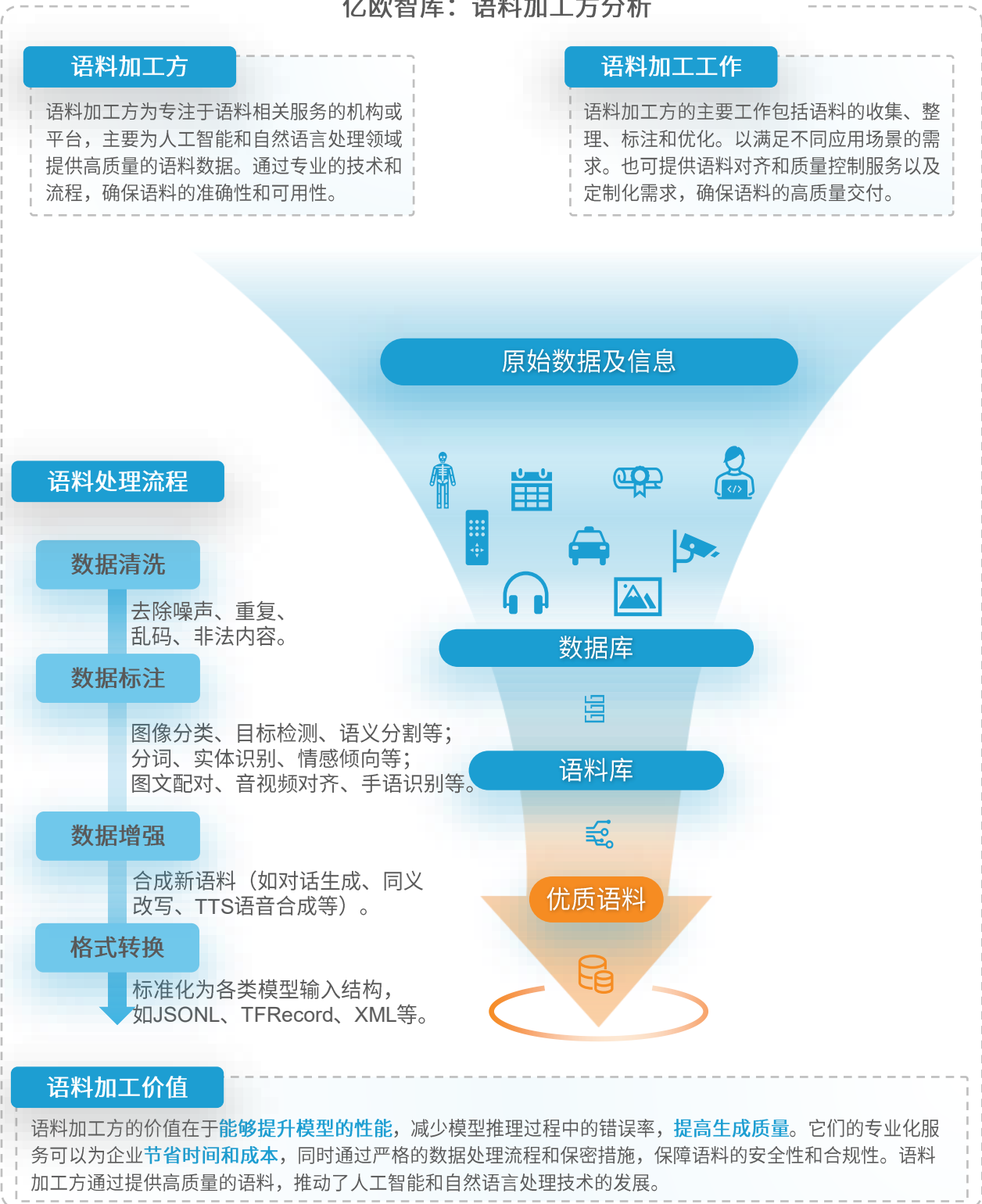
亿欧智库：语料供给方分析



2.1 语料加工方

◆ 语料加工方是专注于为人工智能和自然语言处理领域提供高质量语料及相关服务的专业机构。通过收集、整理、标注和优化语料，确保语料准确可用，并根据客户需求提供定制化的语料加工服务。语料加工方的价值在于能够显著提升模型性能，减少错误率，同时为企业节省时间和成本，并保障数据安全，推动人工智能技术的发展。

亿欧智库：语料加工方分析



2.1 语料辅助方

- ◆ 语料辅助方在整个数据产业链中承担合规审查、版权管理、数据交易撮合、安全保障等职责，是保障数据合法合规流通的关键力量。在“数据要素×”上升为国家战略的背景下语料辅助方得到了进一步的发展。
- ◆ 语料辅助方提供数据脱敏、匿名化、合规审计服务；协助客户应对GDPR、数据安全法等法律监管。提供数据集买卖、语料订阅、数据许可授权等服务。也有提供数据集买卖、语料订阅、数据许可授权等服务。



2.2 语料加工解决方案

- ◆ AI驱动的工具链涵盖数据合成、模型优化、安全性、场景化仿真、行业定制化和多模态融合，广泛应用于工业、医疗、教育等领域，推动技术与需求的深度融合。其智能化的设计不仅提升了语言数据的处理效率，还推动了跨行业应用的广泛落地。无论是学术研究、商业开发还是技术创新，工具链都展现出强大的适应性和潜力，为人工智能与语言技术的融合提供了坚实支持。

