

亿欧智库 https://www.iyiou.com/research

Copyright reserved to EO Intelligence, July 2026

2026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

- 当人工智能的浪潮席卷全球，算力、算法与数据被视为驱动智能革命的三大引擎。而在这其中，语料作为“数据”的核心形态，正从幕后走向台前，成为决定AI模型智商与情商的“隐形天花板”。中国，作为全球AI竞争的主战场之一，其语料产业正站在从“野蛮生长”向“精耕细作”跨越的历史节点。

- 伴随 AI 技术与产业应用的深度推进，语料数据领域正迎来全方位的变革升级，需求端率先呈现出清晰的演进方向。语料需求正向 “更广、更深、更真” 迈进：“更广” 是持续拓展网络、书籍、行业和真实场景等通用与专业数据；“更深” 是通过多模态加工和专家知识提炼，将复杂逻辑与隐性经验转化为模型可学习的数据；“更真” 是围绕代码、智能体和真实业务任务，沉淀可运行、可交互、可反馈的过程数据。

- 需求升级倒推着生产端迭代，语料数据生产方式正从传统的数据筛选，向知识转换和环境合成延伸。一方面对海量数据去伪存真、保留核心知识价值；另一方面将逻辑、背景和专家经验等隐性知识转化为显性知识，并通过代码沙盒、模拟环境等方式生成任务轨迹和交互数据。专家众包与 AI4Data 技术协同发力，正在重塑高质量语料的生产范式。

- 链路变革最终传导至服务端，语料服务模式正从一次性交付，转向 “采集、加工、使用、评测、再加工” 的持续运营闭环。数据团队更加深入行业现场，面向模型和智能体提供规划设计、产品交付和持续迭代等全链路服务，推动语料服务从 “卖数据” 向 “做运营” 全面升级。

目录

CONTENTS

1 人工智能语料的发展现状

1.1 人工智能语料发展规划
1.2 中国人工智能语料发展现状
1.3 中国人工智能语料发展阻碍与机遇

2 人工智能语料的行业需求特征

2.1 行业语料需求潜力象限分析
2.2 重点领域需求象限分析
2.3 新兴领域需求象限分析
2.4 语料需求潜力小结
2.5 AI4S语料特征、痛点与范式变化
2.6 AI4S语料特征、痛点与范式变化趋势

3 语料风云榜榜单

3.1语料风云榜榜单评选细则
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 人工智能语料优秀案例分析

4.1国家管网
4.2量子城市
4.3厦门供应链数智创新
4.4谱蓝科技
4.5脉搏深处
4.6飞数数据

1.1 人工智能语料发展规划

- 从各级政策演进能够看出，国内中文语料建设由零散自发建设转向国家统筹布局、省市落地配套的系统化发展路径，政策重心逐步从规范数据安全边界过渡到优质语料规模化供给与要素市场化并行。国家层面持续完善顶层制度，围绕数字中文、数据要素与生成式 AI 接连出台指导性文件，明确通用、行业、民族语言、传统文化多品类语料建设任务与标准体系，不断打通公共数据转化合规语料的制度通道。各地方依托数据条例与 AI 扶持政策，落地语料专区交易、财政奖补、公共数据授权运营等落地举措，形成东部先行试点、多省份跟进落地的区域格局。整体行业呈现合规化、精品化、产业化同步推进的特征，通用通识语料夯实底座，医疗、工业、法律等垂直领域语料成为政策扶持重点，版权确权与分级管理日趋成熟，政策导向持续推动公共资源与市场化资源互补共建。

亿欧智库：中国语料发展规划

	2022	2023	2024	2025	2026.07
顶层规划	《“十四五”数字经济发展规划》	《生成式人工智能服务管理暂行办法》 《数字中国建设整体布局规划》	《“数据要素×”三年行动计划（2024—2026 年）》 《中办国办关于加快公共数据资源开发利用的意见》 《关于促进数据产业高质量发展的指导意见》	《教育部 国家语委 中央网信办关于加强数字中文建设 推进语言文字信息化发展的意见》 《公共数据资源登记管理暂行办法》 《人工智能高质量数据集建设指南》	《人工智能语料库基础术语》 《国家数据局关于培育数据流通服务机构加快数据要素市场化意见（2026.2）》
地方法规		《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023-2024 年）》 上海数据交易所设立全国首个语料交易专区配套《语料数据合规指引》 《北京市促进人工智能产业发展条例》	《上海市促进工业服务业赋能产业升级行动方案（2024-2027 年）》扶持工业领域语料 《深圳市打造人工智能先锋城市的若干措施》语料补贴	《浙江省人民政府关于支持人工智能创新发展若干措施的通知》 《山东省省级“语料券”奖补资金管理办法及实施细则》财政补贴语料库建设	《广西深入实施“人工智能+”三年行动方案（2026—2028 年）》扶持本土民族语料库建设

数据来源：中国政府、亿欧智库

1.2 中国人工智能语料发展现状

- 中文语料体系已形成通识语料为主体、垂直语料为补充的成熟格局，整体规模与质量实现大幅跃升，依托全网文本、版权图书、百科资讯等资源搭建起大模型通用训练底座，当前行业呈现通用语料体量远超垂直语料、企业私域数据主导、开源数据为辅的特征，同时合成数据逐步成为新增量，语料建设从规模扩张转向合规化、精细化、场景化升级，为中文大模型产业化落地筑牢数据根基。



*语料为公开资料来源的原始未去重未精洗语料，语料语言为中文，Token计算为1中文约等于2.5Token

数据来源：亿欧智库、CSDN

1.3 中国人工智能语料发展阻碍与机遇

◆ 中国人工智能语料发展正处于关键转型期。在政策强力引导与垂直应用爆发的机遇下，国家级平台建设加速，出版传媒业作为高质量语料“富矿”的价值凸显。但发展同时面临多重阻碍：版权壁垒与利益分配机制缺失导致“数据孤岛”难以打通；专业加工人才匮乏及行业标准不统一制约了语料质量；加之高质量垂直语料历史积累薄弱、数字化程度低，使得核心语料获取成本极高。整体而言，行业正面临“有数据难流通、有需求难满足”的结构性困境。

阻碍一：法律法规与共享机制尚不健全

在具体落地层面，语料涉及的版权归属、用户隐私保护等法律法规细则仍不够明确。由于缺乏合理的利益分配机制和统一的共享协同机制，许多高质量语料仍被封闭管理，形成了“数据孤岛”，导致“有数据难流通”。

机遇一：政策强力引导与国家级平台建设

国家层面高度重视数据要素，将数据作为生产要素，并密集出台政策支持高质量数据集的开发。国家发改委、国家数据局等部门不仅印发了指导意见，还在北京、上海等地推动建立数据交易所和语料交易专区，甚至牵头设立国家级语料库联盟，为行业发展注入了强劲的制度动能。

阻碍二：版权壁垒突出与授权模式不成熟

出版传媒资源丰富，但这些高质量内容往往分散在不同的机构手中，且受到严格的版权保护。当前市场上缺乏成熟、统一的语料交易与授权模式，版权授权壁垒突出，侵权风险高发。这导致大模型企业难以合法、低成本地获取这些核心语料，造成了“有需求难满足”的行业困局。

机遇二：出版与传媒业成为高质量语料的“富矿”

随着互联网公开数据逐渐耗尽，行业开始挖掘高价值密度的垂类语料。出版业和传媒业拥有海量经过专业编辑审核、逻辑链完整的图书、期刊和音视频素材。这些内容具备高可信度和系统性，是训练大模型复杂推理能力的绝佳原料。目前，已有大量出版社和传媒上市公司牵头，开始将专业资源转化为AI语料。

阻碍三：专业标注人才缺口与加工标准不统一

高质量语料离不开精细化的人工加工，但目前国内数据标注行业面临百万级的人才缺口，且整体流程规范性不强。同时，不同机构在分词、句法结构、语义标注等方面缺乏统一的国家或行业标准，导致产出的语料库结构差异大、难以兼容互通，严重影响了模型训练的一致性和准确性。

机遇三：本土数据加工与标注技术的迭代升级

面对中文语料短缺，国内正在加速研发更精准的中文自然语言处理工具，并推动数据标注行业从劳动密集型向知识密集型转变。通过自动化标注、密态计算等新技术的应用，不仅能提升语料处理效率，还能在保障安全的前提下挖掘数据价值，构建自主可控的中文语料体系。

阻碍四：高质量垂直语料稀缺且获取难度极高

中文在垂直领域的优质语料历史积累非常薄弱。许多专业领域的语料仍以纸质或非结构化形式存在，数字化程度极低。此外，专业语料对标注者的领域知识要求极高，导致加工成本高昂，且容易受到生成式AI“幻觉”内容的反向污染，进一步加剧了纯净专业语料的稀缺。

机遇四：垂直领域应用爆发催生定制化语料需求

随着AI在医疗、金融、法律、工业等行业的深度渗透，市场对定制化、高安全性的行业专属语料需求迫切。这种明确的市场需求正在倒逼产业链上游进行资源整合，促使企业、高校和科研机构联手，针对特定场景开发高精度的工业级语料池，为AI在尖端领域的落地提供支撑。

