

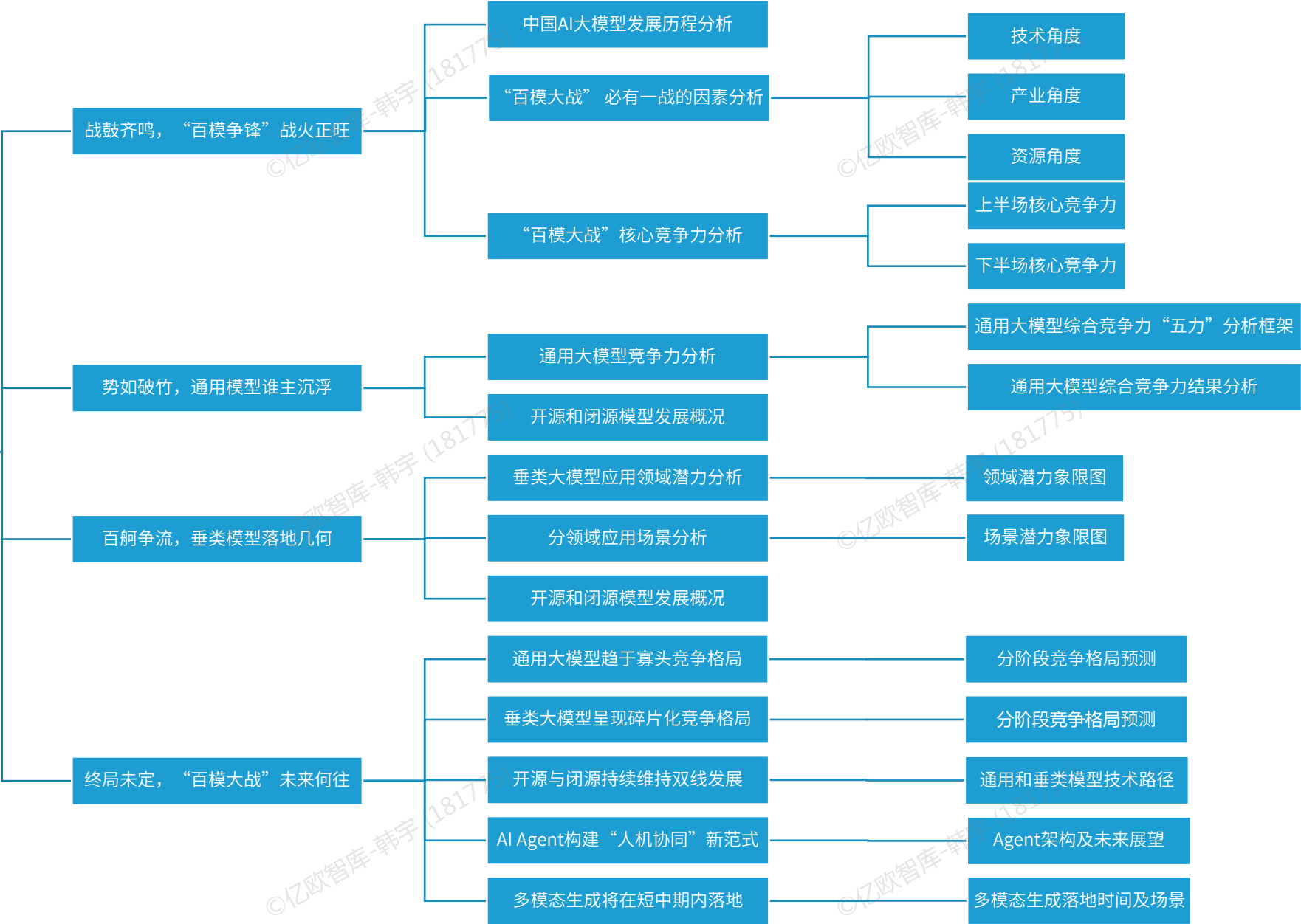
战鼓震天动地响，铁蹄踏破万里疆

-2024中国“百模大战”竞争格局分析报告

亿欧智库 <https://www.iyiou.com/research>

Copyright reserved to EO Intelligence, March 2024

2024中国“百模大战”竞争格局分析报告



目录

CONTENTS

01 战鼓齐鸣，“百模争锋”战火正旺

- 1.1 中国AI大模型发展历程分析
- 1.2 “百模大战”必有一战的因素分析
- 1.3 “百模大战”核心竞争力分析
- 1.4 大模型基础层图谱

02 势如破竹，通用模型谁主沉浮

- 2.1 通用大模型图谱
- 2.2 通用大模型竞争力分析
- 2.3 开源和闭源模型发展概况

03 百舸争流，垂类模型落地几何

- 3.1 垂类大模型图谱
- 3.2 垂类大模型应用领域潜力分析
- 3.3 开源和闭源模型发展概况
- 3.4 优秀企业案例

04 终局未定，“百模大战”未来何往

- 4.1 通用大模型趋于寡头竞争格局
- 4.2 垂类大模型呈现碎片化竞争格局
- 4.3 开源与闭源持续维持双线发展
- 4.4 AI Agent构建“人机协同”新范式
- 4.5 多模态生成将在短中期落地

2023 - 2024年中国AI大模型发展历程

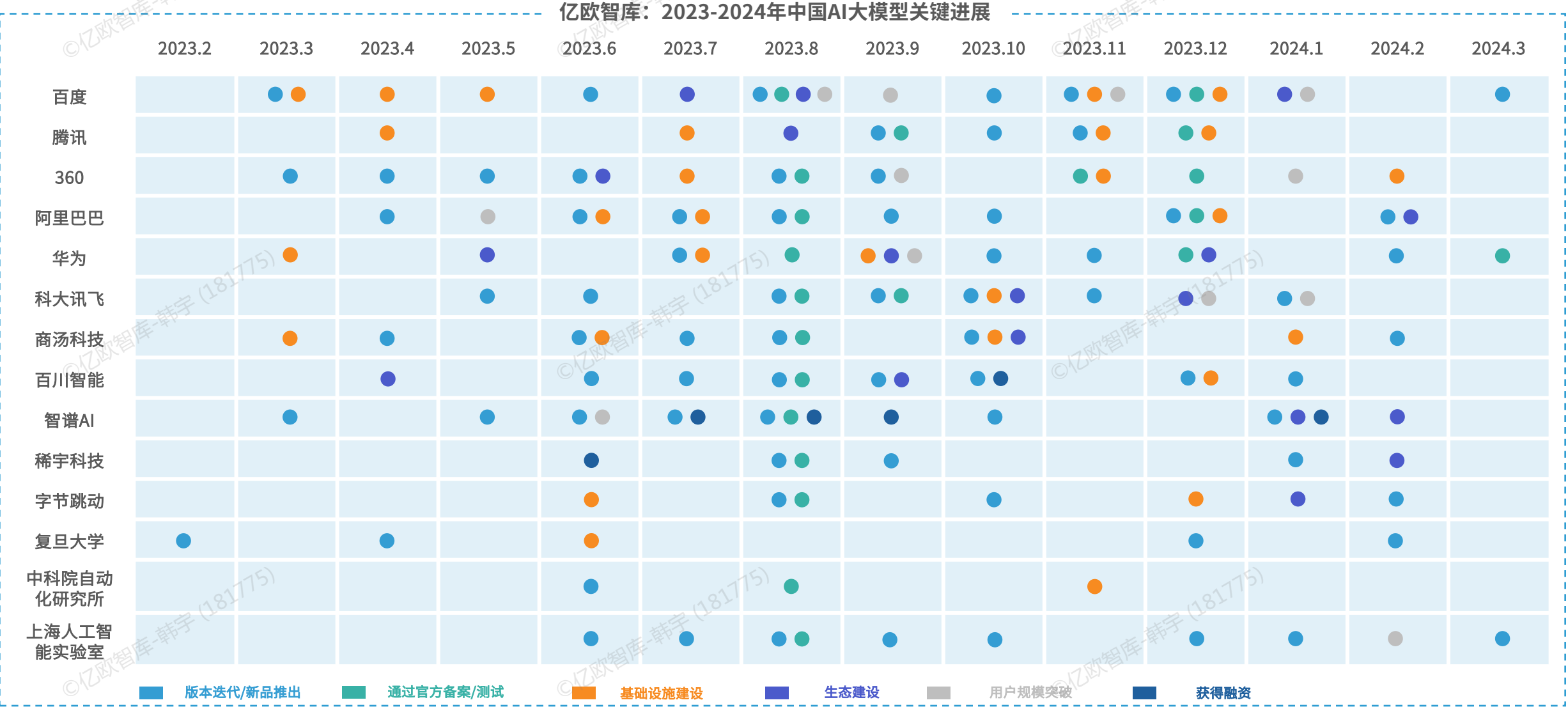
◆ 2022年12月ChatGPT-3.5的发布引发了全球范围内对大模型的广泛关注，2023年2月国内高校率先发布产品，2023年3月开始以百度、阿里、360和商汤等为代表的企业陆续推出各自的大模型，2023年6月开始国内涌现出众多的通用大模型和垂类大模型。目前国内公布的大模型数量已超过300个，市场竞争激烈，大模型行业呈现出“百模大战”的竞争格局。



资料来源：《中国人工智能大模型地图研究报告》、公开资料、亿欧智库

2023 – 2024年中国AI大模型关键进展

◆ 2023 – 2024年，中国大模型产业快速发展，科技大厂、AI公司、创业公司和研究机构等先后入局。各类玩家在产品迭代、官方备案、基础设施建设、生态建设、用户建设和融资方面积极推进并取得进展。



资料来源：公开资料、亿欧智库

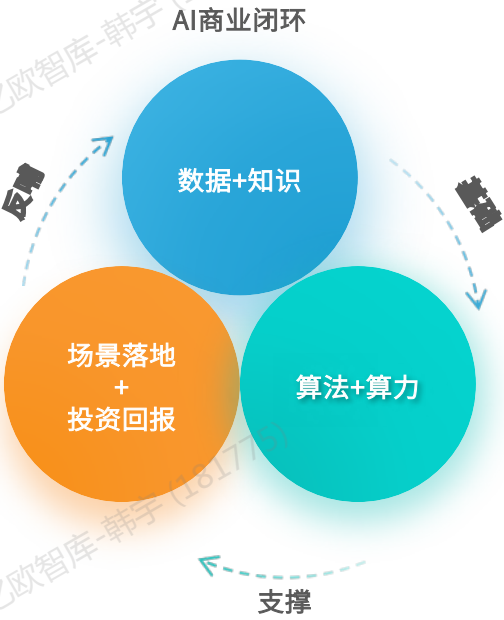
◆ 从20世纪50年代人工智能技术诞生至今，先后四次引发大规模的关注，分别是专家系统时代、机器学习时代、深度学习时代和大模型时代。前三个阶段由于数据、算力、场景落地和投资回报等方面的原因未能形成AI产业的商业闭环。基于大模型通用性、涌现性和强计算能力的特征，大模型时代有望形成商业闭环，因此基于技术可行性和未来巨大想象空间，众多玩家涌入大模型赛道。

亿欧智库：大模型时代技术进步有望实现商业闭环

AI发展史中四次引发社会大规模关注的时代

数据+知识、算法+算力、场景落地+投资回报构成了AI产业商业闭环的三个核心环节，在AI商业闭环中：

- 数据和知识是基础
- 在算法和算力的支撑下实现场景落地，下游客户投资场景获得正向回报
- 场景落地后叠加客户持续投资又为产业输送新的知识和数据
- 闭环形成后，AI产业将实现长期有序发展



专家系统时代

- 专家系统是早期人工智能的代表，其核心思想是将领域专家的知识通过规则和推理方法进行表示和处理
- 专家系统通过将专家知识编码为一系列的规则或者知识库，然后利用推理引擎来模拟专家的决策过程，以解决特定领域的问题

知识构建成本高
知识表达困难
技术泛化能力弱

数据+知识

仅完成商业闭环中的“数据+知识环节”

机器学习时代

- 机器学习是基于数据驱动的人工智能方法
- 核心思想是通过从数据中学习模式和规律，从而使计算机系统具有智能行为

应用场景有限
训练数据依赖性强
算力资源有限

数据+知识

算法+算力

走到商业闭环中的“算法+算力”环节

深度学习时代

- 深度学习是机器学习的一个分支，其核心是使用深层神经网络来进行学习和模式识别
- 深度学习利用多层次的神经网络结构来学习高层次的抽象特征表示，从而实现了对复杂数据的有效表示和处理

场景泛化能力弱
场景投资回报低

数据+知识

场景

算法+算力

场景投资回报低，未形成商业闭环

大模型时代

- 大模型是指具有庞大的参数规模和复杂程度的机器学习模型，具有通用性、涌现性和强计算能力

场景泛化能力强
技术涌现能力强
数据和算力质量提升
多任务属性提高投资回报率

数据+知识

场景+回报

算法+算力

有望形成商业闭环

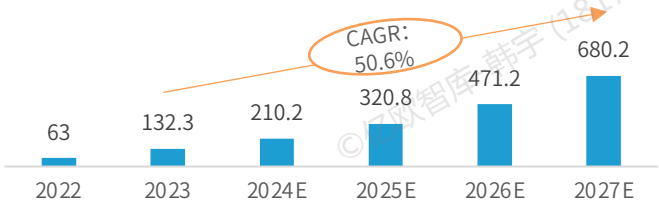
“百模大战”必有一战的因素分析 - 产业角度

◆ 市场规模持续扩大叠加资本市场关注度不断提升，日益火热的大模型赛道涌入各类玩家。从新兴技术发展周期视角来看，历经多玩家涌入和投资热度提升的期望膨胀期后市场会逐渐出清，战况激烈的“百模大战”将在短中期内步入该阶段。因此从产业发展规律角度来看，大模型赛道目前已走过创业高峰期，随之而来的市场出清趋势将激化各产业玩家间的竞争，推动新一轮的市场洗牌。

亿欧智库：市场火热引发众多玩家入局

- 中国大模型市场发展迅速，2023年中国语言大模型市场规模突破**100亿元**，未来将保持高速增长，随着机器视觉、多模态大模型等技术的持续成熟，未来大模型市场规模有望持续增长。2023年人工智能领域的资本热度相对下滑，**但AIGC/大模型在资本市场的关注度持续提高，并出现了智谱AI、百川智能等融资金额超过20亿人民币的事件。**
- 科技领域FOMO（Fear Of Missing Out）错事恐惧效应明显，叠加大模型成为炙手可热的赛道，AI企业、科技公司、初创企业和研究机构等纷纷入局，**国内大模型数量已从2023年6月的不足80个增长到2024年2月的300多个**，市场竞争激烈。