亿欧智库：语料加工分析工具



2.2 语料加工解决方案

- ◆ 语料加工作为一种专注于为人工智能、自然语言处理及相关领域提供高质量语料及相关配套服务的专业化解决方案。通过系统化的语料收集、整理、标注、优化和质量控制流程，将原始文本、语音、图像等多种形式的数​​据转化为准确、可用、合规的训练资源。语料加工的核心价值在于显著提升人工智能模型的性能，降低错误率，同时为企业节省时间和成本。
- ◆ 国内领先的人工智能大模型数据平台推出了一系列创新成果，包括数据集描述标准语言和面向大模型语料处理的全流程开源工具体系。实现了语料的标注化、可视化，支持语料的灵活检索和高速下载。

亿欧智库：语料服务解决方案相关案例（以OpenDataLab为例）



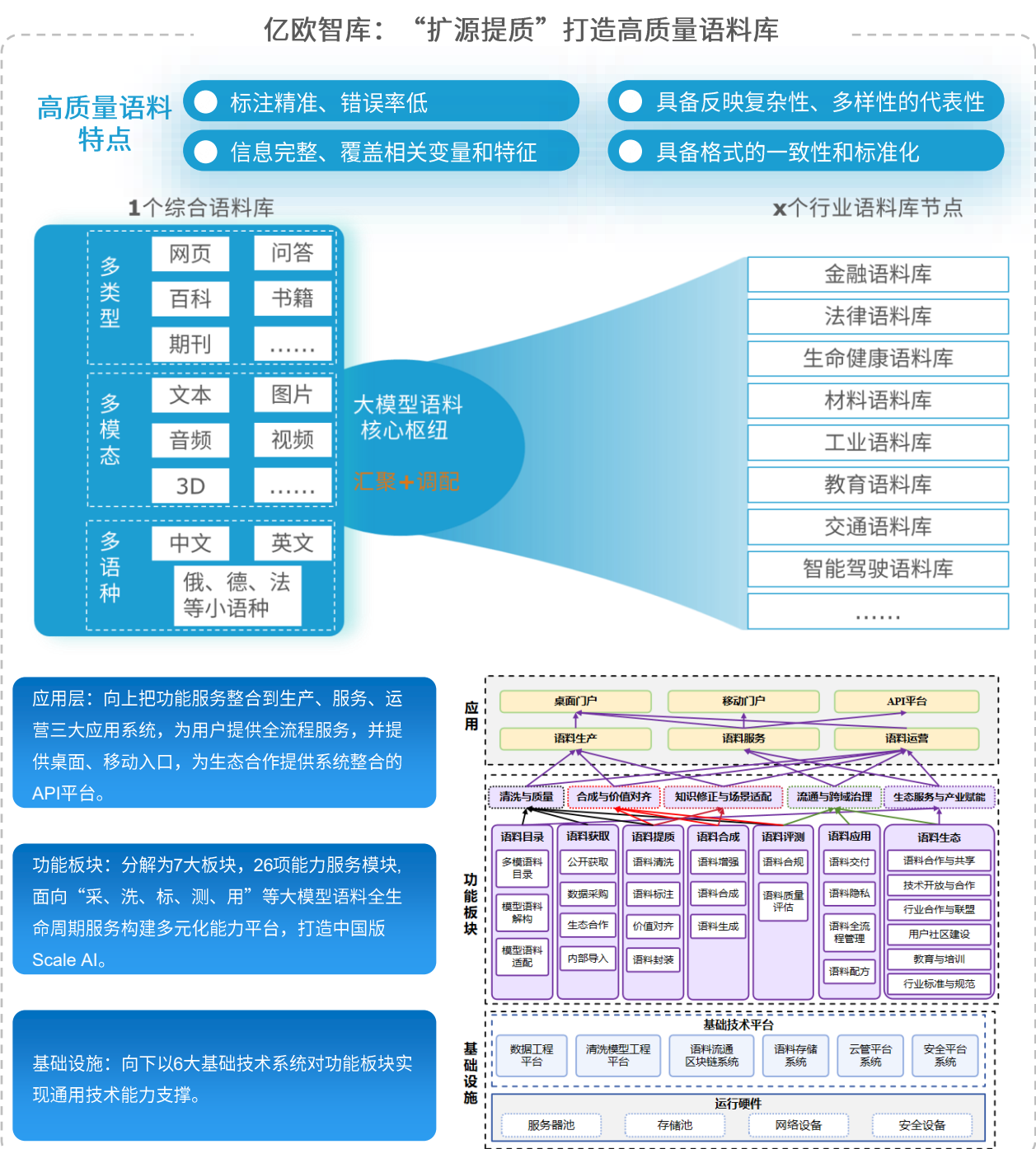
2.3 语料平台建设路径

- ◆ 语料库运营平台是抢抓AI发展战略机遇，构筑我国AI竞争优势的重要突破口。
- 平台的建设致力于打造全国范围内最权威、最全面、最精准的语料数据和服务提供载体。因此，平台建设应当突出国家战略部署和基础服务功能，强化其公共属性和公益定位；同时，考虑大规模语料汇聚、治理、开发等工作所需要的巨大资源投入，平台可以通过语料产品的开发来获取运营收益，反哺平台的建设运营。
- 平台应兼顾汇聚和运营，不仅能够采集、汇聚和存储海量的语料数据，还应通过数据治理，形成对外提供语料检索、分析和应用的服务能力，以支持自然语言处理、机器学习等领域的研究与应用。
- 平台应以需求为导向，面向AI企业、AI训练基地等具有高质量语料的需求方提供数据服务或产品。
- 平台应着眼于产业发展和生态构建，在数据治理和数据服务等环节，发挥平台优势，充分链接更多市场参与主体，通过专业化、链接型、前瞻性的战略布局，推动市场构建语料生态。