数据来源：亿欧智库、CSDN

目录

CONTENTS

1 人工智能语料的发展现状

1.1 人工智能语料发展规划
1.2 中国人工智能语料发展现状
1.3 中国人工智能语料发展阻碍与机遇

2 人工智能语料的行业需求特征

2.1 行业语料需求潜力象限分析
2.2 重点领域需求象限分析
2.3 新兴领域需求象限分析
2.4 语料需求潜力小结
2.5 AI4S语料特征、痛点与范式变化
2.6 AI4S语料特征、痛点与范式变化趋势

3 语料风云榜榜单

3.1语料风云榜榜单评选细则
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

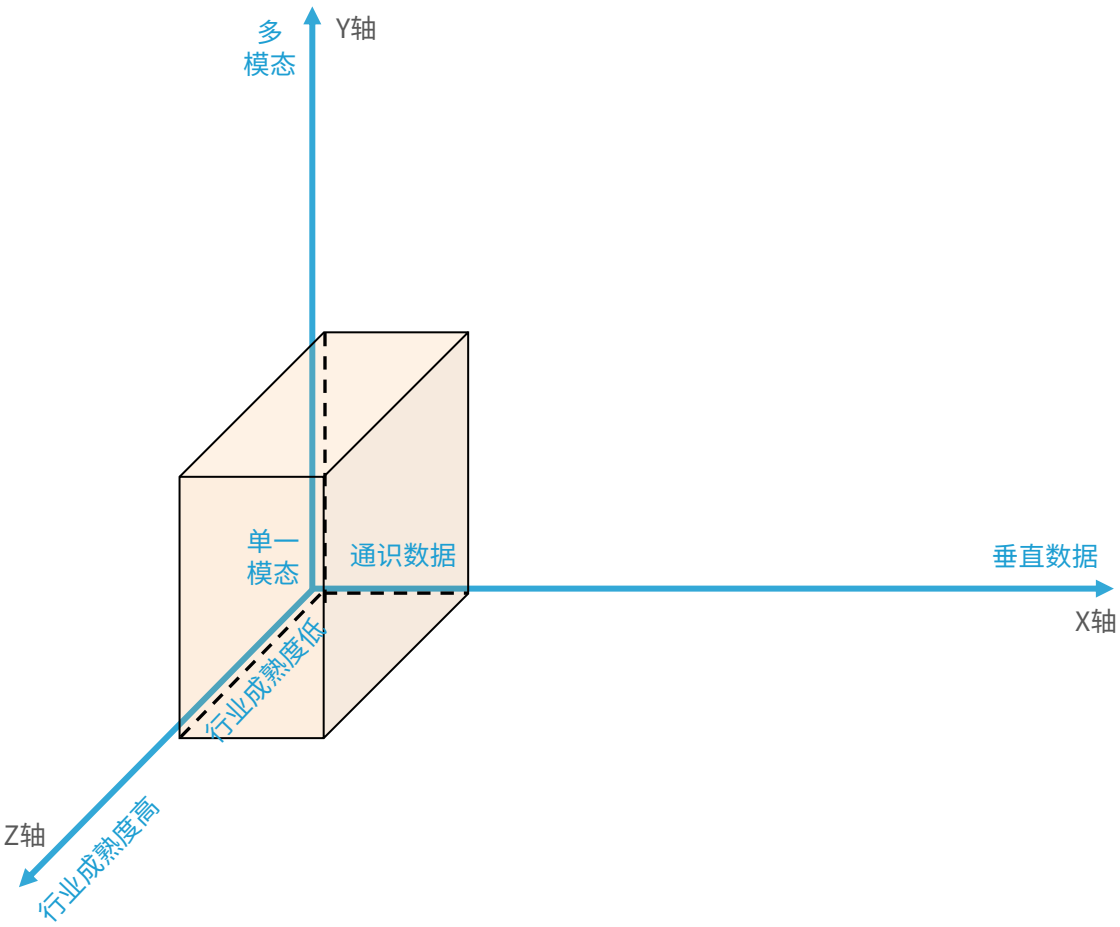
4 人工智能语料优秀案例分析

4.1国家管网
4.2量子城市
4.3厦门供应链数智创新
4.4谱蓝科技
4.5脉搏深处
4.6飞数数据

2.1 行业语料需求潜力象限分析

◆ 语料的价值密度与其垂直程度成正比。面对特定行业的深层需求，通用的公开数据往往显得力不从心，唯有高密度的垂直领域语料才能破解专业场景下的“知识匮乏”难题。与此同时，多模态语料打破了传统单一文本的局限，以其多变的形态和更立体的信息承载能力，展现出远超常规语料的核心价值。但需要强调的是，语料本身并不是万能的，它的最终价值无法脱离应用行业的成熟度与AI技术的适配度而独立存在。只有在技术可行、场景成熟的闭环中，优质语料才能真正成为驱动行业智能化跃迁的核心引擎。

亿欧智库：语料需求潜力象限分析模型



面积衡量语料未来的商业化需求潜力。

X轴衡量语料的行业深度，国家数据局明确将高质量数据集分为通识数据集、行业通识数据集和行业专识数据集。随着X轴向行业深处延伸，数据的敏感程度与专业门槛同步提升，对语料合规性的要求也愈发严苛。因此，X轴不仅直接对应了这一核心划分标准，体现了AI从通用能力向专业深度演进的趋势，更深刻反映了在专业知识沉淀过程中，合规治理必须贯穿始终的内在逻辑。

Y轴衡量语料的模态复杂度，决定了语料采集、标注和处理的难度。随着具身智能和多模态大模型的兴起，单纯的文字语料已无法满足需求，音视频、交互指令等高复杂度语料正成为新的竞争高地。

Z轴衡量需求语料的行业自身AI商业落地成熟度。

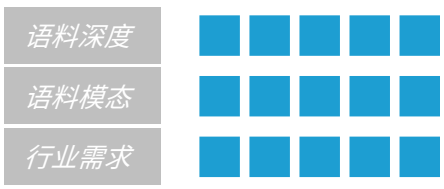
数据来源：亿欧智库

2.2 重点领域需求象限分析

◆ 医疗以100分居首，其专有知识体系严密、多模态语料（文本/影像/语音等）覆盖广，且直接关联生命健康，高质量语料需求迫切；金融（65分）依赖专属语料构建合规逻辑，需支撑高频交易与风控迭代；自动驾驶（57分）聚焦交通场景，多源传感器数据与时空同步要求高；智能制造（51分）依托工业机理与产线工艺知识，适配工业4.0智能化升级；教育（39分）覆盖K12至高等教育学科谱系，需个性化标注语料支撑AI教学交互。各领域因知识壁垒、模态复杂度、行业刚需差异，呈现梯度化语料需求特征。

亿欧智库：重点领域语料需求潜力分析

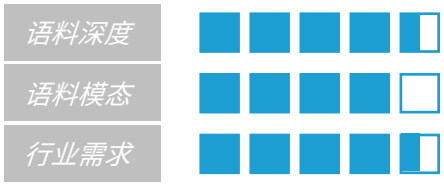
医疗



需求潜力评分 **100**

- 医疗领域拥有极其严密的专有知识体系（如ICD-11疾病标准、临床指南、药品库等），必须依赖高质量的垂类语料进行深度训练。
- 语料涵盖了文本、2D影像（X光、病理切片）、3D影像（CT、MRI）、视频（超声、手术录像）以及语音等多种模态，且需要跨模态的精准对齐与推理。
- 医疗领域直接关系到生命健康，对高质量、经过脱敏和专业标注的语料有着极度迫切的需求，以支撑辅助诊断、药物研发等核心场景。

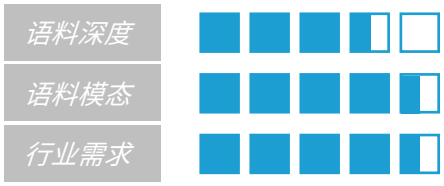
金融



需求潜力评分 **65**

- 具有独特的“行话”体系（如久期、回撤）、复杂的反讽修辞以及严格的风控合规要求，通用模型常出现幻觉或误读，需要大量金融专属语料来构建内生思维链。
- 除了常规的研报文本、路演音频外，还涉及大量非结构化的票据、流水、合同等图像数据的OCR识别与交叉验证。
- 市场对高频交易、智能投顾和信贷风控的迭代速度要求极快，需要持续的高质量语料输入以保持模型的敏锐度和合规性。

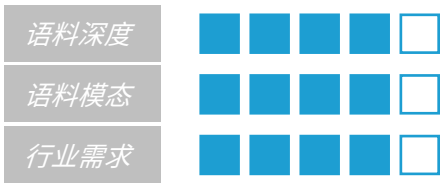
自动驾驶



需求潜力评分 **57**

- 专注于交通法规、车辆动力学及极端路况的处理，具有较强的场景专属性，不过视觉特征属于物理世界的通用常识，该领域拥有大量世界级的开源多模态数据集（如KITTI、nuScenes、Waymo Open Dataset等）
- 数据融合了激光雷达、毫米波雷达、摄像头视频流、高精地图等多源传感器数据，对时空同步要求极高。
- 安全是底线，需要海量的长尾场景语料来不断迭代算法，提升感知的准确性和决策的安全性。