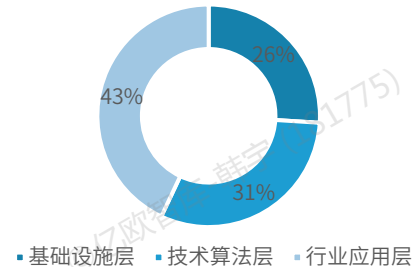
中国大语言模型市场规模（亿元）及增速



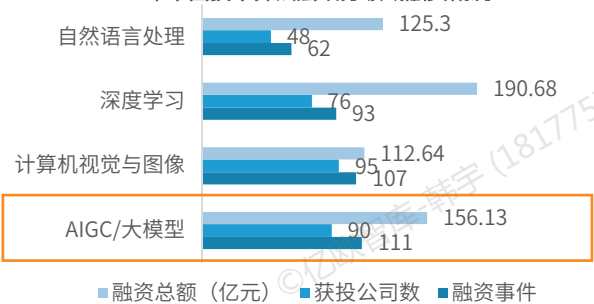
- 2023年中国语言大模型市场规模达**132.3亿元**，未来将保持高速增长
- 预计2027年突破600亿元，2023 - 2027的复合年均增长率达**50.6%**

- 2023年AI领域融资事件总数为585件，技术算法层融资事件182件，占AI领域的31%
- 在技术算法层细分领域的融资中，**AIGC/大模型累计融资156.13亿元，获投公司数90家，融资事件数111件**，在包括基础设施和行业应用的所有AI细分领域中，三项指标均领先与行业平均水平

2023年中国AI领域一级市场股权融资事件类型分布



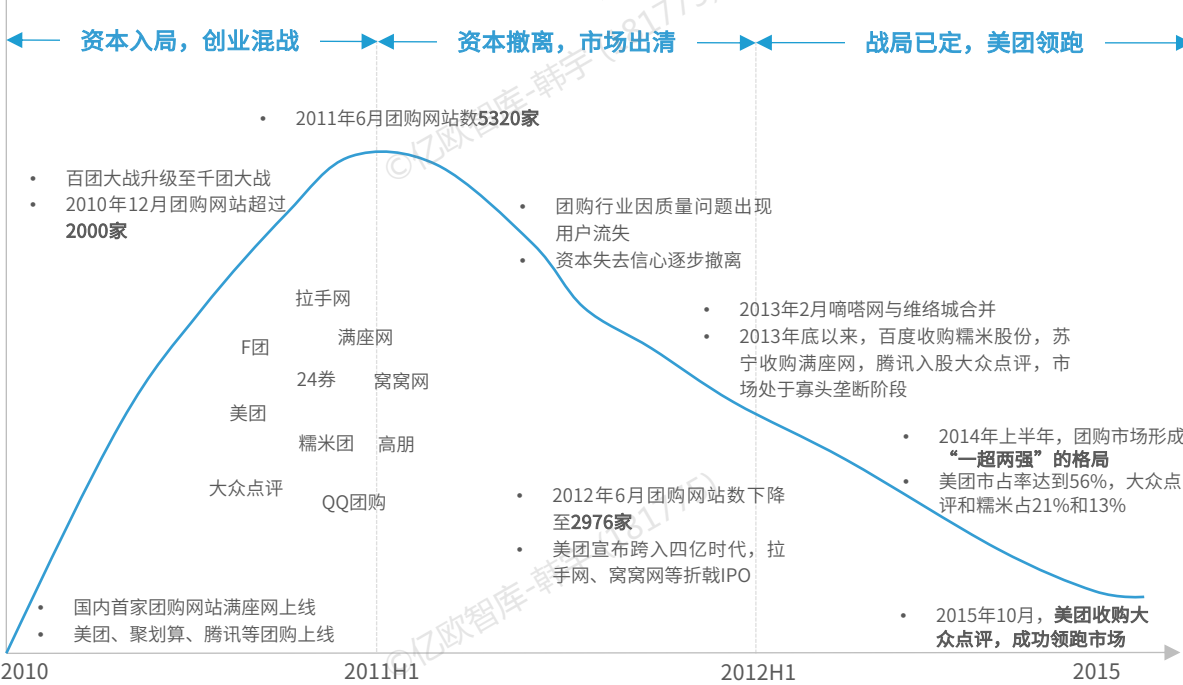
2023年中国技术算法层细分领域融资概况



亿欧智库：产业发展规律引致竞争格局优化

- 新兴技术的发展通常会经历技术萌芽期、期望膨胀期、泡沫破裂低谷期、稳步爬升复苏期和生产成熟期，历经多玩家涌入和投资热度提升的期望膨胀期后市场逐渐出清，头部玩家成为市场的主导角色。
- 2010 - 2015年互联网行业爆发的千团大战历经资本大量涌入的创业混战期、资本撤离后的市场初步洗牌期和行业整合并购期，最后走向先对稳定的寡头竞争格局。**百模大战对比千团大战在竞争策略、商业模式和技术影响等方面存在差异，但混战后趋于稳定的市场规律仍具有借鉴意义。**

2010 - 2015年互联网“千团大战”历程



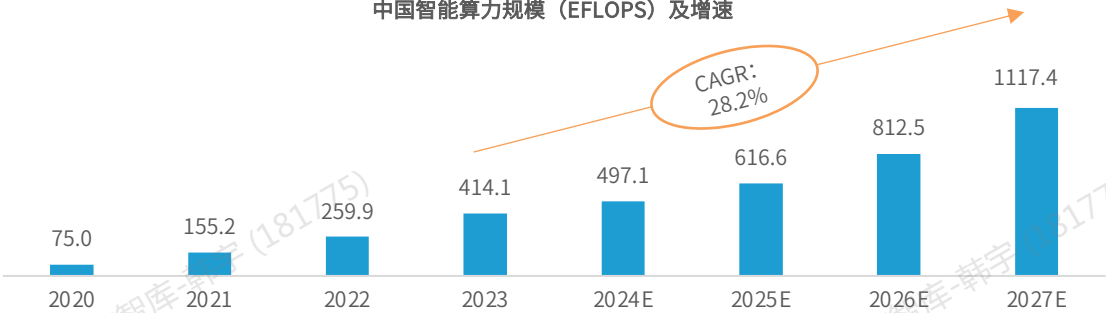
资料来源：IDC&浪潮《2023 - 2024年中国人工智能算力发展评估报告》、公开资料、亿欧智库

◆ 大模型赛道处于高拥挤状态，有限性是算力等资源的本质特征。随着市场竞争的日益激烈，算力资源会呈现供需失衡的状态并出现资源配置倾斜，叠加日益上涨的算力成本，行业玩家只有积极应战并构建起竞争优势，才能获得产业长期发展的资源支持。

亿欧智库：算力供需失衡引发资源配置倾斜

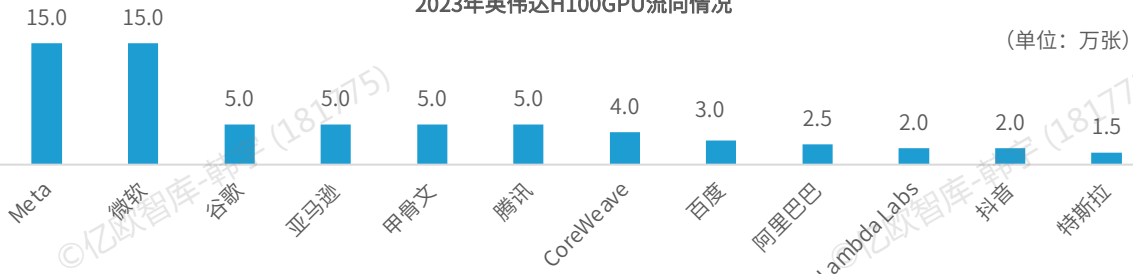
- 工业和信息化部预测，到2024年年底我国将有5% - 8%的企业大模型参数从千亿级跃升至万亿级，**算力需求增速会达到320%**，而**中国智能算力规模2023 – 2027年复合年均增长率仅为28.2%**，将出现算力供需不平衡。

中国智能算力规模（EFLOPS）及增速



- 英伟达A100/H100/A800/H800等**高算力专业级芯片是业内大模型训练的首选**。
- 大模型产业爆发后专业级GPU供不应求，美国对中国的芯片封锁增加了国内公司购买芯片的难度，目前可通过英伟达合作伙伴网络内的代理商或英伟达OEM整机合作伙伴采购GPU，两种形式都倾向于优先供货于阿里和字节等订单量大的大厂，**2023年英伟达H100 GPU在国内大部分流向字节跳动、阿里、百度和腾讯等大厂，中小公司和创业公司芯片购买难度增加**。

2023年英伟达H100GPU流向情况



亿欧智库：成本高昂加速市场出清

- 大模型训练和推理阶段都需要消耗较高算力，训练阶段GPT3.5同级别模型需要消耗数千张A100 GPU，**GPT4同级别模型则需耗费近万张A100 GPU**。在推理阶段，**仅考虑问在问答场景需消耗数万张H100**，多模态使用场景下算力消耗量随之增加。
- 市场需求高涨叠加产能限制导致GPU市场严重供需失衡，**高性能GPU价格持续增长**，2023年5月A100和A800市场单价分别达到15万元/张和9.5万元/张，显著提高大模型产业入局和维持成本。

部分大模型消耗卡的数量及训练时长

模型	消耗卡数量	训练时长
LLaMA-65B模型	2048张A100 80GB GPU	21天
GPT3-175B模型	1024张A100 40GB GPU	24天
GLM-130B模型	768张A100 40GB GPU	2个月
Falcon-40B模型	384张 A100 40GB GPU	2个月
Inflection模型	3500张H100 GPU	-
Inflection-2模型	5000张H100 GPU	-
Megatron-Turing模型	4480张A100 GPU	-
GPT4-1800B模型	25000张A100 GPU	90-100天

A100



15万元/张
2023年5月市场价
37.5%
2022年12月份至2023年4月累计涨幅

A800



9.5万元/张
2023年5月市场价
20%
2022年12月份至2023年4月累计涨幅

资料来源：Omdia、中泰证券、国盛证券、IDC、公开资料、亿欧智库
图片来源：英伟达官网

- ◆ 大模型是一个重资源禀赋和高进入门槛的赛道，“百模大战”是一场重投入和高消耗的持久战役。“百模大战”的上半场是资源和技术的碰撞，入局企业需要长久的资源支持获得高密度的人才、高质量的数据和大规模的算力，同时需要持续的研发投入，以科技创新和技术积累构建竞争壁垒。
- ◆ 在“百模大战”下半场的竞争中，需要重点关注商业落地能力和客户投资回报价值。创新商业模式形成商业闭环实现公司有序发展，同时提升大模型技术为下游客户创造的实际价值，持续巩固企业的竞争优势。

亿欧智库：“All in AI”的资源支持

- 作为技术密集型产业，进入大模型赛道和在激烈的市场竞争中获得优势都需要**大量资金支持**，用于**人才团队构建、高质量训练数据获取和大规模算力保证**，“All in AI”的资源支持和战略方向是在“百模大战”中持续生存的关键要素之一。



资金

- 大模型的**研发、训练及推理**需要大量的资金投入，需要一定规模的**市场推广费用**在市场竞争中保持声量
- 研发团队中，计算机视觉、深度学习、语音识别和自然语言处理等领域的人才**薪酬高于平均水平**
- **高质量数据集的构建**需要投入大量的资金，覆盖数据采集、清洗、标注和处理等流程的成本
- **算力成本居高不下**，GPT-3同级别模型训练一次的成本超100万美元



人才

- 大模型是典型的技术密集型产业，需要**高密度的人才支持**
- 大模型的开发和应用需要跨学科的专业知识，包括机器学习、数据科学、软件工程、系统架构等领域的专家，**人才的多样性和专业性**是推动技术创新和应用落地的关键



数据

- 数据是训练大模型的基础，高质量的数据集对于模型性能至关重要；拥有**大量、多样化且高质量的数据集**，能够使模型更好地理解 and 生成语言，提高其在特定领域的应用效果
- 目前海外高质量数据集数量较多，**国内开源数据集在数据规模和语料质量上仍有较大差距**，厂商自有数据和数据处理能力是形成差异化优势的关键



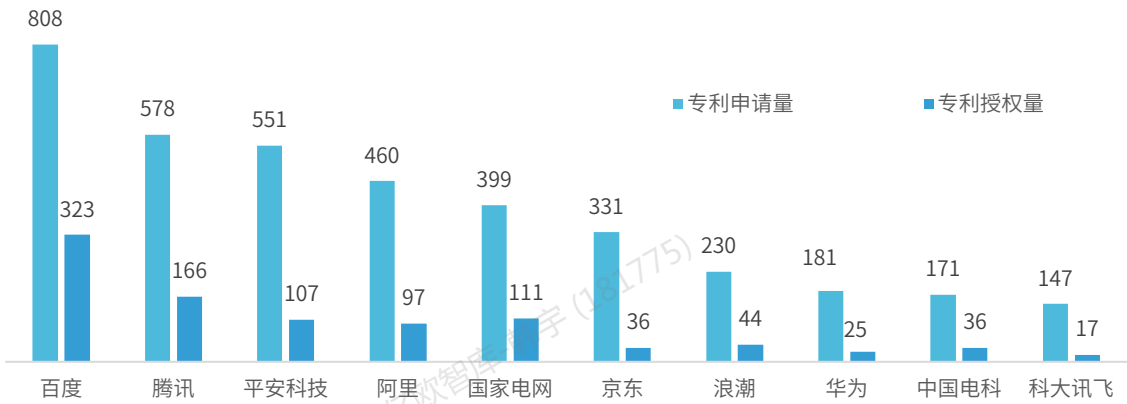
算力

- 算力是发展大模型产业的必备基础，随着模型参数和复杂度的提升，大模型对算力的需求也在加速增长，**大模型训练所需算力每3-4个月增长1倍，增速远超摩尔定律**
- 大模型训练和推理必须的**专业级GPU目前处于供不应求的状态**，资源主要倾斜于实力雄厚的大厂；对于采取算力租赁形式切入赛道的玩家，**市场需求旺盛推动租赁成本日益高涨**

亿欧智库：科研创新与技术积累

- 大模型产业的技术进步来源于科研创新和技术积累，从全球范围来看中国、欧洲和美国在AI for Science领域的积累较多，三地AI for Science论文发表量超过全球总数的80%，**我国AI for Science论文发表数量最高**。聚焦国内除清华大学、北京大学和浙江大学等科研院校外，华为和百度等机构都在积极推动全球AI for Science发展。
- 除前言理论研究外，专利的积累和沉淀也能有效帮助企业构筑竞争壁垒。**截至2023年底，中国大模型专利申请总数突破4万件**，百度、腾讯和阿里等头部科技企业积累了大量创新成果。专利数量持续增加表明中国在大模型领域的科研投入已逐步形成技术积累，科研创新能力持续提升。目前大部分专利主要集中在互联网和科技企业等大厂，初创企业的专利积累相对较少。