2.3 语料平台建设路径

- ◆ 语料平台建设按照综合与专业统筹兼顾、可智能配比的原则，打造大模型语料汇聚与调配的核心枢纽，按照“1+X”整体架构搭建高质量大模型语料资源体系，更好实现与行业场景的适配。
- ◆ 通过汇聚海外语料、小语种等语料资源，不断充实世界知识体系建设，服务于万亿级基础大模型的训练与微调；
- ◆ 通过优化配比综合语料与行业语料资源，更多服务于垂直领域中小企业大模型技术和产品的创新创业，赋能培育一批垂类大模型。



目录

CONTENTS

1 语料市场概况

- 1.1 人工智能语料发展历程
- 1.2 语料发展中外对比
- 1.3 中国人工智能语料主要阻碍

2 语料市场生态发展

- 2.1 语料生态三方分析
- 2.2 语料加工解决方案
- 2.3 语料平台建设路径

3 语料风云榜榜单

- 3.1 评选方式与规则
- 3.2 语料风云榜转型典范Top 10
- 3.3 语料风云榜新锐势力Top 10

4 语料优秀案例分析

- 4.1 中文在线
- 4.2 中启易联
- 4.3 OpenDataLab
- 4.4 Scale AI
- 4.5 CloudFactory

3.1 语料风云榜榜单评选细则

- ◆ 在数字化浪潮的推动下，人工智能已成为推动经济社会发展的重要引擎，作为三驾马车之一“数据”的高质量提速建设，将大大缓解算力压力，助力AI高效发展。
- ◆ 在2025世界人工智能大会期间，将以“语料筑基，智生时代”为主题举办语料专题论坛，围绕高质量语料数据如何高效赋能大模型产业发展，按照专业化、链接型、前瞻性三个维度，向市场传递重构语料生态的顶层设计理念。2025语料风云榜、行业优秀案例也将在该论坛上正式发布，主办方希望通过本次风云榜与案例集的评选，遴选语料行业优秀企业和案例，打造标杆示范，鼓励更多的市场主体投身于语料产业生态布局，推动语料全行业提质、增效、降本，有效缓解语料供给难、供给贵问题。

亿欧智库：2025语料风云榜榜单评选流程

➤ 申报

2025年5月21日—6月9日

申报。开展2025语料风云榜及案例集评选内容征集与申报，并提交申报承诺函

➤ 复评

2025年6月16日-6月20日

参考评价体系，对候选推荐名单进行逐一打分，并通过多项指标的综合考量与评估

➤ 预选

2025年06月10日—06月15日

参赛企业需提供相关企业、产品、技术等介绍（含重点文字介绍、基础信息表，可选提供图片或PPT辅助材料）

➤ 公示

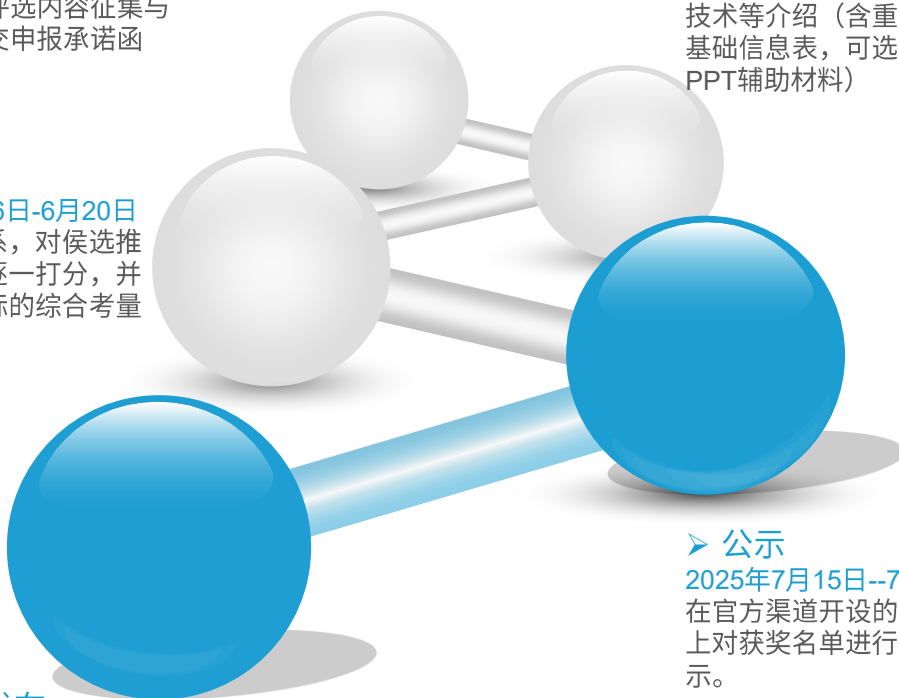
2025年7月15日--7月22日

在官方渠道开设的评选专题页面上对获奖名单进行为期七天的公示。

➤ 颁奖与发布

2025年07月27日

在2025世界人工智能大会语料专题论坛进行发布，并启动各大主流媒体官宣。



3.1 语料风云榜榜单评选细则

- ◆ 预赛要求该申请公司需为面向国内外人工智能语料的代表企业，围绕经营能力、品牌能力、产品能力、创新能力、基础能力、规范能力等指标进行评价。
- ◆ 复赛阶段，申报企业根据申报内容进行线上评选。组委会经过多位专家评委的定性定量复选评审，按照综合得分排名“2025语料风云榜转型典范Top 10”，“2025语料风云榜新锐势力Top 10”。