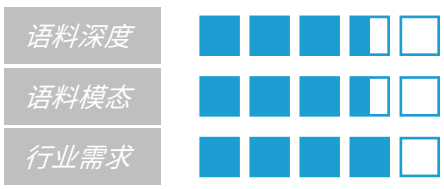
智能制造



需求潜力评分 **51**

- 包含深厚的工业机理、材料科学及特定产线的工艺流程知识，壁垒较高。
- 涉及工业视觉（缺陷检测）、时序传感器数据、设备日志以及流体仿真等多种形态。
- 随着工业4.0的推进，对基于机理的智能体控制和无人化生产需求增加，推动了对高质量工业数据集的建设。

教育



需求潜力评分 **39**

- 涵盖K12到高等教育的学科知识图谱，虽然部分基础内容与通用知识重叠，但教学逻辑和辅导策略具有垂直性。
- 数据多模态包括教材文本、教学音视频、手写笔迹识别以及互动问答等。
- 个性化教学和AI助教需要大量经过教育学标注的语料，以实现更拟人化和启发式的交互。

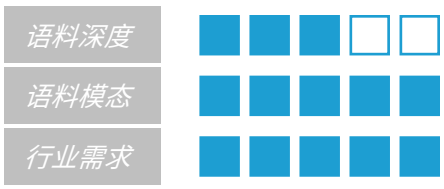
数据来源：亿欧智库、凤凰资讯、CSDN、财联社

2.3 新兴领域需求象限分析

◆ 具身智能以80分的需求潜力评分位居首位，其高度依赖三维点云、关节轨迹等多模态数据的实时融合，且真实世界的高质量交互数据极度稀缺，对仿真数据和自动化处理平台有强依赖。文化影视（32分）作为多模态密集型行业，涉及剧本、分镜、音视频及3D资产，在AIGC辅助创作方面潜力巨大，但亟需带版权许可的高质量素材。出版传媒（25分）则侧重于内容规范与事实核查，正积极利用大模型进行辅助写作和审校，对高质量中文语料库有稳定需求。总体来看，新兴领域呈现出从物理交互到内容创作的多层次语料需求特征。

亿欧智库：新兴领域语料需求潜力分析

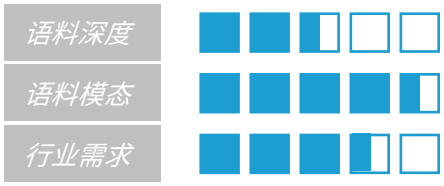
具身智能



需求潜力评分 **80**

- 聚焦于物理世界的交互逻辑、运动控制与空间感知，与传统互联网文本数据差异巨大。
- 高度依赖三维点云、关节轨迹、力觉反馈、视觉图像等多模态数据的实时融合，是目前多模态复杂度最高的领域之一。
- 这是一个新兴且高速发展的赛道，真实世界的高质量交互数据极其稀缺，目前极度依赖仿真数据和自动化处理平台来生成海量训练语料。

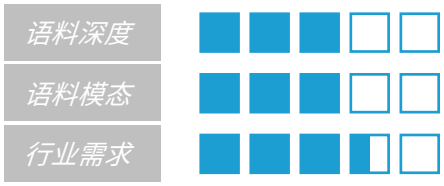
文化影视



需求潜力评分 **32**

- 创意产业主观性强，虽然有其行业流程，但对绝对标准化知识的依赖度相对较低。
- 典型的多模态密集型行业，涉及剧本（文本）、分镜（图像）、影视原片（视频）、音效配乐（音频）以及3D建模资产。
- 在AIGC辅助美术创作、音乐生成和影视制作环节有较大潜力，需要大量带版权许可的高质量素材语料。

出版传媒



需求潜力评分 **25**

- 侧重于内容创作规范、事实核查与版权保护，对语言的准确性和逻辑性有较高要求。
- 主要是图文混排，逐渐向短视频脚本和数字人生成拓展。
- 行业正积极寻求利用大模型进行辅助写作和内容审校（如公文、稿件纠错），对高质量的中文语料库有稳定需求。

数据来源：亿欧智库、凤凰资讯、CSDN、财联社

2.4 语料需求潜力小结

- 梯度划分反映了该行业语料的紧迫性与商业价值，第一梯队的医疗与具身智能凭借极高的垂直壁垒、多模态复杂度及迫切的语料需求，成为当前高质量数据建设的绝对核心战场，掌握其核心语料即意味着占据了行业制高点。第二梯队的金融、自动驾驶和智能制造具备深厚的数字化基础与极高的商业价值。第三梯队各自在视频生成等单一维度上爆发力极强，且作为大模型最贴近C端用户的应用场景。

第一梯队

医疗　　具身智能

三项指标几乎没有短板，且互相放大效应极强

无论是医疗的严谨性命攸关，还是具身智能对物理世界交互数据的极度饥渴，都决定了它们是当前高质量语料建设的绝对核心战场。谁掌握了这些高门槛的垂直多模态语料，谁就掌握了行业的制高点。

第二梯队

金融　　自动驾驶　　智能制造

具备极高的商业价值和复杂的数据形态

行业已经具备了较好的数字化基础，对大模型的需求非常务实（如风控、安全驾驶、良品率提升）。它们的语料建设重点在于如何将复杂的非结构化数据（如传感器数据、金融票据）转化为机器可理解的高质量指令。

第三梯队

教育　　文化影视　　出版传媒

相对更依赖通用语言能力，或者垂直壁垒略低于上述硬核科技行业

总分不高，但部分行业在单一维度上表现极其突出。例如文化影视的多模态程度高达9分，这意味着它在视频生成（如Sora类模型）、3D资产构建等AIGC领域拥有巨大的爆发潜力，是未来内容生产力的核心赛道。

数据来源：亿欧智库

2.5 AI4S语料特征、痛点与范式变化

◆ 决策式AI学习条件概率分布，依赖强监督与高信噪比的金标准数据，面临小样本与高维灾难痛点，其应用范式为替代模型以加速科学预测。生成式AI学习联合概率分布，具备无监督/自监督与多模态特征，痛点在于需满足物理约束与有效性验证，其范式转变为逆向设计与假设生成，成为创造引擎。两者在数据处理方式、应用场景上存在显著差异，分别推动科学研究从预测到创造的范式演进。

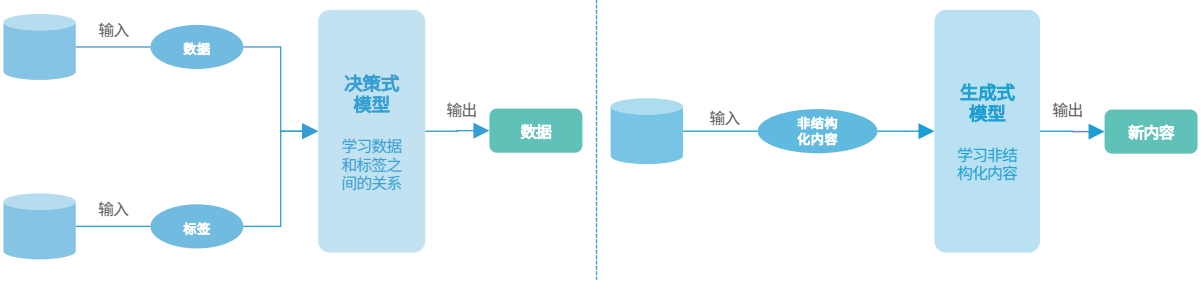
亿欧智库：决策式AI和生成式AI4S数据语料特征、痛点与范式变化

决策式AI

学习数据中的条件概率分布，即一个样本归属于特定类别的概率，在标记数据点的数据集上进行训练，学习数据点的特征和标签之间的关系，模型训练完成后可对新的场景进行判断、分析和预测

生成式AI

学习数据中的联合概率分布，即数据中多个变量组成的向量的概率分布，对已有的数据进行总结归纳，并在此基础上使用深度学习技术等，创作模仿式、缝合式的内容



数据特征
强监督与高信噪比

决策式AI（如用于预测材料带隙、蛋白质结合亲和力的模型）需要的是经过严格清洗、标注明确的“金标准”数据。这类数据通常来自权威数据库或经过同行评审的实验记录。语料的质量直接决定了模型预测的物理准确性，数据中的任何系统性误差都会被模型放大。

数据痛点
小样本与高维灾难

科学领域的真实实验数据往往极其昂贵且稀疏。决策式AI面临着“数据量远小于特征维度”的困境。因此，在语料处理上，极度依赖特征工程和降维技术，以从有限的语料中提取出最具物理意义的特征变量。

数据应用范式
替代模型

决策式AI将复杂的科学仿真（如流体力学方程、量子化学计算）转化为数据驱动的预测模型。其语料价值在于“替代高成本实验”，通过训练数据来逼近真实物理规律，从而在毫秒级内完成原本需要超算数天的预测任务。

数据特征
无监督/自监督与多模态

生成式AI不再局限于精准的“输入-输出”配对，而是能够消化海量的、未标注的科学文献、化学式、基因序列甚至科研论文文本。它通过自监督学习，从庞大的异构语料中捕捉科学实体之间的深层关联与潜在规则。

数据痛点
物理约束与有效性验证

生成式AI面临的最大挑战是“生成结果是否符合科学真理”。如果训练语料缺乏严格的物理化学约束，模型极易产生“科学幻觉”。因此，其语料处理的核心在于将物理定律（如对称性、守恒律）作为约束条件嵌入数据生成过程，确保生成的数据不仅在统计上合理，在物理上也是合法的。