中国AI大模型专利企业创新主体排名



资料来源：《中国AI for Science创新地图研究报告》《中国AI大模型创新和专利技术分析报告》、公开资料、亿欧智库

- ◆ 商业落地能力是连接技术创新与市场需求的桥梁，是企业实现可持续发展的关键。在“百模大战”中，商业落地能力是决定企业能否长期保持竞争力和市场领导地位的重要因素。
- ◆ 技术实现商业落地后需要密切关注和有效提升产品为下游客户创造的价值，不断提高产品附加的战略价值、经济价值和成本价值，以提高客户投资回报为手段提升客户粘性和市场竞争力。

亿欧智库：商业落地能力

- 大模型应用领域广泛，可面向企业和个人实现变现。按量收费、订阅模式、定制化服务和广告变现是目前常见的商业模式，在**长期竞争中合理且符合企业自身发展条件的商业模式**是大模型厂商保持竞争优势的关键要素。

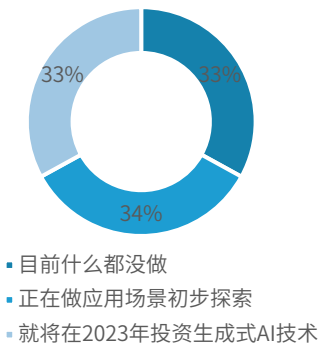
AI大模型商业模式列举

客户类型	商业模式	定义及案例
B端、C端	按量收费	- 定义：根据用户使用量进行收费，比如按tokens计费、按照API调用次数收费，按照产出内容量收费 - 案例：科大讯飞目前对外开放了星火大模型V1.5、V2.0和V3.0等三个版本，价格分别为0.18元/万tokens、0.36元/万tokens、0.36元/万tokens；AI图像生成平台大部分按照图像张数进行收费
B端、C端	订阅模式	- 定义：依托内容生成平台和工具，通过收取使用费用获得收益 - 案例：百度大模型文心一言已上线会员模式，定价为59.9元/月，连续包月优惠价为49.9元/月，开通会员后，可解锁文心大模型 4.0
B端	定制化服务	- 定义：根据客户的特定需求，提供定制化的大模型服务，服务内容不仅包括大模型系统，还包括服务器等硬件产品等一站式解决方案 - 案例：为高校提供的大模型系统教学实践平台，不仅包括需要大模型系统，还需要包含服务器等硬件产品
B端	广告变现	- 定义：当大模型终端平台或工具达到一定体量后，相关商家会选择在其上进行广告投放，大模型厂商获得广告费用 - 案例：Ask AI在App Store中下载量巨大，吸引了商家投放广告，每月获得数百万美元收入

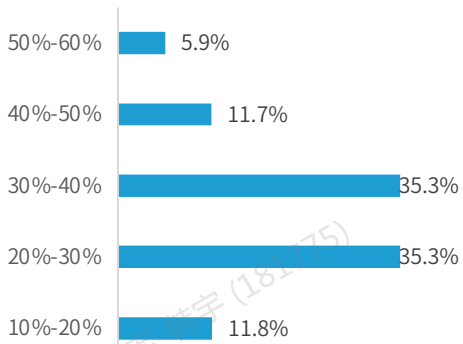
亿欧智库：投资回报价值

- 中国企业对于生成式人工智能的接受度较高，一半以上的企业计划已开始相关尝试，其中**34%企业正在做一些应用场景的初步探索**，**33%的企业将在2023年投资生成式AI技术**。对于有投资生成式AI技术意愿的企业，35.3%的企业在未来3年对于生成式AI的投入增幅在20%-30%之间，35.3%的企业预计增加30%-40%的投资。
- 整体来看国内企业对于大模型的接受度和投资意愿度较高，**但当大模型技术商业落地后，需要考虑技术为下游客户创造的实际价值**，不断优化大模型技术为客户创造的战略价值、经济价值和成本价值，以价值提升为手段增加客户粘性，从而实现企业长期有序发展。

中国企业对生成式AI的态度



未来3年公司在生成式AI的投入增幅



大模型为下游客户创造的价值





注：图谱未详尽，图谱中的成本占比指构建大模型时不同基础层环节投入的占比

目录

CONTENTS

01 战鼓齐鸣，“百模争锋”战火正旺

- 1.1 中国AI大模型发展历程分析
- 1.2 “百模大战”必有一战的因素分析
- 1.3 “百模大战”核心竞争力分析
- 1.4 大模型基础层图谱

02 势如破竹，通用模型谁主沉浮

- 2.1 通用大模型图谱
- 2.2 通用大模型竞争力分析
- 2.3 开源和闭源模型发展概况

03 百舸争流，垂类模型落地几何

- 3.1 垂类大模型图谱
- 3.2 垂类大模型应用领域潜力分析
- 3.3 开源和闭源模型发展概况
- 3.4 优秀企业案例

04 终局未定，“百模大战”未来何往

- 4.1 通用大模型趋于寡头竞争格局
- 4.2 垂类大模型呈现碎片化竞争格局
- 4.3 开源与闭源持续维持双线发展
- 4.4 AI Agent构建“人机协同”新范式
- 4.5 多模态生成将在短中期落地

通用大模型图谱

业务落地

 文心一言	 讯飞星火	 天工大模型	 星辰语义大模型	 Yi 开源模型
 混元大模型	 云天天天书	 紫东太初	 Dongni	 XVERSE
 360智脑 360奇元	 封神榜系列	 盘古	 孟子 GPT	 元乘象 ChatImg
 通义千问	 西湖	 日日新 商量	 CharacterGLM	 元识
 夸克大模型	 ABAB	 智谱清言	 云雀大模型	 AndesGPT
 书生·多模态 书生·浦语	 Kimi Chat	 源	 百川大模型	 面壁露卡Luca
	 海河·谛听	 悟道大模型	 小冰大模型	 百灵大模型

业务试点

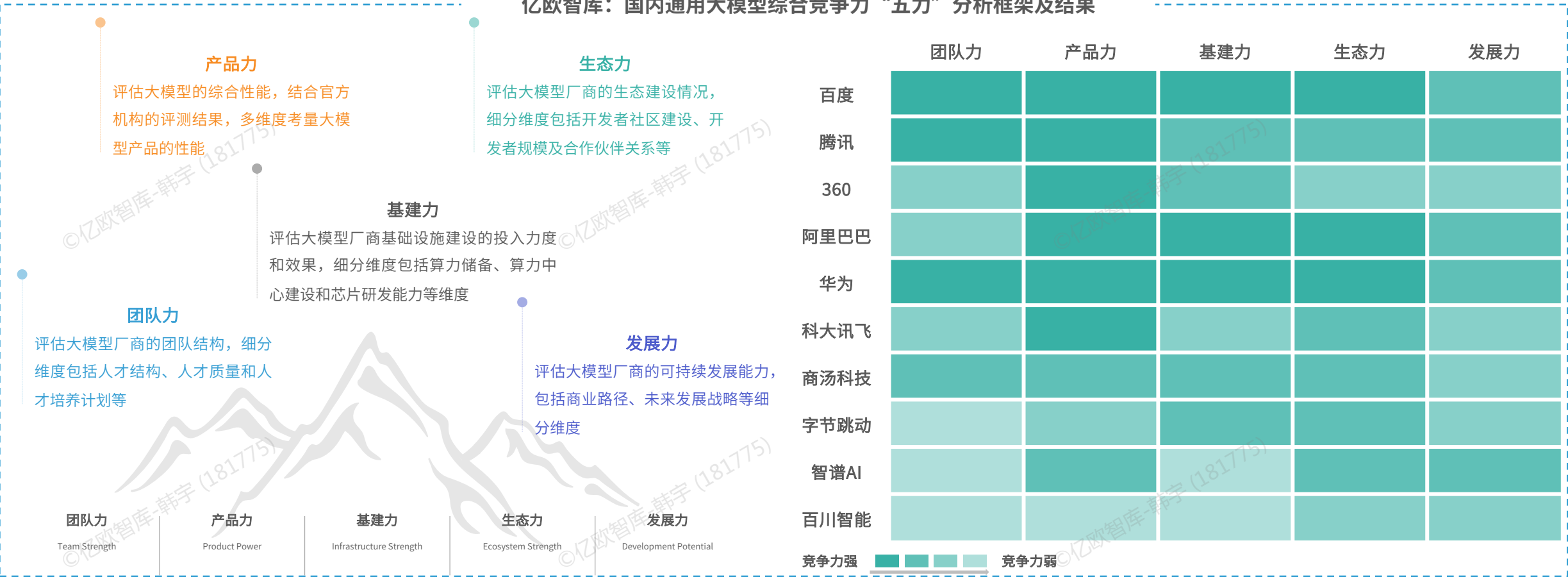
 MOSS	 鸿鹄图文大模型	 星河通用视觉大模型	 欧姆	 智象
 美团大模型	 九天·众擎基座大模型	 APUS天燕大模型AiLMe	 小白	 Yan
 玉言	 快意	 鹏城·脑海	 蓝心	 ChatYuan
	 天河·天元	 华藏	 活字	 语鲸大模型

概念与研发

 中国科学技术大学 UniDoc	 华中科技大学 Monkey	 浙江大学 KnowLM	 清华大学 Chat-UniVi	 南方科技大学 SUS-Chat-34B	 香港中文大学 PointLLM	 北京理工大学 MindLLM (明德)
--	--	--	--	--	--	--

- ◆ 本报告结合国产大模型通过备案名录、官方评测结果和市场调研等内容，选取百度、腾讯、360、阿里巴巴、华为、科大讯飞、商汤科技、字节跳动、智谱AI和百川智能十家发布了通用大模型的厂商为研究对象，从“团队力、产品力、基建力、生态力和发展力”五个维度构建“五力”模型，评估通用大模型厂商综合竞争力。
- ◆ 不同类型的大模型厂商各具优势，百度和华为的五项能力均领先于行业平均水平；腾讯、360、阿里巴巴和科大讯飞在产品力方面表现优异，商汤科技和字节跳动的基础设施投入和建设处于行业中上水平。智谱AI和百川智能作为大模型初创企业的代表，智谱AI持续发力产品创新，生态力和发展力处于行业中上水平；百川智能整体竞争力仍有提升空间。

亿欧智库：国内通用大模型综合竞争力“五力”分析框架及结果

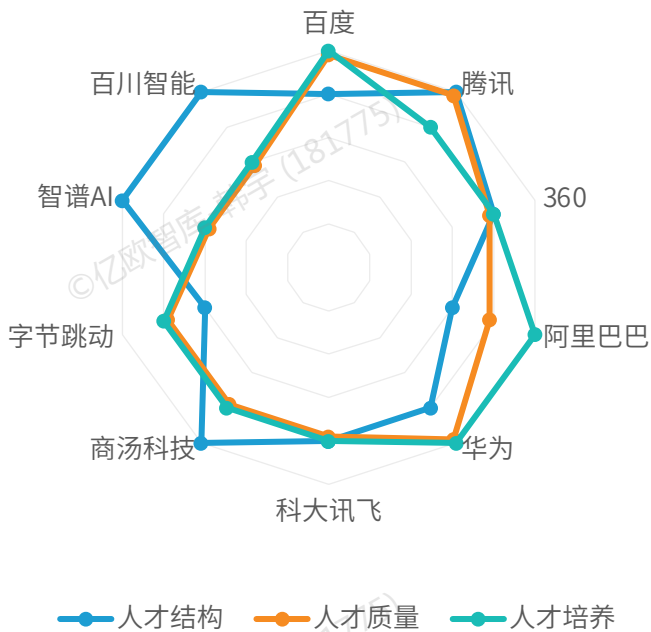


资料来源：公开资料、亿欧智库

◆ 人才结构、人才质量和人才培养方式是衡量大模型厂商团队力的关键指标，整体来看大厂在团队建设方面占据先发优势，百度、腾讯和华为在团队力方面整体实力较强，以高占比的研发人员、高创新力的人才团队和长效的人才培养机制为抓手，多措并举提高人才队伍素质，为大模型产品的技术创新和产品落地提供源源不断的人才支持。

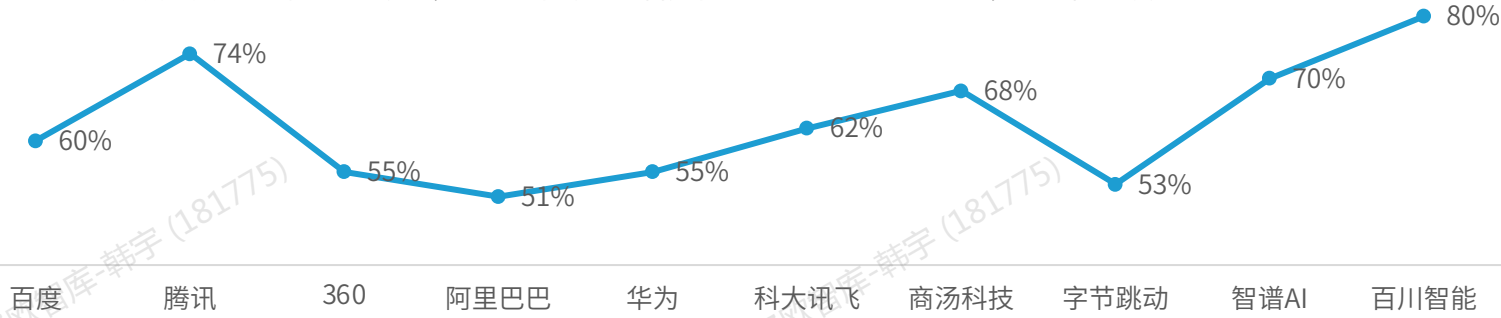
亿欧智库：通用大模型“团队力”评估结果

- 人才结构：细分维度包括厂商研发人员占比、AI人才规模等
- 人才质量：评估人才创新潜力，细分维度包括创新主体人才团队竞争力得分和高端人才储备情况等
- 人才培养：评估厂商在人才培养方面的布局和发展潜力