亿欧智库：2025语料风云榜评选维度

一级指标	二级指标	三级指标
好企业	经营能力	市场表现
		行业标杆案例
		客户体量
	品牌能力	品牌建设
		生态合作伙伴
		带动就业
好产品	产品能力	行业覆盖度
		语料库质量
		DCMM评估 (加分指标)
	创新能力	研发团队实力
		专利和知识产权
		模式与技术特点
好规则	基础能力	安全性要求
		内控管理
	规范能力	参与制修订标准与规则
		社会责任与规范发展

3.2 语料风云榜转型典范Top 10

 上海报业集团 SHANGHAI UNITED MEDIA GROUP	上海报业集团
 TIMEVERSE 天娱数科	天娱数字科技(大连)集团股份有限公司
DCG 数传集团	数传（上海）企业发展有限公司
 光明乳业	光明乳业股份有限公司
 宝信软件 BAOSIGHT	上海宝信软件股份有限公司
 COSCO SHIPPING	中远海运科技股份有限公司
 商汤 sensetime	上海商汤智能科技有限公司
 COL 中文在线	中文在线集团股份有限公司
 YOOZOO 游族网络	上海游族信息技术有限公司
 VIP 维普资讯	重庆维普资讯有限公司

备注：榜单企业排名不分先后

3.3 语料风云榜新锐势力Top 10

	脉策数据	上海脉策数据科技有限公司
	松应科技 ORCA	上海松应科技有限公司
	中国联通 China unicom	联通数据智能有限公司
	本原智数	上海本原智数科技有限公司
	标贝科技 DataBaker Technology	标贝（青岛）科技有限公司
	EASYLINK 中启	成都中启易联科技有限公司
	哈恩德	上海哈恩德技术开发有限公司
	中城交科技 China Urban Transport Technology	中城交（上海）科技有限公司
	文鰲智链	文鰲智链（上海）科技有限公司
		上海未甲文化传媒有限公司

备注：榜单企业排名不分先后

目录

CONTENTS

1 语料市场发展概况

- 1.1 人工智能语料发展历程
- 1.2 语料发展中外对比
- 1.3 中国人工智能语料主要阻碍

2 语料市场生态发展

- 2.1 语料生态三方分析
- 2.2 语料加工解决方案
- 2.3 语料平台建设路径

3 语料风云榜榜单

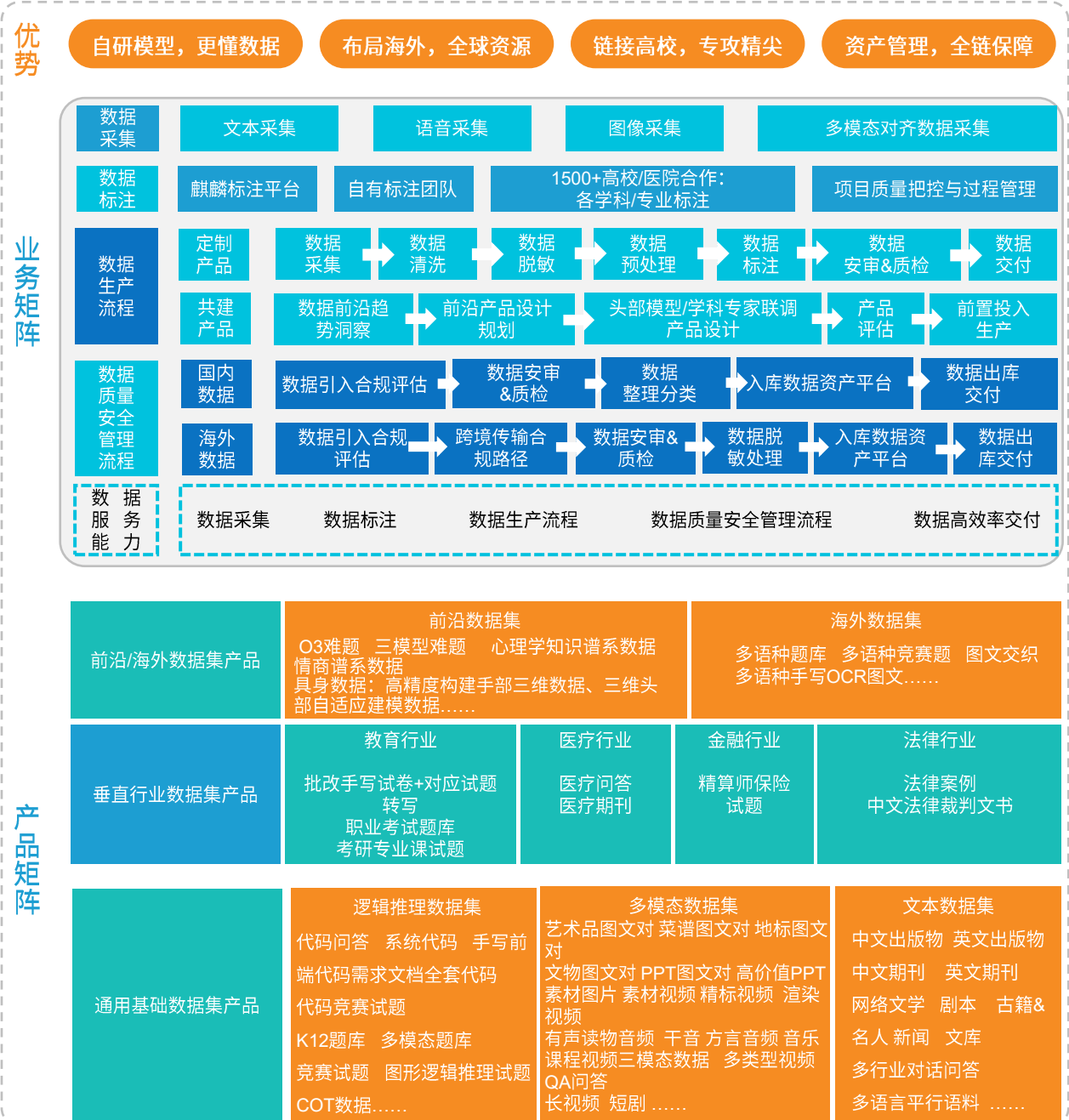
- 3.1 评选方式与规则
- 3.2 语料风云榜转型典范Top 10
- 3.3 语料风云榜新锐势力Top 10