数据应用范式
逆向设计与假设生成

生成式AI将语料数据从“预测工具”转变为“创造引擎”。它不再是对已有数据的简单拟合，而是通过探索训练数据未覆盖的广阔化学空间或生物空间，反向生成具有特定功能的新材料、新药物候选分子，甚至直接为科学家提出全新的科研假设。

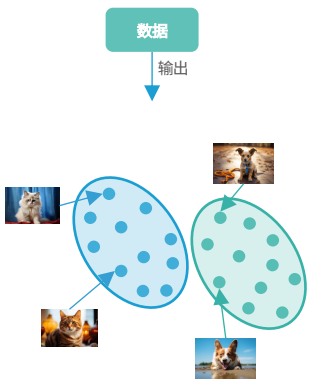
数据来源：亿欧智库

2.6 AI4S语料特征、痛点与范式变化趋势

◆ 在AI4S 的未来发展中，决策式AI与生成式AI对语料数据的利用将不再局限于简单的“投喂与学习”，而是向着更深层次的物理融合、跨模态理解以及自动化闭环演进。

亿欧智库：决策式AI和生成式AI4S数据语料特征、痛点与范式变化

决策式AI



高质量小样本驱动

面对科学数据昂贵且稀缺的现实，未来的决策式模型将不再依赖数以亿计的盲目数据堆砌。通过精心挑选的高质量科学训练样本与透明的训练策略，利用“质量优于数量”的技术路径，即可在多项科学基准测试中实现突破。

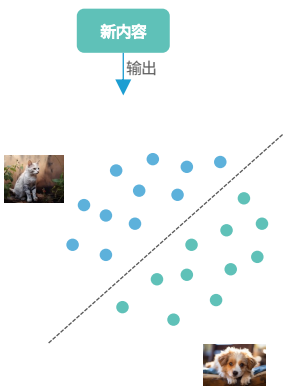
嵌入物理先验约束

为了减少“黑盒”预测的不可解释性，决策式AI将在训练中引入物理先验约束。通过将领域知识（如物理、化学定律）与算法深度融合，降低对数据规模的依赖，从而提升模型在真实科学场景中的预测准确性与泛化能力。

从“唯数据量论”转向“物理约束与高质量小样本”

未来，决策式AI在利用语料数据时将更加注重数据的“含金量”与科学规律的嵌入，而非盲目追求海量数据的堆砌。

生成式AI



高跨模态的全场景理解

未来的生成式AI将能够深度解析从微观世界到宏观宇宙，再到数理逻辑的跨尺度、跨学科全场景语料。这种全维度的语料消化能力，将使其成为能够参与科学问题定义与重构的“认知参与者”。

合成数据生成与数据飞轮

生成式AI将利用深度生成模型（如扩散模型等）学习底层数据分布，主动生成逼真的合成数据以补充训练集。这不仅解决了存量数据匮乏的问题，还能通过“自动化实验系统”与“科研数据管理架构”的结合，形成持续产生新数据的“数据飞轮”。

迈向“跨模态全场景理解”与“合成数据扩充”

生成式AI将突破单一模态的限制，利用多源异构语料构建更广阔的科学认知空间，并主动为科研生态生产数据。



数据来源：亿欧智库

目录

CONTENTS

1 人工智能语料的发展现状

1.1 人工智能语料发展规划
1.2 中国人工智能语料发展现状
1.3 中国人工智能语料发展阻碍与机遇

2 人工智能语料的行业需求特征

2.1 行业语料需求潜力象限分析
2.2 重点领域需求象限分析
2.3 新兴领域需求象限分析
2.4 语料需求潜力小结
2.5 AI4S语料特征、痛点与范式变化
2.6 AI4S语料特征、痛点与范式变化趋势

3 语料风云榜榜单

3.1语料风云榜榜单评选细则
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 人工智能语料优秀案例分析

4.1国家管网
4.2量子城市
4.3厦门供应链数智创新
4.4谱蓝科技
4.5脉搏深处
4.6飞数数据

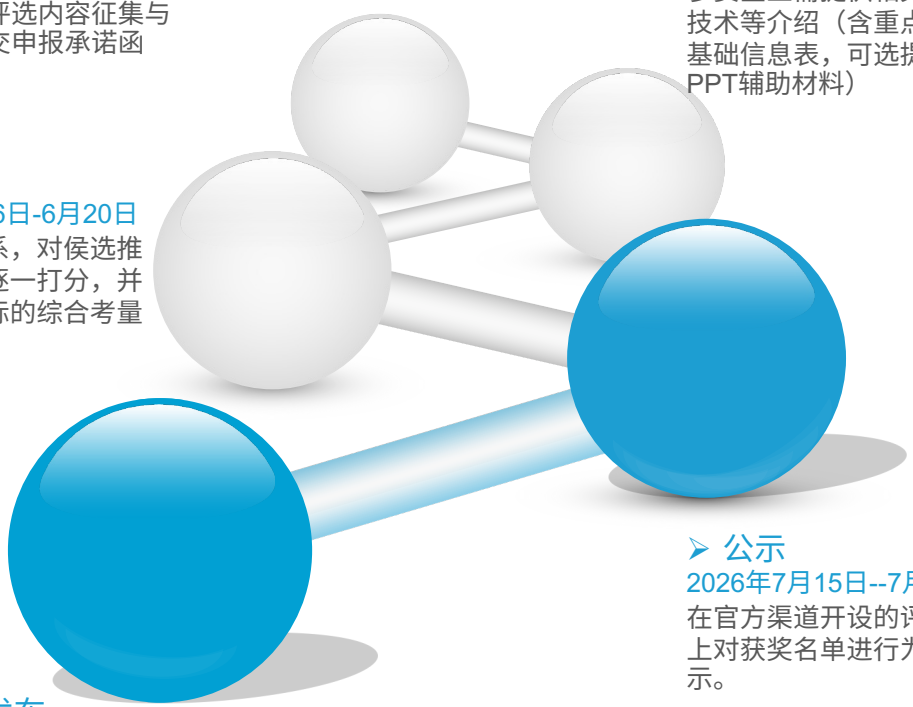
3.1 语料风云榜榜单评选细则

- 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2026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。2026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：2026语料风云榜榜单评选流程

➢ 申报
2026年5月21日—6月9日
申报。开展2026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

➢ 预选
2026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）



➢ 复评
2026年6月16日-6月20日
参考评价体系，对侯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

➢ 公示
2026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
2026年07月27日
在2026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“2026语料风云榜转型典范Top 10”，“2026语料风云榜新锐势力Top 10”。

亿欧智库：2026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

国家管网 PipeChina	国家石油天然气管网集团有限公司
中国交通信息科技集团有限公司 CHINA COMMUNICATIONS INFORMATION & TECHNOLOGY GROUP CO.,LTD.	中国交通信息科技集团有限公司
SIPG 上港集团	上海国际港务（集团）股份有限公司
华建集团 ARCPLUS	华东建筑集团股份有限公司
ESI 上海赛西	上海赛西科技发展有限公司
上海市数字城市规划研究中心 SHANGHAI DIGITAL CITY PLANNING & RESEARCH CENTER	上海市数字城市规划研究中心
TARS 它石智航	上海它石智航技术有限公司
ACE ROBOTICS	上海大晓无限机器人有限公司
数智创新 SMARTCHAIN INNOVATIONS	厦门供应链数智创新有限公司
HONKVANTA 横空科技	武汉横空科技有限责任公司

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

谱蓝科技 PLAN TECHNOLOGY	长沙谱蓝网络科技有限公司
FEIDATA	安徽飞数信息科技有限公司
SHP 数谱集团 SHP DataGRP	数巨有谱数据集团有限公司
LinkWise 临科智华	上海临科智华数字科技有限公司
BYTETREE 柏通数据	苏州柏通数据科技有限公司
EviDoc	上海脉搏深处科技有限公司
	湖南梦云科技有限公司
无锡快数智能科技有限公司	无锡快数智能科技有限公司
艺世科技 InnoVision Tech	上海艺世信息科技有限公司
joy new life	上海乐语新生科技有限公司

备注：榜单企业排名不分先后

目录

CONTENTS

1 人工智能语料的发展现状

1.1 人工智能语料发展规划
1.2 中国人工智能语料发展现状
1.3 中国人工智能语料发展阻碍与机遇

2 人工智能语料的行业需求特征

2.1 行业语料需求潜力象限分析
2.2 重点领域需求象限分析
2.3 新兴领域需求象限分析
2.4 语料需求潜力小结
2.5 AI4S语料特征、痛点与范式变化
2.6 AI4S语料特征、痛点与范式变化趋势

3 语料风云榜榜单

3.1语料风云榜榜单评选细则
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 人工智能语料优秀案例分析

4.1国家管网
4.2量子城市
4.3厦门供应链数智创新
4.4谱蓝科技
4.5脉搏深处
4.6飞数数据

4.1 国家管网：筑牢储运数据底座

- 国家管网面向核心业务场景，构建了“归集-开发-反馈-迭代”的闭环方法论，并确立了“场景牵引、价值优先、技术适配、流程合规”的建设导向。依托高质量数据集平台，国家管网创新性地将数据语料建设融入AI顶层设计，深度融合多模态数据，提供人机协同加工与精细化版本管理，实现全链路质量把控。
- 国家管网聚焦多模态数据全生命周期治理，确立“场景牵引、价值优先、技术适配、流程合规”的建设导向，创新性将数据语料建设融入AI应用顶层设计与架构标准。坚持发挥各企业差异化优势，构建内外协同的开放运营生态，重点面向场站阀室作业监测、内检测智能判图、压缩机故障预测、无人机智能巡检等高价值领域，推动多模态数据标准化入库、精细化标注与模型持续优化，实现数据资产与业务价值的深度耦合。