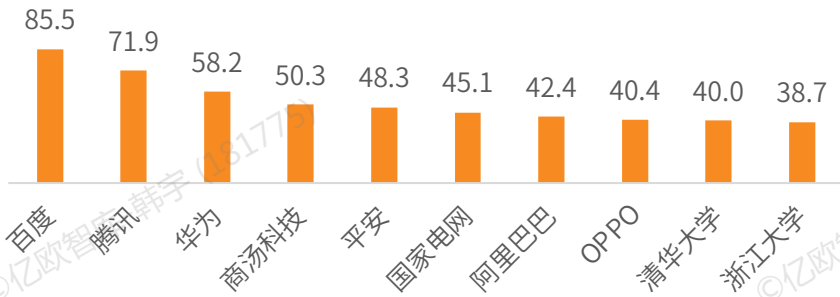
大模型厂商研发人员占比*

- 大模型产业需要高素质的人才支持，研发人员占比是考察人才机构的重要指标之一。
- 大模型厂商研发人员占比均在50%以上，腾讯、智谱AI和百川智能等企业占比高于70%，处于行业领先地位。



创新主体人才团队竞争力

- 人才竞争力是大模型产业创新发展的关键驱动要素，从团队竞争力来看，百度、腾讯、华为和商汤科技的得分在50以上，领先于行业平均水平人才质量较高。



厂商人才培养计划列举

- 百度：联合全国11家职业教育集团，发布“**大国工匠**”人才培养计划，该计划旨在产业数字化转型的时代背景下，培养复合型人工智能技术人才，弥补国内AI人才缺口
- 阿里巴巴：2023年底达摩院（湖畔实验室）博士后科研工作站正式揭牌成立，**2024年目标引入30名应届博士毕业生入站研究**
- 华为：针对**高校师生、开发者、科研人员**量身定制**培养体系**，将不断通过ICT学院和智能基座2.0项目、昇腾AI创新大赛等系列活动为AI生态的技术创新提供动力

资料来源：《中国人工智能产业创新人才竞争力报告（2023年）》、公开资料、亿欧智库
注：数据来源于企业年报/公开报告，字节跳动未公开最新数据，报告中数据为预估结果；另由于篇幅限制评分依据仅作部分展示

◆ 大模型性能的持续优化能够有效提升用户体验和保持厂商业务的长期增长，是大模型厂商构筑竞争壁垒的有效手段。工信部、信通院、新华社和上海人工智能实验室等相继推出大模型性能评估模型，从不同维度对主流大模型产品的性能进行评测，百度、腾讯、360、阿里巴巴、华为和科大讯飞在大部分评估体系下处于领先行业平均水平的地位。

亿欧智库：通用大模型“产品力”评估结果

工信部电子标准院官方“大模型标准符合性评测”

- 介绍：国内首个官方评测结果，该测试由工信部中国电子技术标准化研究院发起，评测围绕多领域多维度模型评测框架与指标体系，从大模型的通用性、智能性、安全性等维度开展，涵盖语言、语音、视觉等多模态领域
- 结果：仅有**百度文心一言**、**腾讯混元大模型**、**360智脑**、**阿里云通义千问**四款国产大模型首批通过测试



上海人工智能实验室“OpenCompass”

- 介绍：上海人工智能实验室发布大模型开源开放评测体系“司南”，全面量化大模型在知识、语言、理解、推理和考试等五大能力维度的表现，本报告数据来源2023 OpenCompass年度榜单，累计评估35个国内外主流模型，GPT-4-Turbo排名第一（61.8/1）
- 结果：智谱AI GLM4（58.9/2）、阿里巴巴Qwen-Max（56.8/3）、百度Erniebot-4.0（54.7/4）、百川智能Baichuan2-Turbo（46.4/12）



中国信通院“可信AI大模型标准符合性验证”

- 介绍：中国信通院结合《生成式人工智能技术及产品评估方法》《大规模预训练模型技术和产品评估方法》标准，开展生成式人工智能标准符合性验证工作，从模型开发、模型能力、模型运营和模型应用等维度展开评估
- 结果：
 - 百度：完成模型开发和模型能力两部分测评，结果表现优异
 - 腾讯：在模型开发和模型能力两个重要领域的综合评价均获得4+级别最高分
 - 360：顺利完成模型开发和模型能力两大模块验证，并获得当前模块4+级最高评分
 - 华为：生成式人工智能技术能力和产品能力标准符合性验证中，获得优秀级（4+级）评分，成为业界首个获得卓越级（5级）评分的大模型产品
 - 科大讯飞：顺利完成模型开发和模型能力两大模块验证，并获得当前模块4+级最高评分
 - 智谱AI：ChatGLM3-130B获大语言模型专项4+级评分



新华社研究院中国企业发展研究中心

- 介绍：研究中心通过构建主流大模型综合指数，从基础能力、智商、情商和工具提效四个维度对主流大模型进行评估，截至2024年2月累计发布3次评估结果

人工智能大模型体验报告1.0

文心一言：1112/3
通义千问：1020/4
讯飞星火：1010/5
商量：967/6
ChatGLM：943/7

评估8家国内外主流大模型，
ChatGPT4以1246分排名第一

人工智能大模型体验报告2.0

讯飞星火：1013/1
文心一言：1010/2
商量：983/3
ChatGLM：983/3
360智脑：951/5
通义千问：935/7

专注国内市场，评估8家国内大模型，
讯飞星火排名第一

人工智能大模型体验报告3.0

讯飞星火：1775/1
商量：1746/2
ChatGLM：1729/3
360智脑：1711/5
字节豆包：1710/6
通义千问：1634/7
腾讯混元：1609/8

专注国内市场，评估10家国内大模型，
讯飞星火排名第一



- ◆ 算力是大模型训练和运行的核心资源，芯片是提供算力的硬件基础，数据是训练大模型的原材料，构建“大规模算力 – 高性能芯片 – 高质量数据”等多维一体的大模型基建设施是保障大模型技术创新和产品商业落地的关键。
- ◆ 大厂相对于初创企业在基建力方面占据比较优势，通过自建算力中心、自研芯片和构建高质量数据集的方式巩固大模型产业发展的底层基础。

亿欧智库：通用大模型“基建力”评估结果

大模型厂商芯片自研情况



大模型厂商算力中心建设情况

企业	算力中心	建设情况
百度	百度智能云济南智算中心	2023年3月揭牌
	百度智能云福州智算中心	2023年4月与福州市达成合作
	百度智能云沈阳智能计算中心	2023年5月开工建设，总体规划算力500P
	百度智能云昆仑芯盐城智算中心	2023年9月正式揭牌上线，算力规模预计达200P
腾讯	百度云算（阳泉）中心	亚洲最大的单体智算中心，算力规模达4EFLOPS
	松江GPU智算中心	2023年11月联合松江落地国内最大规模GPU智算中心，部署了新一代的HCC高性能计算集群
	腾讯长三角（上海）人工智能先进计算中心	2020年6月6日开工建设，预计建成后，先进计算中心机柜数量接近4万架，服务器数量将达到80万台
	腾讯智慧产业长三角（合肥）智算中心	2021年7月9日正式揭牌，采用腾讯第四代T-Block等高端模块化技术
360	360中原智算中心	一期项目已建成投产，本地存储服务器30000台，智能算力1000P；未来将拥有超过5万台服务器，2000P的智能算力
阿里巴巴	阿里云张北超级智算中心	2022年8月30日正式启动，总建设规模为12EFLOPS
	阿里云乌兰察布智算中心	2022年8月30日正式启动，建设规模为3EFLOPS
	阿里云华东智能算力中心	2020年开工建设，2025年达产
商汤科技	商汤科技人工智能计算中心	截至2023年第四季度，商汤人工智能计算中心算力规模高达6800P
	商汤科技（西南）人工智能计算中心	2023年10月24日正式落地重庆，首期建设以国产化算力为主峰值算力达100P，中长期规划建设算力1000P
华为	我国政府主导的智算中心供应商以华为为主，包括北京昇腾人工智能计算中心、天津人工智能计算中心和杭州人工智能计算中心等众多项目	

- ◆ 生态力是指大模型厂商在构建和维护其生态系统方面的能力，包括合作伙伴组建、开发者社区打造和用户群体维护等。
- ◆ 业内主要采取构建开发者社区、建立战略合作伙伴关系和依托自有平台赋能大模型产业等方式发力生态建设，百度、阿里和华为的生态建设能力领先行业平均水平。华为昇腾社区建设日益繁荣，百度依托飞桨平台助力大模型技术落地，阿里建立AI大模型社区吸引众多开发者入驻。

亿欧智库：通用大模型“生态力”评估结果

百度飞桨平台助力大模型落地

- 飞桨深度学习平台为大模型落地提供了技术支持、开发工具/平台支持和社区支持，能够有效支持大模型的开发、高性能训练和部署，降低模型研发成本，优化研发流程。
- 同时文心一言是飞桨模型库的重要组成部分，起到应用示范和推动产业落地的作用。二者互相协同，共创繁荣健康的生态。

飞桨深度学习平台架构

服务平台	EasyDL 零门槛AI开发平台	AI Studio 学习与实训社区	EasyEdge 端计算模型生成平台
工具组件	PaddleHub 预训练模型应用工具	PaddleX 全流程开发工具	PaddleFL 联邦学习
开发套件	PaddleDetection 目标检测	PaddleHelix 螺旋桨生物计算平台	PaddleOCR 文本识别
基础模型库	PaddleNLP 自然语言处理	PaddleCV 视觉模型库	Wenxin Big Model 文心大模型
核心框架	Paddle 飞桨训练框架	Paddle Lite 轻量化推理引擎	PaddleSlim 模型压缩工具

阿里魔搭社区

ModelScope社区成立于2022年6月，是一个模型开源社区及创新平台，由阿里巴巴通义实验室（Institute for Intelligent Computing），联合CCF开源发展委员会，共同作为项目发起方。截止2024年2月，魔搭社区已集聚2800多款优质AI模型及上千数据集，用户数超过390万。

模型库

CharacterGLM-6B	通义千问1.5-7B-Chat
MiniCPM-2B-dpo-bf16	MiniCPM-V
internlm2-chat-1_8b-sft	

数据集

微调扩散模型数据集	商品评论情感预测
分词-中文-新闻领域数据集PKU	Anti-UAV数据集

创空间

通义千问1.5-72B-对话-Demo	名家春联生成
.....	



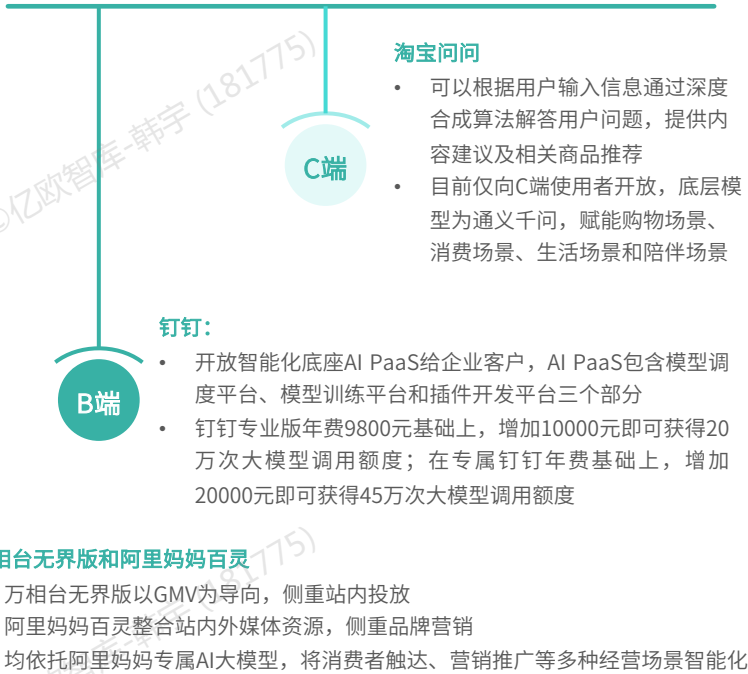
- ◆ 团队力、产品力、基建力和生态力构筑了大模型厂商短中期发展的核心竞争力，考量企业商业路径和战略的发展力是大模型厂商在长期竞争中保持优势的关键要素。目前大模型厂商处于商业化发展初期，部分厂商已形成初步发展路径并进行商业化尝试，但各厂商间的差异相对不明显。
- ◆ 华为构筑三层行业模型走聚焦行业落地的路径；阿里巴巴提出“AI驱动，公共云优先”，B端和C端齐发力；百度上线了付费版文心一言，并且利用大模型重构旗下业务，预计2024年文心一言将带来数十亿元收入；腾讯超过180个内部业务已接入腾讯混元，以微信为入口的Agent为大模型落地带来丰富想象空间；智谱AI已有2000多家合作伙伴，范围涵盖了金融、能源、制造等不同行业，2023年达成亿级签单额。

亿欧智库：通用大模型“发展力”评估结果

华为：构建三层级模型体系，践行AI for Industries

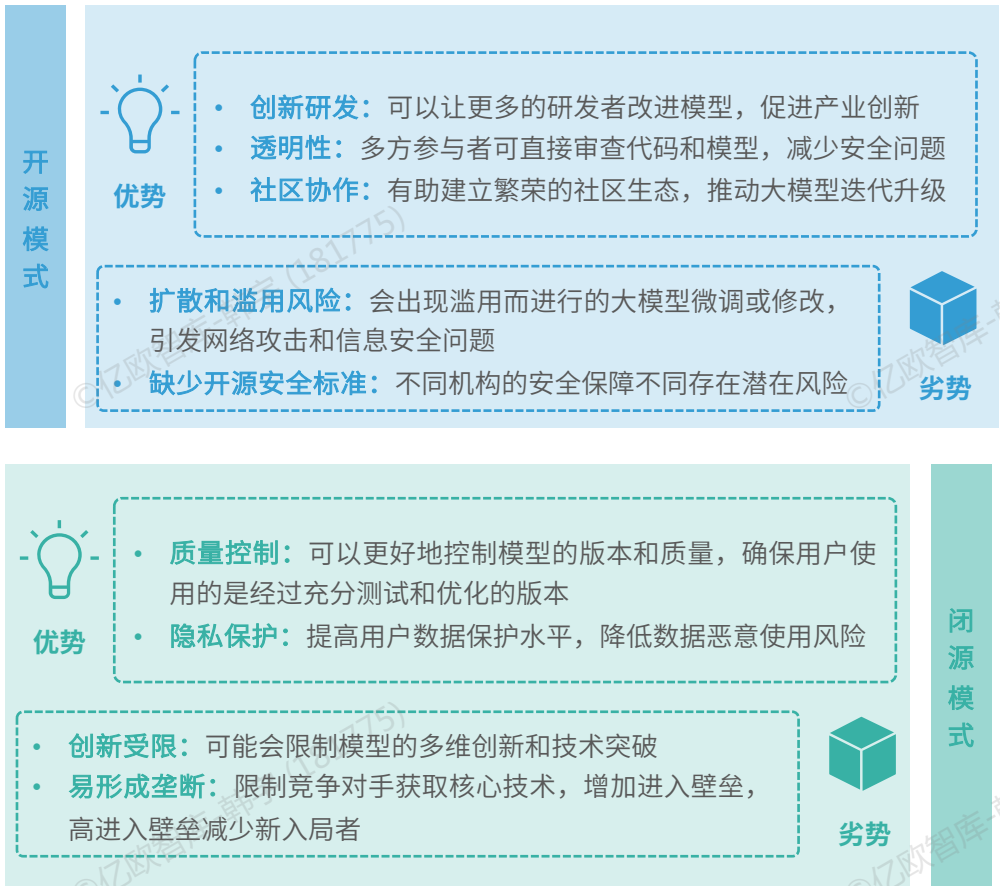


阿里巴巴：B端+C端齐发力，全链条布局AI



- ◆ 基于模型的源代码和训练数据是否公开，可将大模型分为开源和闭源两种形式，随着大模型产业的快速发展，开源和闭源之争日益激烈。开源具有能够有效促进创新和生态繁荣等优点，闭源具有高安全性和强质量保证等优点。
- ◆ 从国内来看，部分通用大模型厂商选择开源路线，但企业在进行决策时需结合自身技术优势和未来发展目标进行选择。比如阿里基于自身算力资源支持和促进生态繁荣的目标选择开源道路，而华为基于“AI for Industries”的发展目标选择闭源路径保护用户数据安全。