4 语料优秀案例分析

- 4.1 中文在线
- 4.2 中启易联
- 4.3 OpenDataLab
- 4.4 Scale AI
- 4.5 CloudFactory

4.1 中文在线：前沿高质量数据赋能AI智能

- ◆ 中文在线集团成立于2000年，2015年1月21日在深交所创业板上市，成为中国“数字出版第一股”（股票代码：300364），是中国最大的正版数字内容提供商之一,也是全球最大的中文数字出版机构之一。
- ◆ AI数据服务方面，中文在线拥有数字内容资源超400万种，借由国际业务布局优势持续引入海外高价值数据，以先进的大模型数据处理技术和高标准服务严苛把控数据质量与交付，持续为数十家AI大模型提供数据服务。
- ◆ AI数据产品方面，中文在线依托中文逍遥大模型研发经验，整合1500+高校及医疗资源聚焦前沿高精专数据产品开发，已成为头部AI大模型高难度高价值语料定制/共建的核心合作伙伴，累积100+种高质量成品语料集。



4.2 中启易联：大模型高质量训练数据服务商

- ◆ 中启易联科技成立于2015年，是一家专为AI企业及算法机构提供高质量训练数据解决方案的服务商，协助AI模型在多领域实现精准理解、智能决策与高效落地，满足不断升级的AI应用需求。
- ◆ 中启易联科技拥有针对大模型训练的高质量自研数据成品库，包含多语种TB级文本数据、大规模语音预训练数据以及多模态图文、视频文本描述数据集，可满足大模型不同阶段、不同行业的训练需要以及定向能力提升的数据需求。
- ◆ 中启易联科技提供高质量数据定制服务，建立从采集-清洗-处理-标注-规整-交付的数据全生命周期管理解决方案，专业高效安全地完成数据生产与制作，助力模型持续发展进化。

服务框架

文本数据集：
语言理解数据集、语言生成数据集、COT思维链数据集、垂类知识库数据集、多语种数据集
语音数据集：
语音识别数据集、语音合成音库数据集、声纹数据集、情感分析数据集、语音翻译数据集
图像数据集：
图像识别数据集、目标检测数据集、图像分割数据集、图像生成数据集、图像理解数据集
视频数据集：
视频理解数据集、动作识别数据集、目标检测与跟踪数据集、视频描述数据集、视频生成数据集

数据采集：
文本数据采集、语音数据采集、图像数据采集、视频数据采集、小语种数据采集、特殊人群采集
数据标注：
文本标注、语音标注、图像标注、视频标注、多模态数据标注、垂类大模型数据标注
数据工具及平台：
数据采集管理系统、数据标注平台、数据众包平台、数据处理工具集、数据质检工具、模型测试工具

数据安全合规保障

数据安全合规保障

敏感数据三级分类管理数据库平台
《生成式AI服务安全基本要求》检测工具

制度完善

《AI内容安全审核规范》
《数据安全应急预案》
《IT及系统安全管理办法》
《人员权限管理办法》
《物理空间保密等级审定》

人员管理

数据参与人员认证PMP证书
数据审核人员认证内容安全师（ICSP标准）
法务、网信顾问专家团队

知识产权

17个数据相关专利
58个数据相关软著
15个自研数据知识产权登记证书

资质认证

ISO9001质量管理体系
ISO27001信息安全管理
ISO27701隐私信息安全管理
ISO14001环境管理体系
ISO45001职业健康安全管理体系
ISO42001人工智能管理体系
ISO200001信息技术服务管理体系
信息系统安全等级保护备案证明三级

核心优势

多模态数据集服务能力

涵盖文本、语音、图像、视频等多种数据类型。中启能够根据客户需求，定制跨模态数据集，支持从单一数据模态到多模态协同训练的任务需求。通过精细的数据标注和深度融合，帮助客户在计算机视觉（CV）、自然语言处理（NLP）、语音识别（ASR）、文本生成（TTS）等领域大幅提升AI模型的能力，尤其在图像与文本的联动分析、语音与视觉的多模态识别等高复杂任务中具有显著优势。

大规模多模态数据集交付能力

凭借丰富的行业经验，积累整合了大规模的多模态数据集，包括2000TB文本数据集、500万小时多语种ASR数据集、2万小时多语种TTS音频数据集、200亿张图像数据集及100万小时视频数据集等。这些数据集涵盖了多种语言与多样化场景，能够支持大模型训练和智能应用开发，并提供强大数据支持。

数据标注与优化技术

在多模态数据标注方面具备深厚的技术积累，能够提供精准的大模型标注、图像标注、视频标注、语音转文本（ASR）和文本转语音（TTS）等服务。特别是在多语种数据集、情感分析、语义理解等高难度标注任务中，能够保证数据的准确性与高效性。