国家管网集团高质量数据集平台架构

构建一体化 AI 高质量数据底座，赋能储运行业高质量发展



国家管网集团高质量数据集服务

提供高质量数据集，对接可信数据空间，赋能AI创新

基础数据供给服务

聚焦国家管网核心业务场景，提供标准化、可直接复用的高质量数据集，适配客户基础数据需求，保障数据专业性与可用性，同步支撑数据要素市场化配置改革需求。

AI模型配套服务

结合国家管网AI应用实践，提供与AI模型全生命周期适配的数据集服务，助力客户提升AI模型性能，赋能油气行业智能化运维与决策优化，推动设备运维从经验驱动向数据智能驱动转型。

行业解决方案与增值服务

立足国家管网行业引领地位，提供一站式行业解决方案与增值服务，覆盖油气产业链全场景，推动行业数据标准化、智能化发展，实现数据价值最大化。

资料来源：国家管网

4.1 国家管网：筑牢储运数据底座

国家管网集团高质量数据集资源目录

全部 | D01战略规划到执行 | D02市场到回款 | D03工程建设 | D04生产运维 | D05安全环保 | D06流程与IT | D07研发 | D08供应链 | D10财经 | 多业务域



建设项目计价依据数据集

建设项目制程规范、企业制度、工程造价管理规定、计价依据文件的工程造价领域知识、专业名词、同义词对照表、业务枚举。

牵头单位： 创新公司
AI场景： 管智算智能体、智能对标智能体
应用方式： 数据挖掘/基础分析/知识库

D01战略规划到执行 | 造价分析数据集



储气库地下工程计价依据数据集

储气库工程制程规范、企业制度、工程造价管理规定、计价依据文件的工程造价领域知识、专业名词、同义词对照表、业务枚举。

牵头单位： 创新公司
AI场景： 管智算智能体、智能对标智能体
应用方式： 数据挖掘/基础分析/知识库

D01战略规划到执行 | 造价分析数据集



LNG接收站计价依据数据集

LNG制程规范、企业制度、工程造价管理规定、计价依据文件的工程造价领域知识、专业名词、同义词对照表、业务枚举。

牵头单位： 创新公司
AI场景： 管智算智能体、智能对标智能体
应用方式： 数据挖掘/基础分析/知识库

D01战略规划到执行 | 造价分析数据集



托运商数据集

企查查等工商信息数据库信息、实时天然气价格（包括LNG价格）、山东地区气象数据等外部数据。

牵头单位： 智网数科
AI场景： 市场洞察智能体-2.0
应用方式： 基础分析/知识库

D02市场到回款 | 市场洞察分析及客户管理数据集



上下载点数据集

企业基本信息，地区公司填报的核实信息，申请上下载点的信息；企业资质类文件，申请类附件，政府支持类文件，技术方案类附件等。

牵头单位： 华南公司
AI场景： 新增上下载点申请辅助审批智能体
应用方式： 基础分析/知识库

D02市场到回款 | 市场洞察分析及客户管理数据集



托运业务数据集

1.托运商结构化信息（企业基本信息、资质信息、法人信息、财务信息及经办人信息）及准入需要上传的证明文件（资质证明、经营许可、银行开户证明及其他辅助材料） ...

牵头单位： 华南公司
AI场景： 托运商准入辅助审批智能体
应用方式： 基础分析/知识库

D02市场到回款 | 市场洞察分析及客户管理数据集



阀室场景视频监测数据集

阀室场景图像数据及其标注信息，阀室安防监测，GOV阀状态监测等场景模型算法建设需要的数据信息。

牵头单位： 智网数科
AI场景： 阀室场景智能识别机器视觉算法
应用方式： 模型训练/机器学习/数据挖掘/基础分析/知识库

D04生产运维 | 油气储运机器视觉数据集



工地场景视频监测数据集

工地场景图像数据及其标注信息，作业坑安全通道堵塞检测等场景模型算法建设需要的数据信息。

牵头单位： 智网数科
AI场景： 工地场景智能识别机器视觉算法
应用方式： 模型训练/机器学习/数据挖掘/基础分析/知识库

D04生产运维 | 油气储运机器视觉数据集

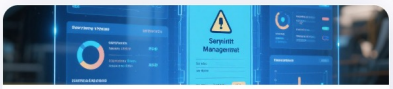


安环场景视频监测数据集

安环场景图像数据及其标注信息，未戴安全帽检测等场景模型算法建设需要的数据信息。

牵头单位： 智网数科
AI场景： 安环场景智能识别机器视觉算法
应用方式： 模型训练/机器学习/数据挖掘/基础分析/知识库

D04生产运维 | 油气储运机器视觉数据集



作业票证数据集

1.集成HSE 管控一体化平台数据，包括安全作业票的相关填报字段和检查记录。 2.作业票证的申请、审批、执行、关闭等全过程记录，包括票证类型、作业内容、安全措施、监...

牵头单位： 智网数科
AI场景： 远程监护辅助智能体-5.0
应用方式： 模型训练/机器学习/数据挖掘/基础分析/知识库

D05安全环保 | 场站巡检管理数据集



作业运营数据集

1.作业违章行为的记录，如违章类型、发生时间、地点、责任人、处理结果等。 2.重大风险的识别、评估、管控措施等数据，以及管控效果评估、持续改进计划等。

牵头单位： 智网数科
AI场景： 作业安全分析智能体
应用方式： 模型训练/机器学习/数据挖掘/基础分析/知识库

D05安全环保 | 作业安全分析数据集



JSA分析案例集

公司的安全作业管理平台、LNG物流平台、雷电预警系统、人员定位系统、各场站智能巡检系统、海南LNG数字孪生平台、北海LNG智能管理平台等等系统的历史数据，以及往...

牵头单位： 智网数科
AI场景： LNG接收站JSA智能分析智能体-5.0
应用方式： 模型训练/机器学习/数据挖掘/基础分析/知识库

D05安全环保 | 作业安全分析数据集

‹ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 › 第 13/19 页，共 110 条

资料来源：国家管网

4.2 量子城市：坤舆经略语料库

- 为贯彻国家发展新一代人工智能的战略部署，落实上海市《关于人工智能“模塑申城”的实施方案》中关于“建立一批通用和专用语料库，打造多层次语料体系，支撑基础大模型研发和垂直应用”的要求，上海市规划和自然资源局牵头推进国内首个超大城市规划资源管理专业语料库建设，旨在为城市治理领域大模型的研发与应用提供高质量数据基础，全面提升模型在空间识别、空间计算、空间理解、空间生成等方面的能力，推动城市治理向智能化转型升级。

坤舆经略语料库

- “坤舆经略语料库”作为在超大城市规划和自然资源领域具有开创性意义的专业语料库，以高质量、体系化、专业化的数据为基础，已深度支撑“云宇星空大模型”等智能应用的研发与落地，全面赋能上海量子城市的核心场景，展现了其作为城市空间智能领域核心数据基座的关键价值。
- 上海市规划资源局将持续深化语料库建设：在规模上，扩充各类语料库覆盖范围，实现规划资源领域全业务、全场景覆盖；在质量上，优化人工标注与评测体系，提升语料精准度；在技术上，面向“文、证、图、表”进行差异化的语料加工生产。同时，拓展语料库在多领域的跨界应用，促进数据融合与知识共享。未来，语料库将与量子城市空间智能建设体系下的各类模型和智能应用深度协同，提供数据保障，为城市治理智能化注入知识动力。






科技公司专家	业务专家				管理专家
	规划设计组	测绘地理组	自然资源组	行政管理组	
库帕思 商汤	市规划院、清华大学建筑学院、同济规划院、清华建筑设计院、同济建筑院、华建集团、北建院、雄大设计港、市园林院、城建院、市政总院、邮电院、电力设计院、上能院、上隧院、上铁院、水利院、上航院、联通	市测绘院、测绘质检站、超图	市资研院、市整理中心、市登记中心、市交易中心、市储备中心、局执法总队	局内各处室及相关事业单位	专家、局机关各业务归口部门负责人组成
语料工作整体推进 模型和语料双向衔接	知识库分类识别 语料高质量标注				高质量语料把关和指导

资料来源：量子城市

4.2 量子城市：坤舆经略语料库

“1+4+X”：规划和自然资源语料体系

- 搭建“1+4+X”规划和自然资源语料体系，形成“1”个系统化知识图谱，建设基石语料库、微调语料集、思维链语料集和评测体系等“4”类核心语料成果，以及“X”个专题领域语料库，构建起覆盖全领域、适配多需求、支撑高精准的语料生态。



1套系统化知识图谱

- 构建了贯通上海规划和自然资源管理和业务的完整知识体系，涵盖规划设计、自然资源、测绘地理等多个知识领域，形成16个一级分类、54个二级分类、165个三级分类的多层级结构。