亿欧智库：开源和闭源的优劣势分析



亿欧智库：通用大模型开源和闭源典型案例



目录

CONTENTS

01 战鼓齐鸣，“百模争锋”战火正旺

- 1.1 中国AI大模型发展历程分析
- 1.2 “百模大战”必有一战的因素分析
- 1.3 “百模大战”核心竞争力分析
- 1.4 大模型基础层图谱

02 势如破竹，通用模型谁主沉浮

- 2.1 通用大模型图谱
- 2.2 通用大模型竞争力分析
- 2.3 开源和闭源模型发展概况

03 百舸争流，垂类模型落地几何

- 3.1 垂类大模型图谱
- 3.2 垂类大模型应用领域潜力分析
- 3.3 开源和闭源模型发展概况
- 3.4 优秀企业案例

04 终局未定，“百模大战”未来何往

- 4.1 通用大模型趋于寡头竞争格局
- 4.2 垂类大模型呈现碎片化竞争格局
- 4.3 开源与闭源持续维持双线发展
- 4.4 AI Agent构建“人机协同”新范式
- 4.5 多模态生成将在短中期落地

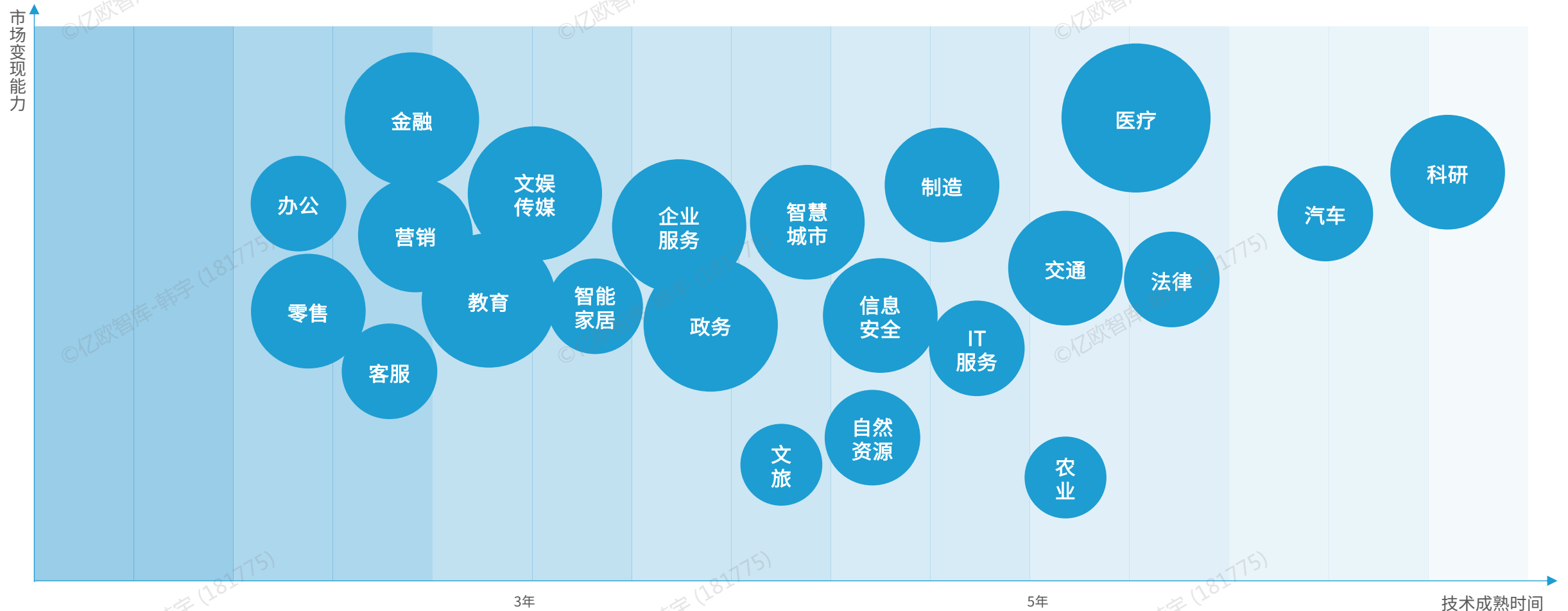
垂类大模型图谱



业务落地 业务试点 概念与研发

注：图谱未详尽

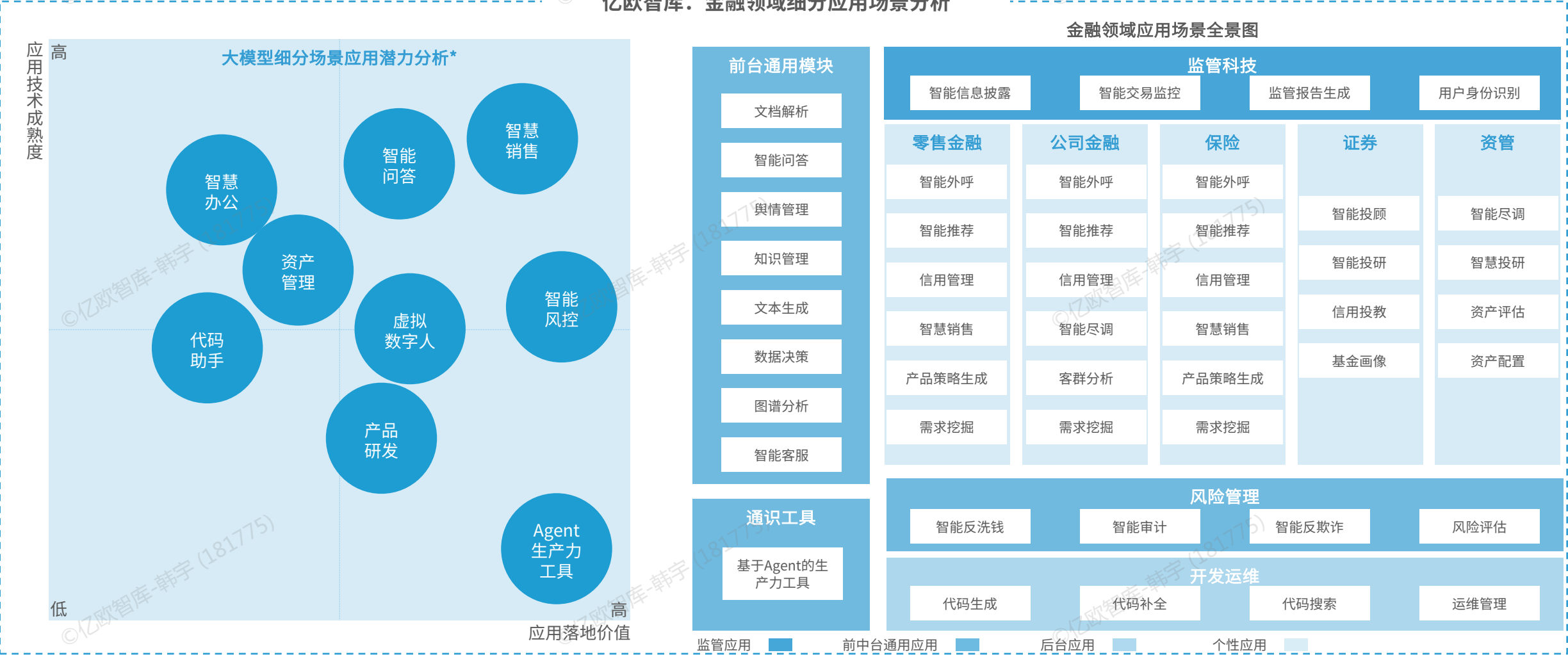
◆ 随着大模型技术的演进，其赋能千行百业的能力不断提升。短期来看，金融、教育、文化传媒、办公和营销等领域能够实现快速场景落地，推动智能化转型。中期来看，政务、制造和智慧城市等领域的市场潜力将不断释放，未来场景想象空间较大。长期来看，大模型将为医疗、汽车和科研等领域带来颠覆性变革，促进高质量发展。



- 横轴表示技术成熟时间，用以衡量大模型技术在细分领域成熟落地所需的时间长短
- 纵轴表示市场变现能力，用以衡量大模型技术未来的市场潜力和场景商业落地价值
- 气泡大小表示玩家规模，用以衡量现阶段布局该垂类赛道的大模型厂商数量

- ◆ 金融业数字化程度相对较高，人工智能的行业渗透率高，也是大模型的优质落地领域，其高质量的数据资源为大模型的训练创造了良好环境。智慧销售、智能问答和智能风控是现阶段的热门应用场景，产品研发和Agent生产力工具是具有发展势能的潜力场景。
- ◆ 销售、风控和运营为大模型在金融领域赋能的重点环节，通过引入前中台通用应用、后台应用、个性应用和监管应用实现流程管理优化、运营成本降低和风控能力提升的目标。