平台与数据管理

建立了完整的数据管理平台，通过集成的数据采集、存储、处理、优化与分析功能，支持AI应用的数据流通与应用。无论是跨模态的数据流转还是大规模数据集的管理与存储，平台都能保证高效、安全、合规地支持AI项目。

4.3 OpenDataLab：上海 AI 实验室开源数据平台

- ◆ OpenDataLab是由上海人工智能实验室（上海 AI 实验室）打造的开源数据平台。注册用户突破10万人，累计提供数据获取服务200余万次，已成为面向大模型语料的国内体量及数据规模综合领先的人工智能大模型数据平台，已构建/收录开源精标数据集7700余个，并打造了面向大模型语料的全链路开源工具，为各类基座模型及专用模型提供了坚实的数据基座。
- ◆ 其中，平台开源的智能数据提取工具MinerU发布12个月，GitHub星标数就突破4万，凭借快速准确、开源易用的能力特性，受到广大用户及大模型开发者青睐。

OpenDataLab 开源数据平台

依托上海人工智能实验室在 InternLM 大模型工程中的技术积累与数据资源优势，MinerU 持续高速迭代。核心团队成员来自苹果、微软、腾讯等科技企业及清华大学、北京大学等顶尖高校，覆盖 NLP、计算机视觉、高性能计算等多个方向，已在大模型与大数据领域发表顶会论文 200 余篇，具备深厚的学术积淀与产业落地能力，相关成果在 GitHub 平台累计获得超过 5 万星标，展现出广泛的国际技术影响力。

同时，OpenDataLab还与北京大学、上海交通大学等高校联合建设“文档智能解析博士联合培养项目”，围绕 BioMiner、医学文献 AI 综述等实际场景，构建“需求—研发—验证”的快速迭代机制，推动关键技术持续演进，确保产品体系与科研前沿、产业需求同步发展。



目前国内人工智能领域综合规模领先的开源数据平台 OpenDataLab



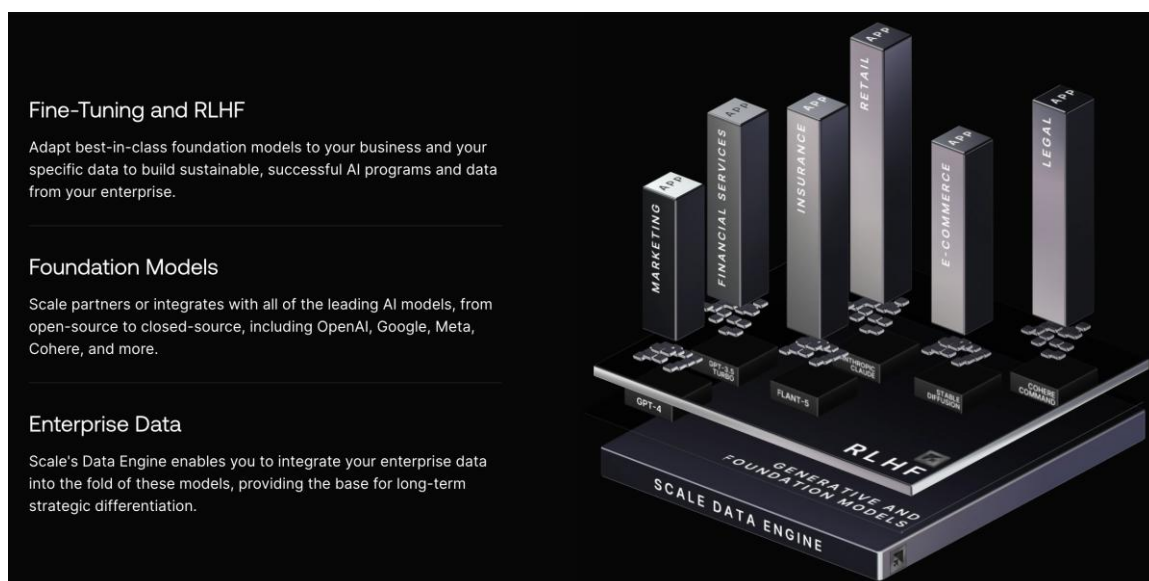
开箱即用，无需复杂配置，即可快速启动。支持批量上传功能，用户可以一次性上传大量文件，大幅提升工作效率。兼容多类文档格式只需一键操作，即可自动完成文档解析，提取关键信息，让数据处理变得简单高效。