4类核心语料成果

- 基石语料库：系统梳理学科教材、政策法规、方案征集、档案成果等规划资源数据资产，实现知识领域全覆盖，为模型预训练提供了高质量的基础语料支撑。
- 微调语料集：聚焦政策查询、知识问答、图纸理解及业务分析等模型核心能力提升，精准匹配训练需求，有效增强模型在关键业务场景的应用精度。
- 思维链语料集：围绕重点项目、关键业务与祼节问题，系统拆解并结构化呈现专业推理过程，为模型处理复杂任务、复刻专家思维提供了关键训练素材。
- 评测体系：建立科学量化的评估标准，对语料质量及模型应用效果进行闭环反馈，推动语料生产标准持续优化与生态体系自迭代。

X个专题领域语料库

- 面向检索增强精细化加工专题领域语料，以提升知识检索召回率。涵盖城市总体规划、城市设计、15分钟社区生活圈、品质市政、存量产业用地盘活、土地供给、生态复合单元和湿地空间、风貌保护管理、优化营商环境、沪派江南等70余个专题领域语料库，达成知识检索召回率稳定在95%以上，支撑模型实测问答能力点赞率超过90%。

资料来源：量子城市

4.2 量子城市：坤舆经略语料库

面向“文、证、图、表”做差异化语料算法



- 文：主要是各类公文，其行文方式、文本结构、目录层级、语言风格、解读要求都不同，需要采用不同类型的算法。如：政府公文/工作签报等文件有明确的目录层级和编写方式；政策解读本身就非常契合问答对的形式；法律法规/办事指南/标准规范等条文不能擅自更改；行业书籍内容太多又需要总结提炼。
- 证：针对建筑许可证、土地使用证、资源开发证、房本等不同的证照。采用“结构化拆分 + 语义标注 + 场景关联” 的思路处理，确保模型能理解许可证的法律要素、字段含义及适用约束。
- 图：规划领域的图则识别是规划领域读图的难点，图则信息密度大，图层复杂，元素多且业务指标信息多。包含了大量的专业信息，对于这类图的处理，需要采用的人工和自动化相结合的模式，通过多轮的评审完善。
- 表：对于表格的处理，需要针对表格进行专项的定制化算法优化，需要结合表格字段的业务属性，面向业务场景做专项的处理。

AI Ready 语料自动生产流程

- 以AI Ready为核心构建自动化语料流水线，形成日常办公文件的语料同步生产机制，推动语料的“生产-管理-使用”的全流程自动化，形成可生长、自迭代的语料生态体系。



资料来源：量子城市

4.3 厦门供应链数智创新：打造国际物流供应链可信数据底座

◆ 厦门供应链数智创新为供应链企业、物流企业、金融机构和研究机构提供可展示、可试用、可集成、可商业化的行业数据服务样板。



科技赋能供应链 · 数智驱动新生态 以高质量数据支撑人工智能进入真实供应链场景

标杆基础：龙头企业共建 × 产业数据汇聚 × 场景应用验证

3家	2.3万+	3大
世界500强供应链龙头联合组建	厦门供应链企业产业基础	农产品·铁矿石·多式联运先导场景

五阶段建设路径：让数据从“可用”走向“可调用、可验证、可进入经营”

L1	L2	L3	L4	L5
原始数据（结构化）	叙事型语料（非结构化）	分析型语料	循证语料	专家经验及技能
表格·数据库·API 统一字段与统计口径	报告·政策·行业资讯 切分重组并保留语境	预训练·SFT·问答对 沉淀指标关系与逻辑	证据链·推理路径·反证 结论可核验、可追溯	专家规则·模板·Skills 可调用、可执行、可复用

案例创新价值

从资源到资产

关键动作：统一目录、字段口径、质量规则
交付成果：数据集、字段字典、质量报告
业务价值：持续供给，减少重复取数与核验

从数据到证据

关键动作：关联来源、时点、支持方向与反证
交付成果：证据卡、证据链、专题数据包
业务价值：结论可核验、可解释、可追溯

从项目到产品

关键动作：标准化服务与企业定制并行
交付成果：订阅、API、图谱、Skills、工作台
业务价值：支持试用、集成及商业转化

资料来源：厦门供应链

4.3 厦门供应链数智创新：从可信数据供给到企业经营工作台

◆ 厦门供应链数智创新围绕五阶段建设路径，形成“可交付、可订阅、可集成、可定制”的产品体系，推动公司从数据资源组织者升级为供应链数据价值共创平台。

数据产品化	知识与证据化	企业场景化
解决数据有没有、能否持续供给	解决AI能否理解、结论能否核验	解决能力能否进入经营流程

01 基础数据供给服务

提供服务：汇聚价格、供需、库存、物流、宏观和政策数据，完成来源标注、清洗、标准化及数据集交付。
服务场景：模型训练、业务补数、品类研究和企业数据资产建设。
解决问题：解决数据分散、来源不清、字段和统计口径不统一、难以直接使用的问题。

02 数据订阅与接口服务

提供服务：按品类和主题提供日、周、月订阅，通过API/SDK持续供数，并提供更新监控与异常告警。
服务场景：业务系统接入、智能问数、行情监控、专题看板和经营分析。
解决问题：解决人工下载和重复搬运、更新不及时、系统接入成本高的问题。

03 AI-ready数据加工服务

提供服务：提供原始数据清洗、切分、标准化和标签化，并交付训练集、评测集、质量报告及版本记录。
服务场景：大模型预训练、SFT、RAG知识库建设、模型评测与持续迭代。
解决问题：解决原始数据无法直接用于AI、标签缺失、质量不稳和版本不可追溯的问题。

04 行业问答对及知识图谱服务

提供服务：建设行业问答对、分类标签、实体关系和知识图谱，并提供持续更新服务。
服务场景：供应链行业问答、关系检索、知识搜索、模型知识增强和业务规则查询。
解决问题：解决AI不理解行业术语、指标关系和业务逻辑，回答零散且缺少关联的问题。

05 证据包与专题数据产品服务

提供服务：围绕专题问题提供证据卡、证据链、专题数据包、来源引用和反证核验服务。
服务场景：市场研究、风险核验、事件跟踪、管理汇报和重大事项分析。
解决问题：解决结论缺少来源、数据口径混用、证据分散以及难以复核和追溯的问题。

06 智能体协同研判服务

提供服务：提供多智能体证据检索、反证检查、交叉复核、推理留痕和研判初稿生成服务。
服务场景：复杂市场分析、风险评估、事件影响研判、专题报告和人工复核。
解决问题：解决人工研究耗时、分析口径不一致、模型结论缺少解释和验证的问题。

07 企业定制化融合服务

提供服务：将企业合同、订单、客商及履约数据与外部数据进行主体匹配和融合，形成专题资产。
服务场景：客商准入、信用与风险评估、履约跟踪、经营分析和管理驾驶舱。
解决问题：解决内外部数据割裂、主体匹配困难、风险和经营全景不完整的问题。

08 业务专家Skills封装服务

提供服务：将专家规则、分析模板、操作步骤和处置经验封装为可触发、可执行、可评测的技能组件。
服务场景：市场研判、客商审查、风险审核、合同审查和业务操作辅助。
解决问题：解决专业能力依赖个人、判断口径不一致、经验难复制和难评测的问题。

09 企业定制化数据工作台服务

提供服务：整合数据、证据、知识图谱和专家技能，提供智能问数、风险预警、报告生成与工单协同。
服务场景：研究简报、风险处置、履约跟踪、经营分析和跨部门协作。
解决问题：解决跨系统查询、人工整理证据、流程脱节以及结论难沉淀复用的问题。

九类服务如何形成客户价值：先供数、再理解与研判、最终进入经营流程

01—03 数据供给
数据集 · 订阅/API · AI-ready加工
形成可信、持续、可调用的数据底座

→

04—06 知识证据
问答图谱 · 证据产品 · 协同研判
把数据转化为可核验的业务判断

→

07—09 企业应用
数据融合 · 专家Skills · 企业工作台
让能力进入真实经营流程并持续复用

场景验证：已面向农产品智慧供应链、铁矿石智慧决策链和多式联运智慧物流链开展应用；智能研判全程保留来源、反证及人工确认记录。

资料来源：厦门供应链

4.4 谱蓝科技：AI高质量数据解决方案供应商

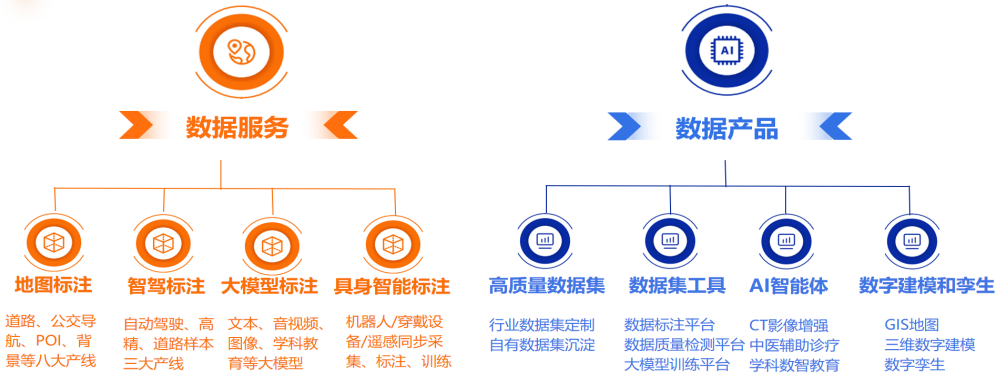
- 谱蓝科技是国内首批数据标注行业探索者，2015年从地理信息标注起步，深耕行业十余载。业务迅速覆盖AI大模型、具身智能、智能驾驶、地理信息四大核心领域。服务合作对象包括腾讯、美团、阿里云、抖音、科大讯飞、高德、百度等一众互联网头部企业，主流大模型覆盖率超过 90%。
- 依托深厚的数据标注行业积淀，公司率先落地各行业高质量数据集应用实践，深度参与能源、电力、交通、民航、应急、政务、医疗、教育、城市治理、工业制造等领域的数据集建设，已沉淀1000TB自有高质量数据集产品。针对行业普遍存在的工具零散、标准缺失、体系不完善等痛点，公司自主研发覆盖采集、预处理、标注、质检、封装——全流程的高质量数据集建设工具链，并内置百种AI算子能力，实现数据智能化预处理与智能标注，显著提升效率、降低成本。
- 面向未来，谱蓝科技将打造赋能产业智能决策的核心知识引擎，布局感知数据领域、构建全域数据基座，成就“智能时代核心数据基础设施领航者”的宏大愿景。