亿欧智库：金融领域细分应用场景分析

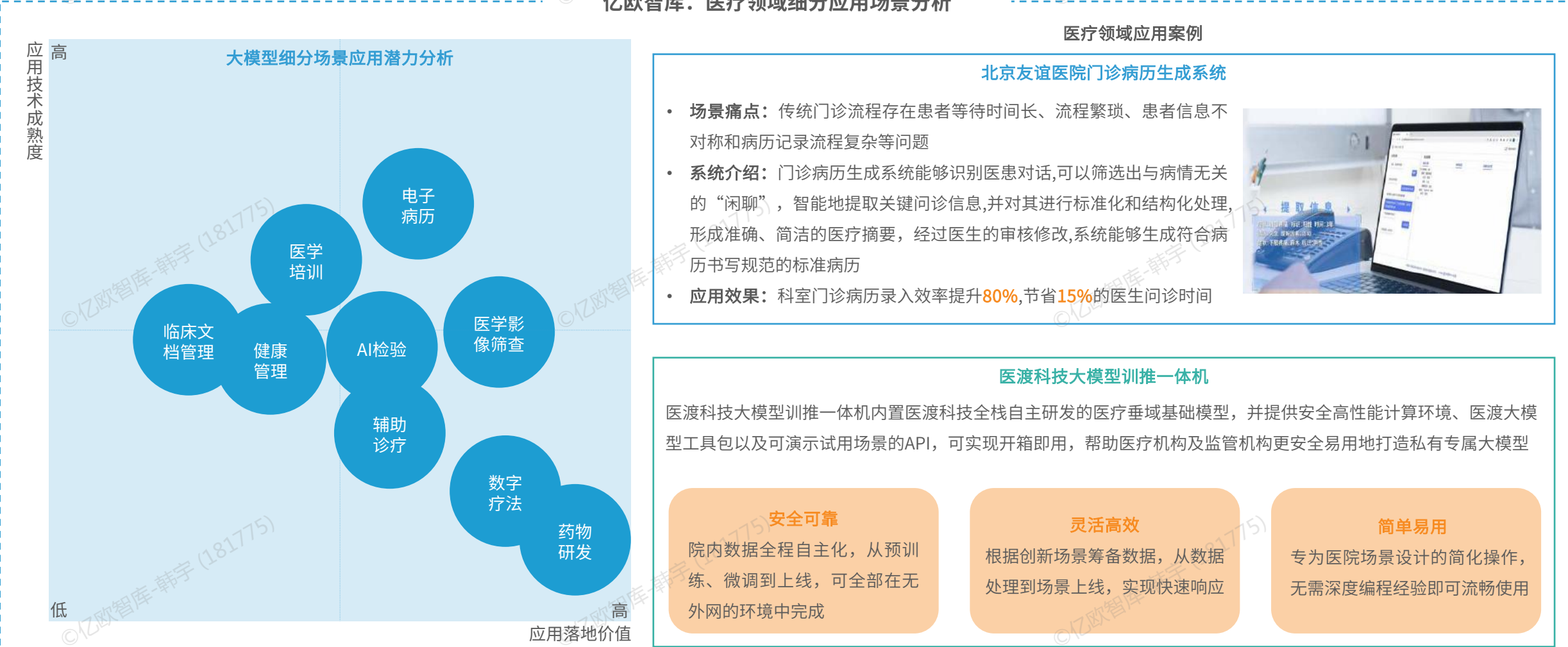


资料来源：中国银保传媒股份有限公司《2023金融业大模型应用报告》、公开资料、亿欧智库

*注：应用落地价值：用以衡量细分场景应用为下游客户带来的经济、成本和战略价值；应用技术成熟度：用以衡量现阶段大模型技术在对应场景中的技术成熟度；下同

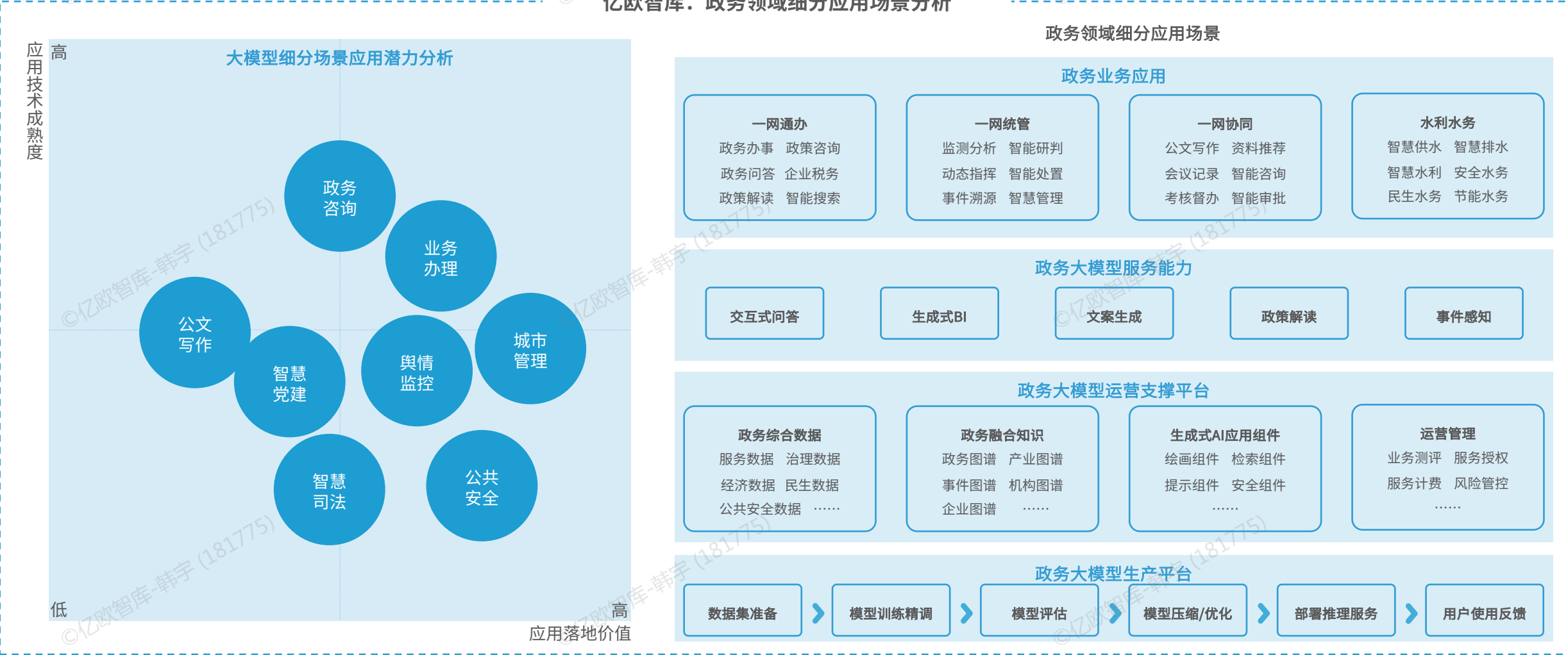
- ◆ 大模型可赋能医疗行业“医、教、研、管”等场景中的各个环节，以提高诊疗效率、诊疗精确度和管理效率等目标为手段，全方位提升诊疗水平。
- ◆ 基于医疗行业应用场景复杂和数据专业度高等原因，大模型的技术成熟度相对较低，目前仅有电子病历等技术要求相对低的场景进入商业化阶段。未来随着智能药物研发、数字疗法和辅助诊疗等技术的成熟将带来跨越式价值提升，有效赋能医疗行业的智能化转型，提升医疗可及性。

亿欧智库：医疗领域细分应用场景分析



- ◆ 随着数字政府建设的持续推进，人工智能等新兴技术在政务领域的渗透率持续提高。目前大模型技术开始赋能业务办理和政务咨询等生成类场景。未来将在城市管理、舆情管理和公共安全等决策类场景中发挥作用。
- ◆ 政务大模型的未来发展需要以大模型生产平台为基础，以运营平台为支撑，融合多元服务能力，持续发掘细分应用场景，赋能政务领域高质量发展。

亿欧智库：政务领域细分应用场景分析



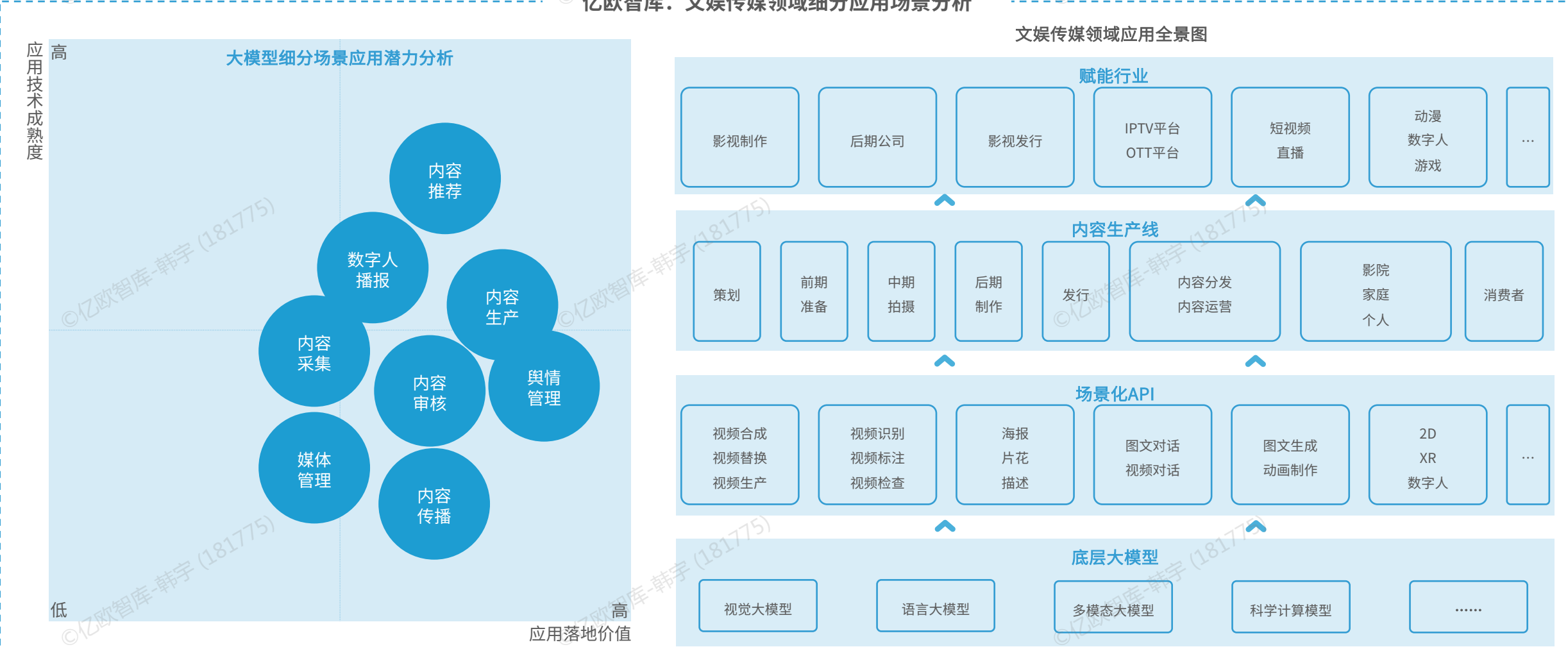
- ◆ 教育大模型C端场景落地速度快于B端，C端场景主要围绕教师和学生两类角色展开，覆盖备课、教研和自主学习等环节。同时大模型的快速发展持续赋能智能硬件产品，以学习机为代表的AI终端硬件迎来新的发展高峰。
- ◆ 目前智能作业和辅助练习场景的应用相对广泛，未来随着大模型技术的成熟，将有效赋能学术科研、智慧管理、学业规划和虚拟教师等细分场景，全方位提升教学质量。

亿欧智库：教育领域细分应用场景分析



- ◆ 影视传媒产业是大模型的优质落地领域，应用广泛覆盖前期内容收集、中期内容制作和后期内容宣传的全流程。目前内容推荐、数字人播报和内容生产等场景开始落地尝试，内容传播、内容审核和舆情管理等场景依托高场景落地价值具有较大的发展潜力。
- ◆ 大模型技术发展迅速，已从单模态生成向多模态生成转变，未来将持续变革影视传媒行业的生产方式，塑造全新的内容生产范式，实现创作效率提升、创作空间拓展和作品质量提升的目标。

亿欧智库：文娱传媒领域细分应用场景分析



资料来源：华为、公开资料、亿欧智库

- ◆ 制造业可划分为研发、设计、仿真、生产、测试、运维、售后等环节，通过利用大模型的多类型生成能力，可有效赋能不同环节的智能化转型。
- ◆ 大模型技术在制造业领域的应用处于初期，目前在质量检测和教学实训等领域开始初步探索。随着技术的成熟未来可依托前瞻设计、高效化仿真、精细化检测、智能调控、科学运维和定制化售后的方式推动大模型技术在制造业领域的深化应用，赋能仿真场景生成和安全监管等多场景落地。

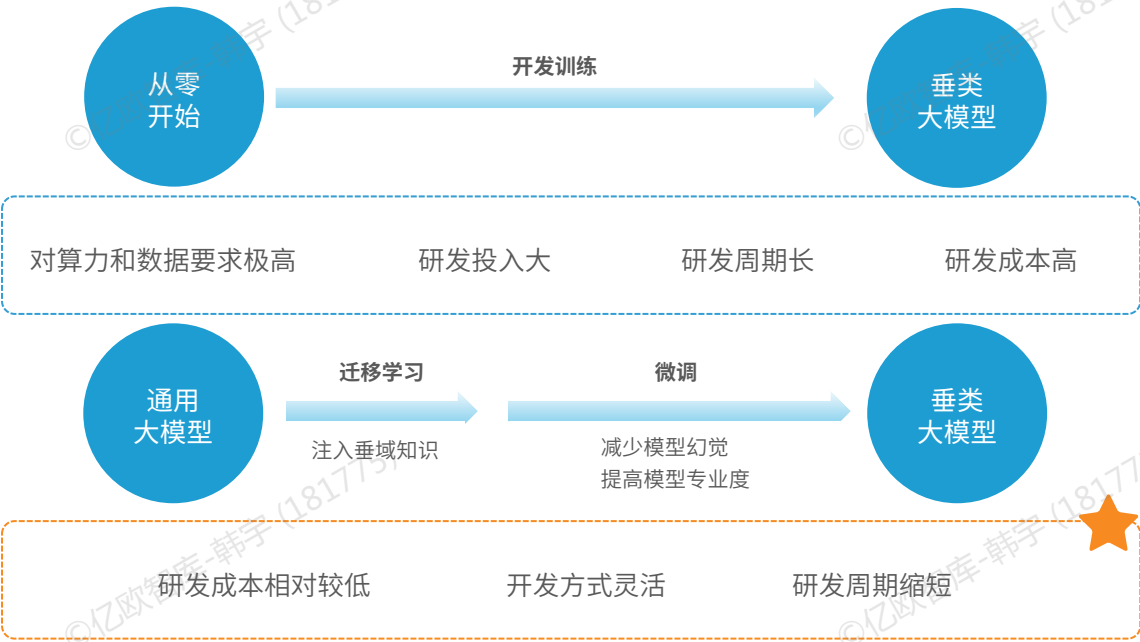
亿欧智库：制造业领域细分应用场景分析



- ◆ 垂类大模型的开发存在从零开发和以通用大模型为底座进行开发两种形式，依托在研发投入、研发效率和开发灵活性等方面的优势，大部分垂类大模型厂商选择基于通用大模型底座进行二次开发的方式。
- ◆ 垂类大模型的开发底座可以选择开源和闭源模型，垂类大模型厂商会从使用成本、输出准确性、技术能力和客户响应等角度出发，选择适合企业发展的模型底座。

亿欧智库：垂类大模型训练方式

- 垂类大模型有两种开发方式，一种是从零开发大模型，另一种是基于已有大模型来开发针对某一垂直领域或垂直场景的垂类模型，这类模型通常基于通用大模型底座，用垂类数据进行训练，进行模型微调后得到细分垂类领域大模型。
- 基于通用大模型底座微调的形式在研发投入、研发周期和灵活性等方面具有优势，垂类大模型厂商大部分选择以该方式切入赛道。



亿欧智库：垂类大模型厂商选择开源/闭源底座模型决策因素

- 采取基于通用大模型底座微调形式开发垂类大模型，模型底座可以选择开源和闭源通用大模型，垂类大模型厂商一般会基于使用成本、输出准确性、技术能力和客户响应等因素进行决策，选择符合自身发展模式的路径。

使用成本

- 闭源大模型有清晰的服务价格，比如按照使用token计费或按照私有化解决方案进行报价
- 开源大模型二次开发成本和企业自身情况相关，包括企业的算力规模、数据质量、开发能力等，每个企业使用成本差异较大

输出质量

- 闭源模型适合场景容错率低的场景
- 目前闭源模型的输出准确性高于开源模型，而开源模型的透明性利于企业快速创造产品和服务，垂类大模型厂商需要根据细分应用场景的要求进行模型选择

技术实力

- 闭源大模型具有完整的工具链和工具平台，企业可以直接调用其接口，同时能够获得大模型企业及时高效的技术支持
- 开源大模型对企业的二次技术开发能力要求高，企业往往需要进行二次开发，对机器学习、自然语言处理、数据科学等方面的专业能力要求高

响应时间

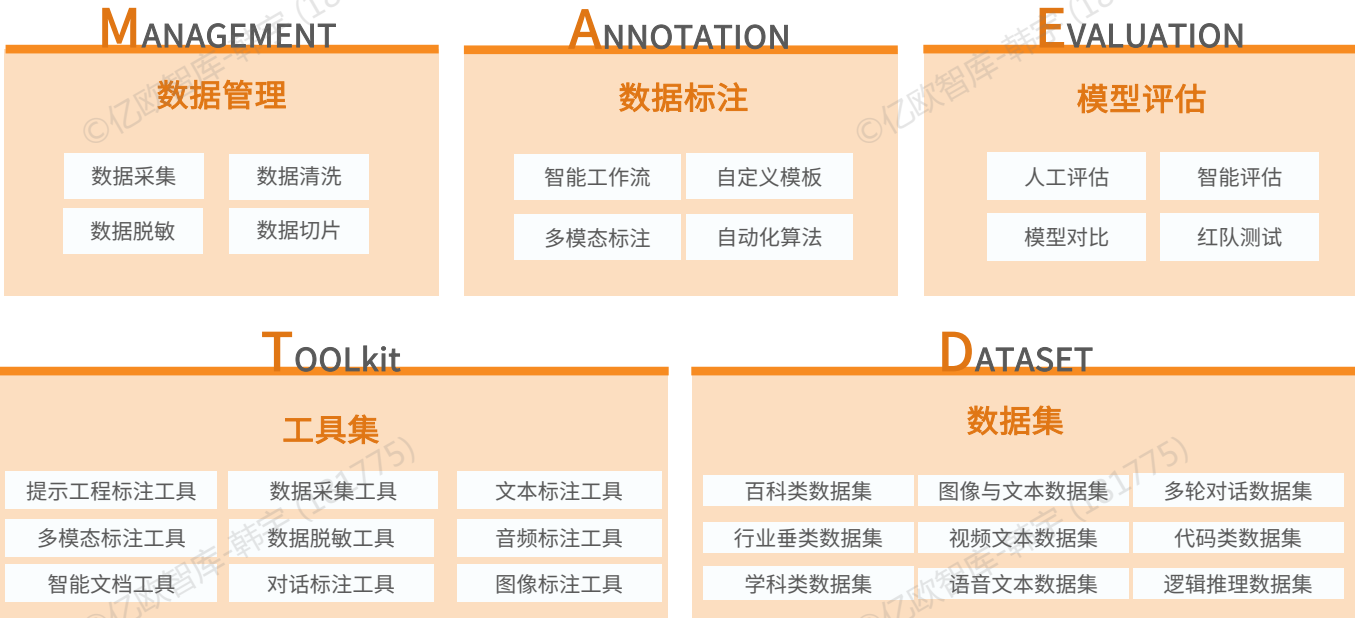
- 开源模型能够减少客户响应时间
- 使用闭源模型时，模型提供者倾向于选择批处理请求，可能导致高延迟；使用开源模型时，垂类厂商可以通过优化模型来降低计算延迟，还可以改变批处理请求的设置以降低网络延迟，从而减小客户响应时间