为满足不同行业开发者在多元技术场景下的语料提取与开发集成需求，MinerU 还提供包括 Web 应用、在线 API 服务以及本地私有化部署在内的多种产品形态。

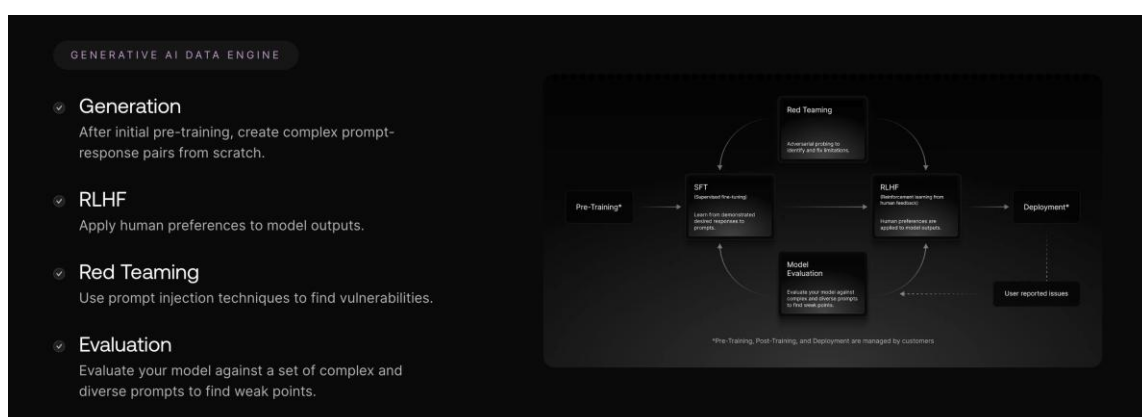
4.4 Scale AI：语料数据注解和模型训练平台公司

- ◆ Scale AI 是一家总部位于美国旧金山的公司，成立于2016年，由Alexandr Wang和Lucy Guo创立，专注于为机器学习模型提供高质量的数据标注和管理服务。该公司通过其先进的平台提供多种数据标注服务，包括图像标注、文本标注、视频标注和3D点云标注等，结合自动化工具与人工审核，确保数据标注的高精度和高效性。技术平台包括Scale Nucleus，一个端到端的数据管理平台，帮助企业从数据采集、标注到分析和模型训练，实现一站式解决方案，以及Scale Rapid，提供快速数据标注服务，满足企业大规模数据集的高效标注需求。
- ◆ Scale AI 的客户遍及多个行业，如自动驾驶、计算机视觉、自然语言处理和机器人技术，知名客户包括OpenAI、Google、Lyft等，其服务在自动驾驶汽车环境感知系统和智能助手的自然语言理解等前沿技术领域得到广泛应用。自成立以来，Scale AI 获得了Y Combinator、Accel和Index Ventures等知名投资者的多轮融资，公司在短时间内迅速成长，估值已达到百亿美元，显示出其在市场上的巨大影响力和发展潜力。

亿欧智库：Scale AI 大模型服务框架



亿欧智库：Scale AI 数据引擎

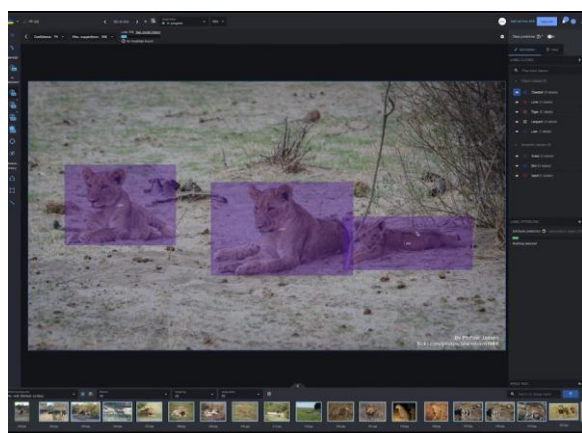


4.5 CloudFactory：语料数据注解和模型训练平台公司

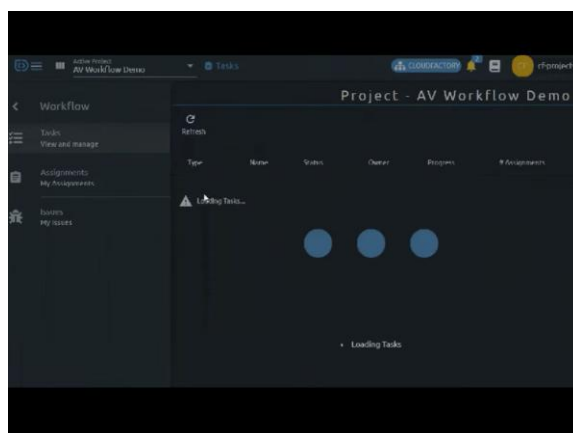
- ◆ CloudFactory 是一家专门提供机器学习和业务流程优化劳动力解决方案的公司。该公司由 Mark Sears 于2010年创立，总部位于北卡罗来纳州的达勒姆市，并在尼泊尔、肯尼亚和英国设有办事处。核心服务包括 AI 辅助的数据标注、数据注释、数据管理以及各种其他数据处理解决方案，这些服务对金融、医疗保健、保险、零售和地理空间等行业至关重要，帮助企业高效、准确地管理和处理大量数据。
- ◆ CloudFactory 商业模式围绕虚拟生产线的概念展开，这一概念借鉴了传统制造业的装配线方法，这种方法使他们能够将复杂的数据任务分解为可管理的步骤，确保高质量和高效的结果。这个模型由他们的专有劳动力管理平台支持，该平台增强了团队协作、实时质量可视化和安全的数据访问，现在为超过700名客户提供服务并处理了数百万的数据任务。

亿欧智库：CloudFactory AI标注工具

AI 自动标注系统



标注系统 workflow



亿欧智库：CloudFactory AI客户



团队介绍

- ◆ 亿欧智库（EO Intelligence）是亿欧旗下的研究与咨询机构。为全球企业和政府决策者提供行业研究、投资分析和创新咨询服务。亿欧智库对前沿领域保持着敏锐的洞察，具有独创的方法论和模型，服务能力和质量获得客户的广泛认可。
- ◆ 亿欧智库长期深耕新科技、消费、大健康、汽车出行、产业/工业、金融、碳中和等领域，旗下近100名分析师均毕业于名校，绝大多数具有丰富的从业经验；亿欧智库是中国极少数能同时生产中英文深度分析和专业报告的机构，分析师的研究成果和洞察经常被全球顶级媒体采访和引用。
- ◆ 以专业为本，借助亿欧网和亿欧国际网站的传播优势，亿欧智库的研究成果在影响力上往往数倍于同行。同时，亿欧内部拥有一个由数万名科技和产业高端专家构成的资源库，使亿欧智库的研究和咨询有强大支撑，更具洞察性和落地性。