资料来源：谱蓝科技

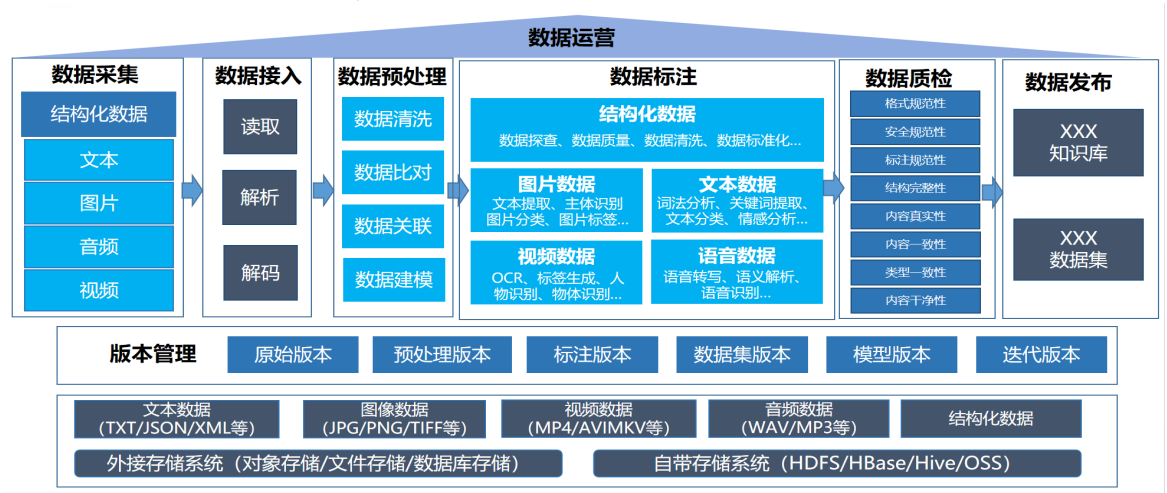
4.4 谱蓝科技：AI高质量数据解决方案供应商

谱蓝科技数据主营业务



谱蓝科技高质量数据集建设的全流程工具链

坚持“价值驱动，场景牵引”，实现数据集建设靶向聚焦，以行业真实业务痛点和应用场景为原点，构建从需求反推数据供给的闭环建设逻辑。



谱蓝科技行业高质量数据集产品



能源

- 电力设备安全巡检高质量数据集
- 油气生产现场监管高质量数据集
- 能源场站视觉识别高质量数据集



交通

- 综合交通空间规划通识数据集
- 基础设施专识高质量数据集



应急

- 安全生产隐患排查标注高质量数据集
- 应急事件处置流程知识高质量数据集
- 灾害场景风险识别高质量数据集



工业制造

- 装备制造设备健康与预测性维护高质量数据集
- 生产质量管控与缺陷追溯高质量数据集



政务

- 政务服务全流程办理要素高质量数据集
- 基层社会治理事件处置标注高质量数据集



城市治理

- 城市公共安全事件识别高质量数据集
- 市容环境问题巡查标注高质量数据集



医疗

- 中医临床诊疗案例高质量数据集
- 中医舌诊 / 面诊视觉识别高质量数据集



教育

- 智慧教学学情分析高质量数据集
- 教育资源知识图谱构建高质量数据集



文旅

- 文旅景区客流行为分析高质量数据集
- 游客服务智能问答语料高质量数据集

资料来源：谱蓝科技

4.5 脉搏深处：知识即服务，激活专家经验价值

- 脉搏深处科技成立于 2025 年 6 月，是库帕思 CVC 投资孵化的 AI 技术公司，定位为中国医疗领域知识基础设施平台，核心使命是破解国产医学大模型「缺乏高质量专业知识与语料」的难题。团队汇聚临床医学、人工智能与数据工程领域人才，入驻国家（医疗）人工智能应用中试基地（上海），并入选基地重点孵化项目。
- 脉搏深处以自主研发的专病知识逆向生成算法，把顶尖专家的诊疗经验转化为可追溯、可验证、可持续更新的专病知识资产，践行「传承于医、加速于药、福泽于人」的发展路径，服务临床能力传承、医学证据生成与药物研发、家庭健康场景解读。EviDoc 是这套能力对外服务的核心产品载体。

核心数据

10+	15+	500万+	70%
已合作顶尖科室	专病管线	高质量临床知识节点	RWE研究平均节省时间

脉搏深处远景：传承于医 · 加速于药 · 福泽于人

传承于医	加速于药	福泽于人
让顶尖专家能力从单点会诊，变成可复制的临床能力。覆盖医生训练、临床辅助、专病中心建设、医联体协同、临床科研。	让临床知识成为医学证据生成与药物研发的底层能力。覆盖真实世界研究、患者分层、适应症拓展线索、指南进入、医生教育。	为家庭健康硬件提供专病专家大脑式的医学解释层。覆盖体检、检测设备、可穿戴设备、家庭监测、慢病随访数据解读。

核心团队 来自复旦大学、上海交通大学等高校，深耕临床医学、人工智能与数据工程领域多年，具备多项AI研究与医学科技顾问背景。

愿景 让顶尖医学知识通过 AI 被沉淀、验证、调用、更新，最终抵达每一个需要它的人。

资料来源：脉搏深处

4.5 脉搏深处：知识即服务，激活专家经验价值

- EviDoc 以「决策链还原」与「推理型专病知识图谱」两项自主技术为核心，将分散于病历、医联体协同和医学文献中的专家隐性经验与权威循证知识，加工为可审核、可追溯、可调用的专家知识资产包，以知识即服务（KaaS）向医院、医联体、基层机构和医疗智能体开放调用。
- 脉搏深处联合库帕思共同构建：脉搏深处负责语料加工、图谱生产与工具支撑，库帕思负责方法论、产品化与商业交付。目前已触达 5 家医院、16 个科室方向、25 个病种方向，覆盖心血管、肾脏、血液、呼吸、消化、皮肤、眼科、肿瘤等专科。

EviDoc 能力架构：从数据要素到知识服务

数据要素层 S0	生产管线层 S1–S3	知识服务层 S4
临床真实世界数据 病历 / 检查 / 随访 / 影像 / 病理	决策链还原 临床实践管线 → 实践型知识单元	专家知识资产包 KaaS 标准 API
权威循证知识 指南 / 共识 / 文献 / 说明书	PICO 要素抽取 公开知识管线 → 循证型知识单元	风险边界校验 超范围 / 证据不足触发转人工
	推理型专病知识图谱 实体-关系-声明 + GraphRAG 召回	应用场景 院内MDT／居家管理／基层协同／智能体

EviDoc 知识生产链路：六环节闭环

语料接入 → 知识加工 → 图谱组织 → 资产包封装 → 接口调用 → 反馈更新

合规底线：原始数据不出院 · 敏感数据不外流 · 知识能力可调用

阶段性触达规模

5	16	25
已触达医院	科室方向	病种方向

资料来源：脉搏深处

4.5 脉搏深处：知识即服务，激活专家经验价值

◆ 以消化内科为试点，基于专家精选病例与权威指南 / 共识作为证据锚点，构建覆盖「病例事实—指南条款—证据等级—适用条件—处置决策」的全链路可追溯知识结构。

案例深度 | 中山医院消化内科：上消化道出血专病知识图谱

80 精选病例	10 部 指南/共识证据锚点	15,515 标准化实体	19,933 知识图谱关系
50 人 内测医生/学员覆盖	10,000+ 累计调用/查询次数	30 倍 交付效率提升	反超 SOTA 内部学术评测领先国际