◆ 澳鹏Appen是全球图像、文本、音频、视频等AI训练数据服务提供商，拥有业内先进的人工智能辅助数据标注平台及自研的大模型开发平台，助力全球7,500+个AI项目的研发及商业化。

公司及大模型智能开发平台介绍

- 数据是决定机器学习模型性能的三大要素之一。随着各类大模型的智能涌现，数据，尤其是高质量的行业数据，正在成为决定大模型高速发展的关键因素。
- 澳鹏智能大模型开发平台提供了完整的产品堆栈，包括数据集管理、数据标注、计算资源调度、模型评估和模型微调等功能，帮助企业轻松采用大模型技术。
- 澳鹏提供超过290种语言和方言的文本语音数据库，并创建了一系列大模型的专用数据集，例如百科类人工泛化文本问答数据集、知识类百科文本语料对数据库、58亿图文对数据库，以及法律问答和医疗问答等。

澳鹏大模型智能开发平台



三大核心技术

自研预标注大模型



自研交互式算法



算法赋能文档智能



解决方案

大语言模型训练数据解决方案

澳鹏基于其大型模型开发平台工具集，为一通用大模型厂商提供了多种训练数据解决方案，包括预训练数据、模型微调和提示词工程等。在此过程中，澳鹏完成了逻辑推理、学科数据、多轮对话、指令标注等多个模块的训练数据生产工作，累计处理数亿量级的语言模型训练数据，其准确率达到96%以上。客户使用澳鹏生产的数据进行模型训练优化后，模型性能指标达到了行业顶尖水平。

多模态大模型数据解决方案

澳鹏基于其大模型平台的多模态数据处理能力，为某AI科技厂商提供图像文本描述数据服务。澳鹏为客户提供了超过50亿对高质量图文对，为客户的图文大模型提供了充足的数据支持。同时，澳鹏还为数个多模态大模型项目提供高质量数据服务，涵盖文本、图像、视频、音频等多种不同模态的数据类型。

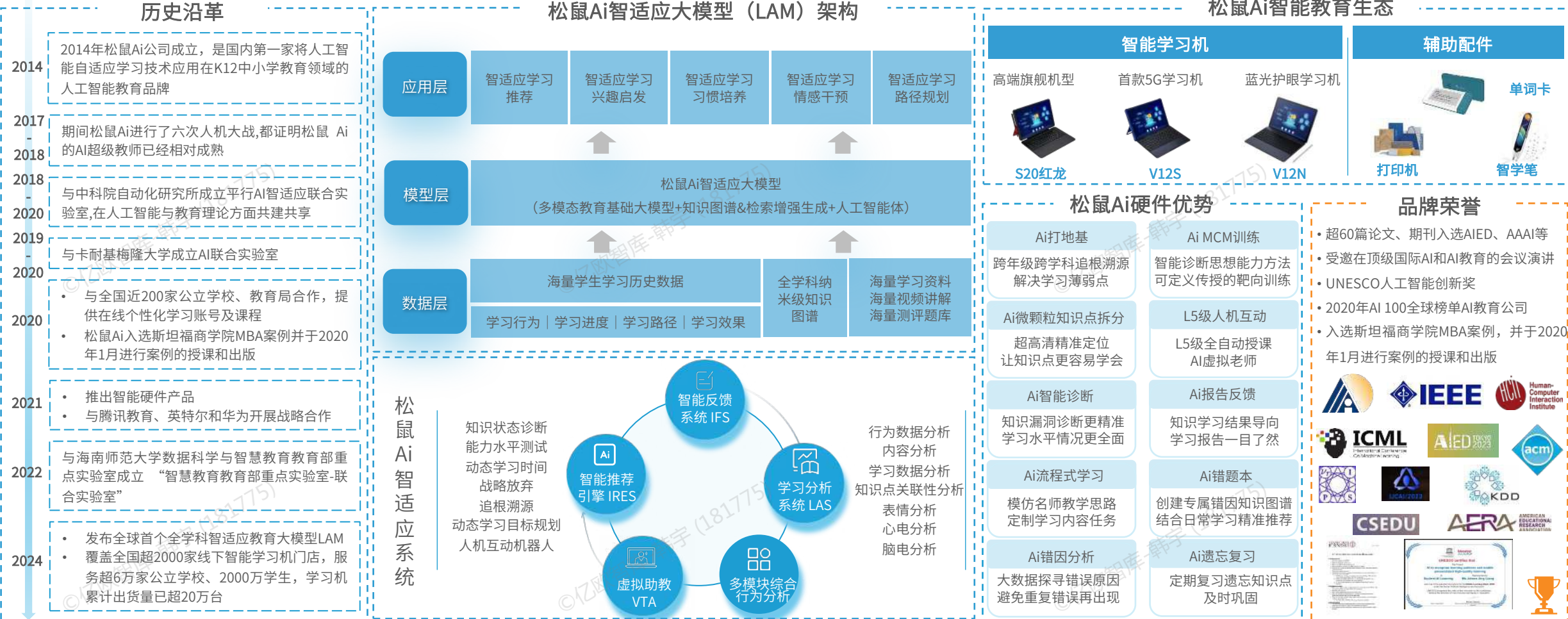
垂类大模型数据解决方案

澳鹏为某科技公司提供大规模代码类大模型训练数据解决方案。利用澳鹏的定制模板引擎和多模态数据编辑器，聚集了行业专家团队，包括各种开发语言的研发工程师、数据分析师和算法工程师，提供高质量的代码训练数据。覆盖了代码编写、逻辑解析、代码调试和单元测试等各个阶段的数据需求，累计提供了50万组以上的高质量代码数据，为代码大模型的应用提供了可靠数据支持。

荣誉



- ◆ 松鼠Ai成立于2014年，是国内较早将人工智能自适应学习技术应用在教育领域的科技创新型独角兽企业，其在人工智能领域的投入已近10年。
- ◆ 松鼠Ai凭借自主研发的全球首个全学科智适应教育大模型（LAM），在智能硬件领域取得了瞩目突破。其AI打地基、智能诊断、知识点精细拆分、流程化学习、MCM学习思想、能力值训练、错因深度分析、遗忘复习机制、错题本整理、报告反馈以及L5级自动驾驶的人机互动教学功能，均彰显其系统的自主进化能力，使得学习内容与教学得以广泛拓展。此外，松鼠Ai还构建了包括学习机、打印机、智学笔、智能单词卡等在内的立体化产品矩阵，充分发挥AI技术的优势，为孩子的成长之路量身打造专属智能生态。



- ◆ 因赛集团是中国第一家以品牌管理为核心主营业务的上市企业，致力于构建以数智化技术、营销服务协同发展的“智慧x智能”营销服务体系
- ◆ 公司大模型产品InsightGPT定位为AIGC驱动的营销全链路应用解决方案，将AIGC技术与营销策略洞察、内容创意生产、传播投放及效果转化等环节深度融合，已具备文生文案、文生图像、视频智剪、图生视频等方面的功能，助推营销行业生产力质效提升

InsightGPT重要事件

2023 10月末 内测版发布 文生图像产品

2024 1月初 正式上线 视频智剪产品

2024 2月初 正式上线 龙年祈福产品

2024 3月初 正式上线 图生视频产品

2024 4月初 预计上线 文生视频产品

InsightGPT赋能全链条营销

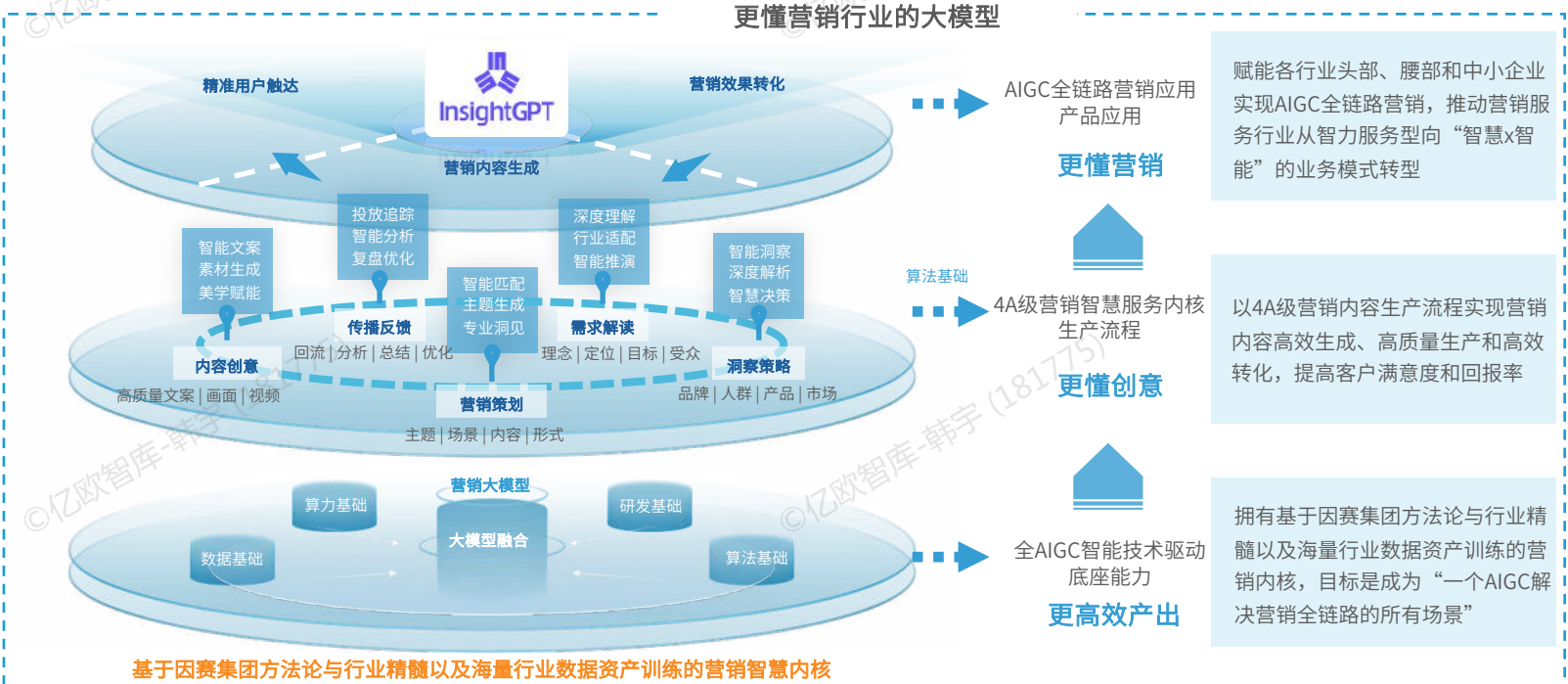
智能化背景 自2023年初起，基于对品牌营销的深入洞察并融合智能化的AIGC技术，自主研发能够生产高品质营销内容的“品牌营销AIGC大模型”和“AIGC营销应用大引擎”，并推出了InsightGPT

产品概述 InsightGPT融合了因赛集团20多年的品牌营销方法论，可以提供全链路营销智慧化服务，以“更具营销洞察力、更具创意沟通力、更具美学表现力”的AIGC内容，为品牌全链路营销赋能

目标与赋能 因赛集团致力于开发一款全AIGC驱动的高品质的营销内容智能创意生成平台，实现数智化驱动的品效合一经营链路闭环

InsightGPT进化方向

品	效	销
品牌价值	效果传播	销售转化
品牌洞察 多维度应用洞察	视频营销 视频生成实现	电商营销 直播场景扩展
整合营销 综合性全案输出	社媒营销 跨平台内容延伸	DTC板块 DTC内容转化
创意板块 高质量创意制作	媒介投放 策略与方案加强	



品

品牌价值

品牌洞察
多维度应用洞察

整合营销
综合性全案输出

创意板块
高质量创意制作

效

效果传播

视频营销
视频生成实现

社媒营销
跨平台内容延伸

媒介投放
策略与方案加强

销

销售转化

电商营销
直播场景扩展

DTC板块
DTC内容转化

图生视频

优秀营销案例

视频智剪 产品/核心卖点/品牌/人群/市场/热点洞察

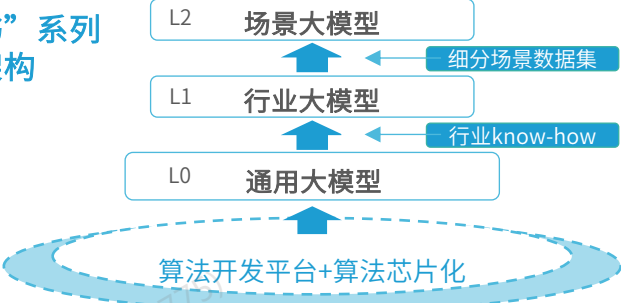
营销创意

海报/裂变营销H5/电商图

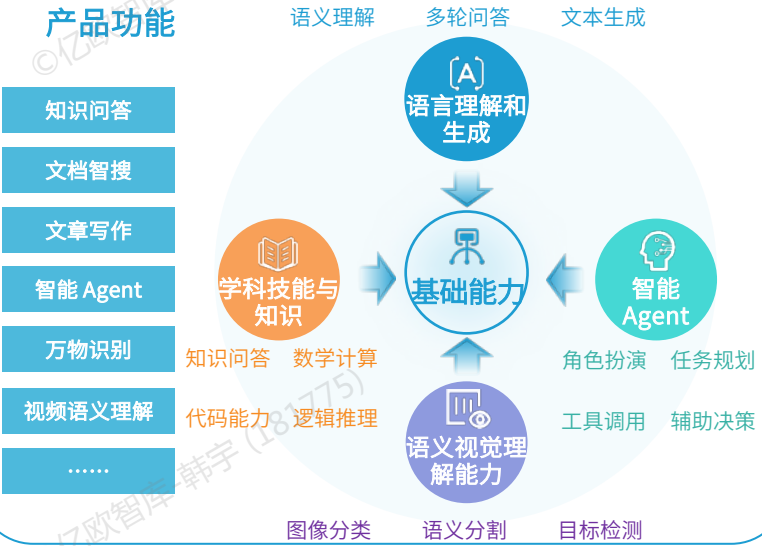
◆ 云天励飞自研大模型“云天天书”包含语言大模型、多模态大模型等不同系列，在C-Eval、CMMLU等权威测试中多次获得第一，具备业界领先的图文数据理解和问答等基础能力，针对行业场景可快速构建出高效、智能的解决方案及系统应用。