报告作者



何舒阳

亿欧智库 分析师

Email: heshuyang@iyiou.com

报告审核



孙毅颂

亿欧智库研究总监

Email:

sunyisong@iyiou.com

版权声明

- ◆ 本报告所采用的数据均来自合规渠道，分析逻辑基于智库的专业理解，清晰准确地反映了作者的研究观点。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。本报告的信息来源于已公开的资料，亿欧智库对该等信息的准确性、完整性或可靠性作尽可能的获取但不作任何保证。
- ◆ 本报告版权归亿欧智库所有，欢迎因研究需要引用本报告部分内容，引用时需注明出处为“亿欧智库”。对于未注明来源的引用、盗用、篡改以及其他侵犯亿欧智库著作权的商业行为，亿欧智库将保留追究其法律责任的权利。

关于亿欧

- ◆ 亿欧是一家专注科技+产业+投资的信息平台和智库；成立于2014年2月，总部位于北京，在上海、深圳、南京、纽约设有分公司。亿欧立足中国、影响全球，用户/客户覆盖超过50个国家或地区。
- ◆ 亿欧旗下的产品和服务包括：信息平台亿欧网（iyiou.com）、亿欧国际站（EqualOcean.com）、研究和咨询服务亿欧智库（EO Intelligence），产业和投融资数据产品亿欧数据（EO Data）；行业垂直子公司亿欧大健康（EO Healthcare）和亿欧汽车（EO Auto）等。
- ◆ 基于对中国科技、产业和投资的深刻理解，同时凭借国际化视角和高度，亿欧为中外客户提供行业研究、投资分析、创新咨询、数据产品、品牌公关、国际化落地等服务。已经服务过的客户包括华为、英特尔、腾讯、百度、一汽解放、理想汽车、京东、微软、安顾集团、统信、中石油-昆仑数智、中电信息、东信集团等。

关于库帕思

- ◆ 上海库帕思科技有限公司是按照市委市政府要求为建设“人工智能上海高地”而成立的功能性语料枢纽，公司按照“开放性”、“链接型”、“市场化”的总体要求，搭建“1125”的语料工作框架。库帕思立足上海、服务全国，旨在赋能千行百业新一轮增长，提供高价值、高维度、超大规模的语料资源以及语料技术解决方案，培育以语料为引领的人工智能创新生态，努力建设服务大模型迭代演进的语料“超级工厂”。

关于OpenDataLab

- ◆ OpenDataLab是由上海人工智能实验室（上海 AI 实验室）打造的开源数据平台。注册用户突破10万人，累计提供数据获取服务200余万次，已成为面向大模型语料的国内体量及数据规模综合领先的人工智能大模型数据平台，已构建/收录开源精标数据集7700余个，并打造了面向大模型语料的全链路开源工具，为各类基座模型及专用模型提供了坚实的数据基座。其中，平台开源的智能数据提取工具MinerU发布12个月，GitHub星标数就突破4万，凭借快速准确、开源易用的能力特性，受到广大用户及大模型开发者青睐。

亿欧服务

- ◆ 基于自身的研究和咨询能力，同时借助亿欧网和亿欧国际网站的传播优势；亿欧为创业公司、大型企业、政府机构、机构投资者等客户类型提供有针对性的服务。

◆ 创业公司

亿欧旗下的亿欧网和亿欧国际站是创业创新领域的知名信息平台，是各类VC机构、产业基金、创业者和政府产业部门重点关注的平台。创业公司被亿欧网和亿欧国际站报道后，能获得巨大的品牌曝光，有利于降低融资过程中的解释成本；同时，对于吸引上下游合作伙伴及招募人才有积极作用。对于优质的创业公司，还可以作为案例纳入亿欧智库的相关报告，树立权威的行业地位。

◆ 大型企业

凭借对科技+产业+投资的深刻理解，亿欧除了为一些大型企业提供品牌服务外，更多地基于自身的研究能力和第三方视角，为大型企业提供行业研究、用户研究、投资分析和创新咨询等服务。同时，亿欧有实时更新的产业数据库和广泛的链接能力，能为大型企业进行产品落地和布局生态提供支持。

◆ 政府机构

针对政府类客户，亿欧提供四类服务：一是针对政府重点关注的领域提供产业情报，梳理特定产业在国内外动态和前沿趋势，为相关政府领导提供智库外脑。二是根据政府的要求，组织相关产业的代表性企业和政府机构沟通交流，探讨合作机会；三是针对政府机构和旗下的产业园区，提供有针对性的产业培训，提升行业认知、提高招商和服务域内企业的水平；四是辅助政府机构做产业规划。

◆ 机构投资者

亿欧除了有强大的分析师团队外，另外有一个超过15000名专家的资源库；能为机构投资者提供专家咨询、和标的调研服务，减少投资过程中的信息不对称，做出正确的投资决策。

- ◆ 欢迎合作需求方联系我们，一起携手进步；电话 010-53321289，邮箱 hezuo@iyiou.com



网址: <https://www.iyiou.com/research>

邮箱: hezuo@iyiou.com

电话: 010-53321289



扫码关注亿欧智库
查看更多研究报告



扫码添加小助手
加入行业交流群