全链路可追溯：从病例事实到处置决策

病例事实 → 指南条款 → 适用条件 → 处置决策

核心机制：专家证据锚定 + 结构化抽取 + 一致性校验 —— 不依赖海量无筛选语料输入，显著提升知识资产的精度与可审计性。

已验证应用场景

出科考　每日挑战　循证地图　药物对比

产业转化　已有企业基于该成果启动商业化开发，其中一家已正式签约入驻中山孵化器，标志着成果从科研验证阶段进入标准化生产与产业化孵化阶段。

该专病语料生产 SOP 及生产 / 验收标准已固化并启动团标申报，具备向其他专病复制扩展的规模化基础。

资料来源：脉搏深处

4.5 脉搏深处：知识即服务，激活专家经验价值

◆ 从单病种验证到多专科规模化拓展。

5	16	25
已触达医院	科室方向	病种方向

核心已验证专病方向（中山医院）

心内科	肾内科	血液科	呼吸科
心衰	IgA 肾病	慢淋 / 边缘区淋巴瘤	慢性咳嗽
同一套架构延伸验证	同一套架构延伸验证	同一套架构延伸验证	同一套架构延伸验证

医院—科室—病种触达矩阵

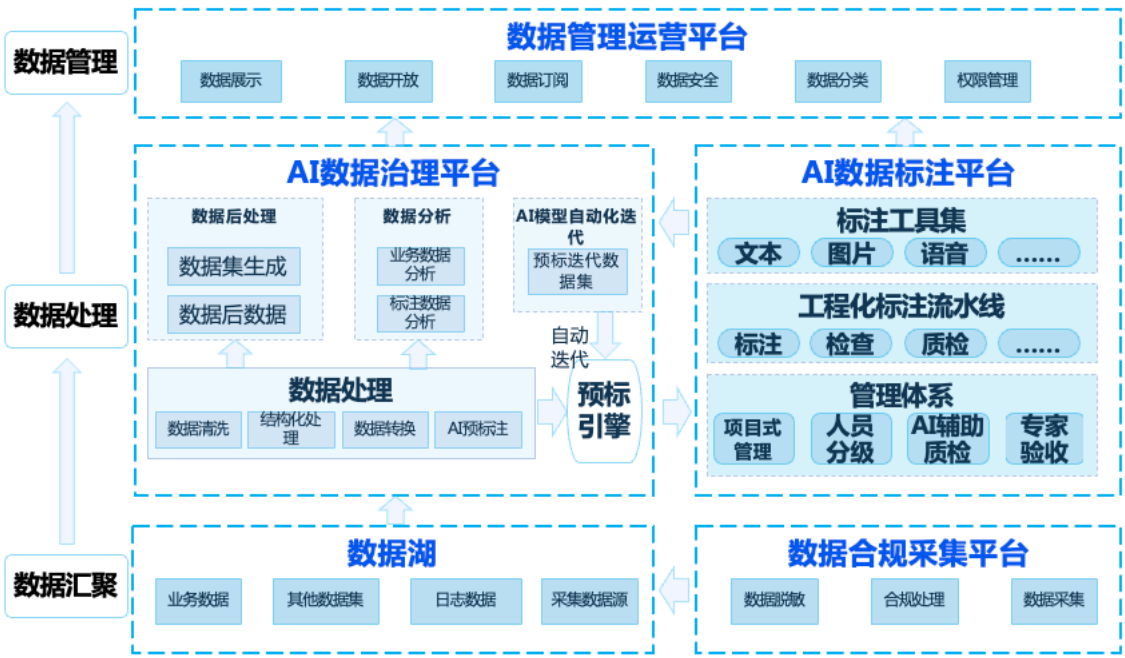
医院	科室方向	已触达专病方向
华山医院	皮肤科	特应性皮炎、脱发
市一医院	眼科	白内障
五官科医院	眼科	干眼病、角膜病
中山医院	呼吸科、肾内科、消化内科、心内科、血液科	慢性咳嗽、慢阻肺、IgA 肾病、急性胰腺炎、上消化道出血、心衰、边缘区淋巴瘤、CLL/SLL
肿瘤医院	妇科、淋巴瘤科、泌尿科、乳腺外科、神经外科、头颈外科、胸外科、胰腺外科	宫颈癌、卵巢癌、弥漫大B细胞淋巴瘤、NK/T细胞淋巴瘤、膀胱癌、前列腺癌、肾癌、乳腺癌、脑胶质瘤、甲状腺癌、食管癌、胰腺癌

资料来源：脉搏深处

4.6 飞数数据：人工智能高质量数据生产者

- 安徽飞数信息科技有限公司成立于2024年4月，是科大讯飞孵化的数据专业服务商，骨干团队均拥有10年以上数据专业服务经验。公司秉持“创造价值数据，赋能千行百业，驱动智能世界”的理念，专注于人工智能高质量数据集建设、数据要素全场景平台开发和企业数据治理服务，致力于成为国内领先的数据综合服务商。

- 安徽飞数是合肥市规上企业、合肥市数据产业协会秘书长单位，积极参与了合肥市国家数据标注基地建设，也是国家数据局关注的民营数据企业，同时积极承担国家数据基础设施建设项目。2025年数据生产规模超5PB，年产值达2.7亿元。公司已在合肥、重庆、武汉、成都、西安、长沙、兰州、长春等10城设立标注职场，链接国内外100余家生态合作伙伴和60万众包人员，初步形成了立足合肥、覆盖全国、辐射一带一路的数据综合服务网络。飞数数据深度合作伙伴覆盖汽车、安防、手机、家居、金融、教育、新零售、生态系统等行业。包含众多世界500强企业、高校科研机构、政府机构、头部AI企业和大型互联网企业，涵盖计算机视觉、语音识别、自然语言处理、知识图谱等AI主流技术领域。



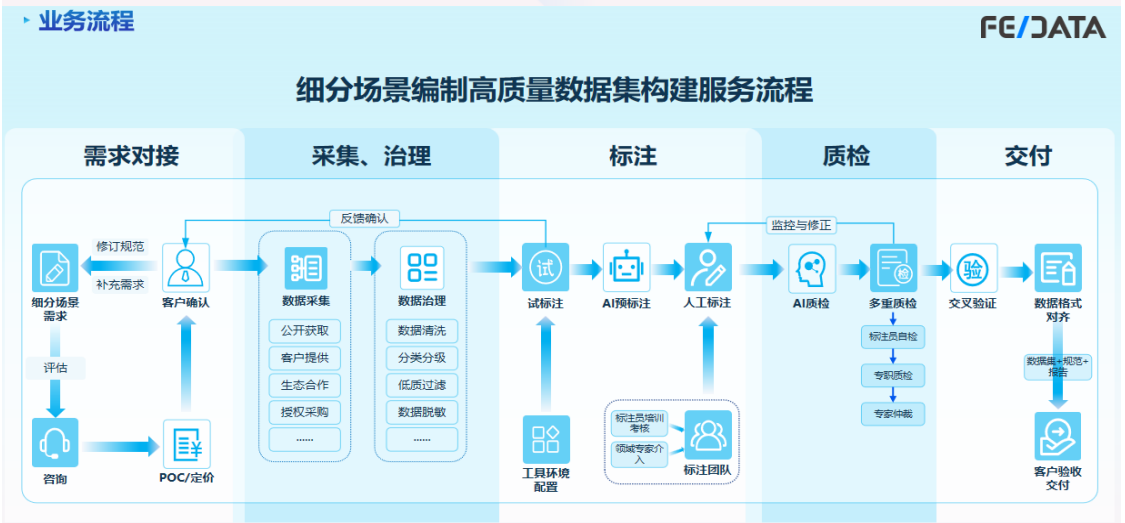


资料来源：飞数数据

4.6 飞数数据：人工智能高质量数据生产者







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成立于2024年3月，是为了贯彻落实国家人工智能战略，繁荣产业生态而成立的功能性和市场化双轮驱动的语料运营枢纽平台，定位于“AI原生的语料数据新基建运营商”，聚焦“建基座、强技术、搭平台、育生态”四项任务，推动“语料数据方法体系、语料数据基础设施、语料行业生态体系”三方面重构，在语料行业系列标准体系、语料运营公共服务平台、语料作价入股创新等方面形成一系列全国首创，并持续面向基模、垂模和智能体开发者提供“AI Ready”的高质量语料数据服务。

- 公司组建以来，搭建了“1125”的语料工作框架。构建“1”个语料基座，统筹自建与链接，布局“1+X”，即1个综合语料库+X个行业语料库，力争做到语料供给既能高效服务基础模型，也能有效降低垂类模型的应用成本；构建“1”个语料运营平台，构建“采、洗、标、测、用”五位一体的工具体系和自动化流水线；构建“2”大战略基础体系，一是立足世界模型演进趋势研究基于世界知识体系的语料构建方法，打造“语料魔方”体系；二是多维度价值对齐的语料体系，对标模型“五有”（有德、有趣、有序、有品、有用）建设基石语料库和测试集；构建“5”项可持续发展能力，即智能清洗与质量保障、智能合成与价值对齐、场景适配与知识修正、可信流通与跨域治理、生态服务与产业赋能等五方面能力。

- 目前，公司对照“模塑申城5+6”整体布局，已经全面启动金融、医疗、工业、教育、具身智能和科学智能等领域的行业语料库建设，按照储备一批、转化一批、应用一批，稳步推进数据语料化工作，不断深化语料FDE工程，推动数模更深层次融合应用。创新产教融合机制，打造语料标注和应用人才“培训-考核-发证-见习-就业”闭环，依托语料基金和语料作价入股培育“建用协同”生态，按照“一业一方法、一业一指引”持续推动语料系列标准研制，已经链接了超过300家的语料生态合作伙伴，大模型语料数据联盟实现了实质性的“提质扩容”。

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的动态和前沿趋势，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

39

亿欧智库

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