“云天天书”系列模型

“云天天书”系列模型架构



产品功能



优势特点

自主可控：

自研大模型训练框架，从训练到推理适配国产算力

创新高效大模型训练技术：

在可伸缩训练、断点续训等业界难点问题上屡次取得创新突破

高性能低成本推理：

首创SPACE推理引擎结合低比特量化，算得快、算得多、算得精，大幅降低应用成本

超长上下文窗口：

基于缓存动态优化技术，可实现300k超长上下文窗口，相当于45万汉字

边缘侧场景理解和开集目标检测：支持边缘侧高效在线训练，秒级微调响应并快速准确扩展到各种应用场景

完备的大模型安全体系：

多级敏感标签、海量敏感词库不断更新、输入输出多重拦截，通过中央网信办备案

“云天天书”产品形态

一体机



- 推理一体机：支持千亿级模型推理应用，支持云天自研芯片
- 训推一体机：支持千亿级模型微调及推理应用，适配华为昇腾生态

系统产品



- 支持模型私有化集群部署，提供训练及推理授权服务
- 支持天书平台丰富的管理、支撑、应用及API等各类服务

细分应用场景

政务公文智写	政务民意速办	警务视频语义检索	企业员工培训	交通城市道路CT巡检
基于公文范式及输入关键信息自动生成提纲、公文进行修改润色	基于政务知识库，通过意图识别、主动反问来实现对民众精准知识问答	基于多模态大模型，可通过自然语言输入对目标场景进行快速识别和检索	通过语言大模型实现自动出题、智能陪练以及科学评估，提升员工培训效率	基于多模态大模型，大幅提升道路病害检测的泛化性、准确率，为科学养护提供决策依据

目录

CONTENTS

01 战鼓齐鸣，“百模争锋”战火正旺

- 1.1 中国AI大模型发展历程分析
- 1.2 “百模大战”必有一战的因素分析
- 1.3 “百模大战”核心竞争力分析
- 1.4 大模型基础层图谱

02 势如破竹，通用模型谁主沉浮

- 2.1 通用大模型图谱
- 2.2 通用大模型竞争力分析
- 2.3 开源和闭源模型发展概况

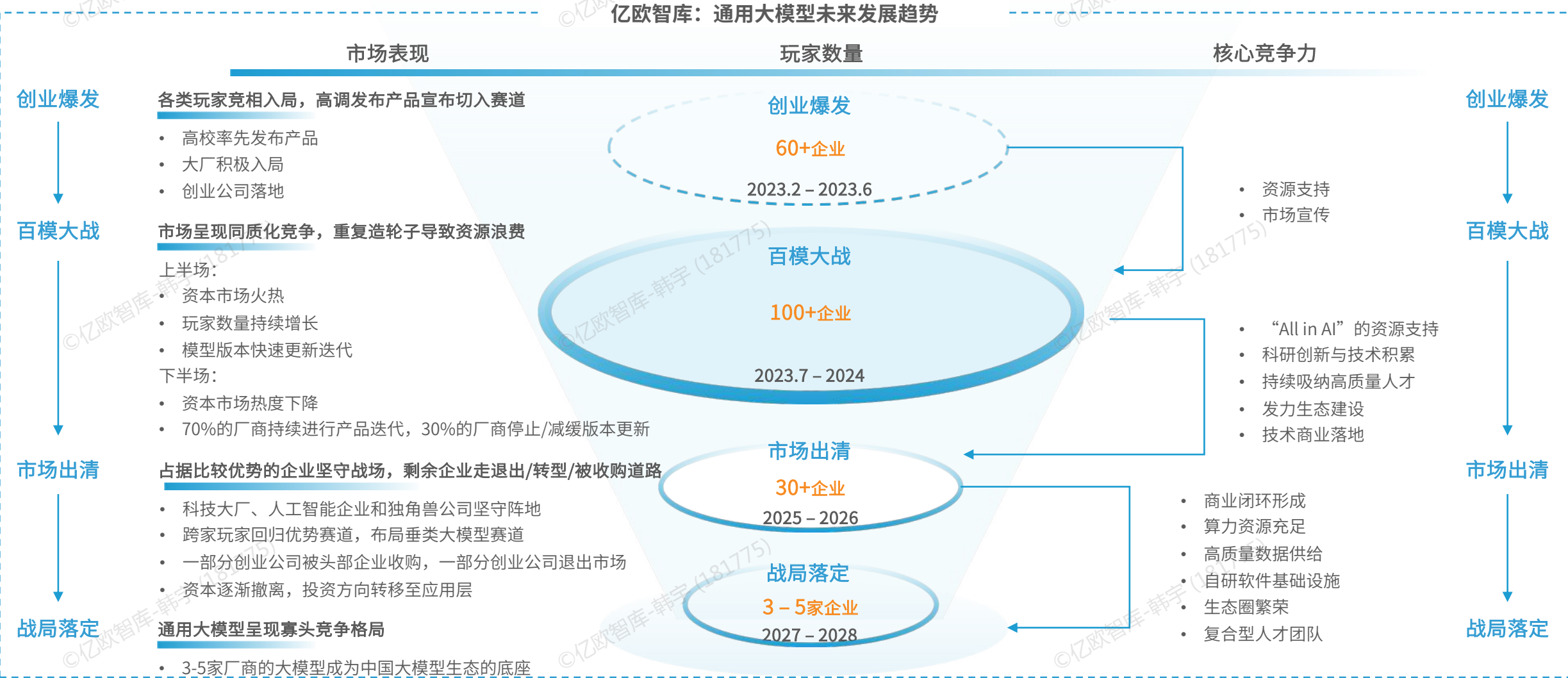
03 百舸争流，垂类模型落地几何

- 3.1 垂类大模型图谱
- 3.2 垂类大模型应用领域潜力分析
- 3.3 开源和闭源模型发展概况
- 3.4 优秀企业案例

04 终局未定，“百模大战”未来何往

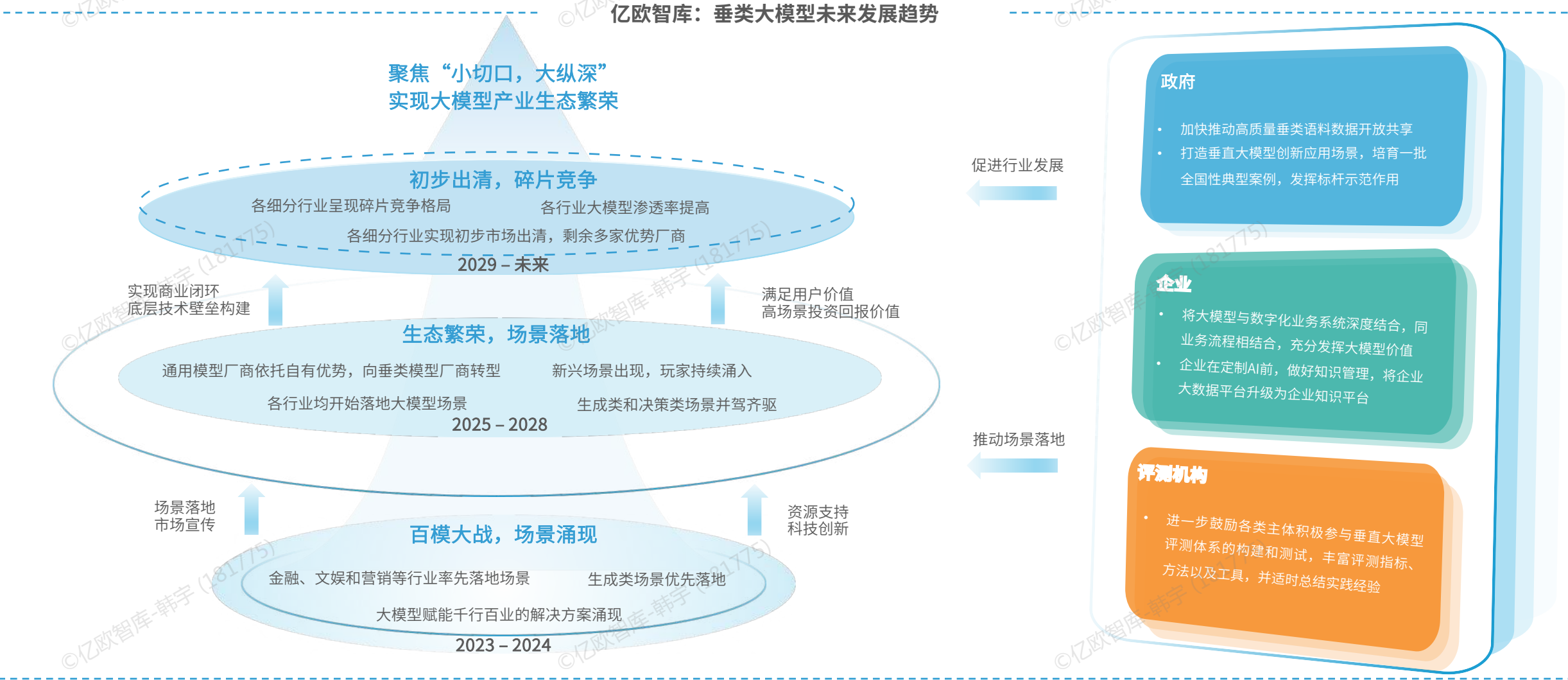
- 4.1 通用大模型趋于寡头竞争格局
- 4.2 垂类大模型呈现碎片化竞争格局
- 4.3 开源与闭源持续维持双线发展
- 4.4 AI Agent构建“人机协同”新范式
- 4.5 多模态生成将在短中期落地

◆ 通用大模型历经创业爆发期，目前正处于“百模大战”深水区，市场同质化竞争严重，资源、技术、人才和商业落地能力成为厂商迈向下一阶段的核心竞争力。预计2025 - 2026年市场逐渐出清，优势企业坚守战场，剩余企业陆续退出竞争。2027 - 2028年预计通用大模型市场将呈现寡头竞争格局，3 - 5家厂商成为中国大模型生态的底层基座。



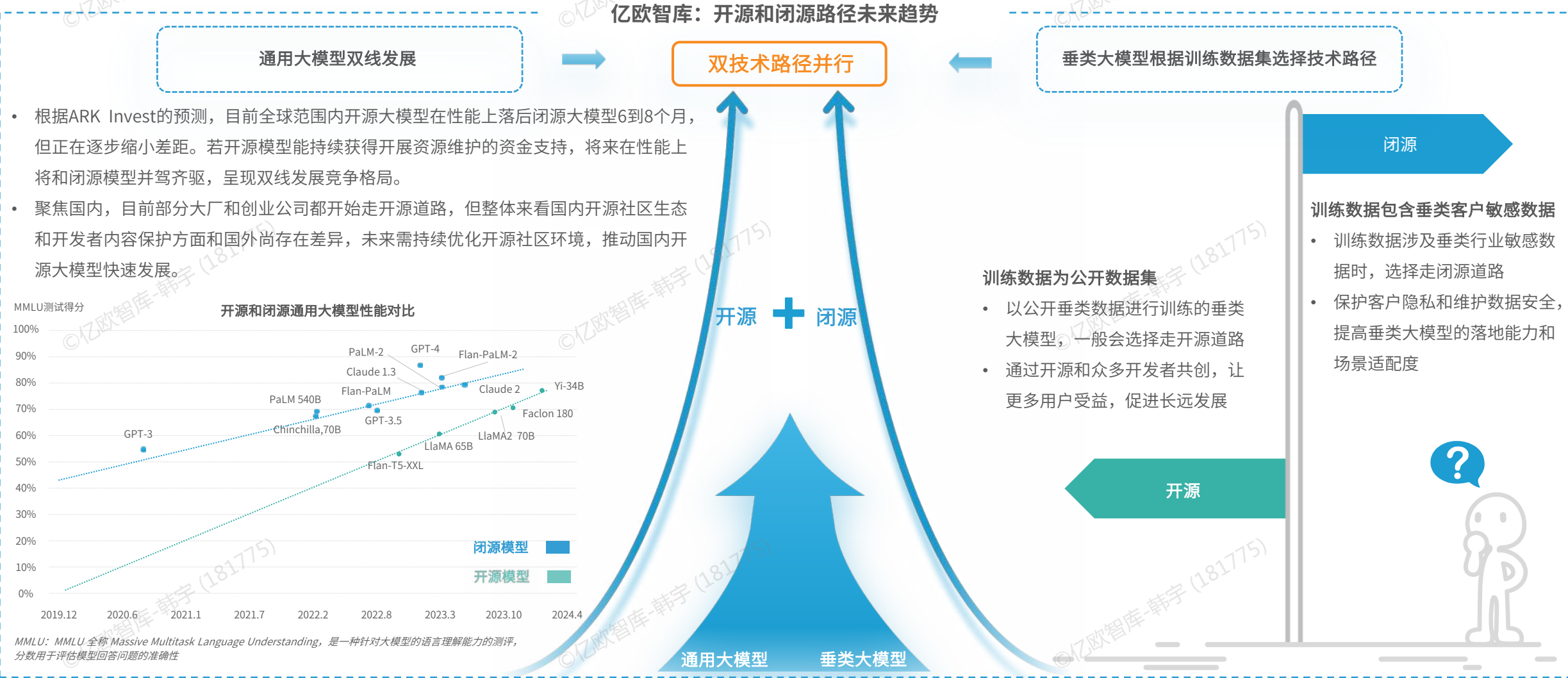
垂类大模型呈现碎片化竞争格局

◆ 垂类赛道将成为大模型产业激烈竞争的领域，预计2023 – 2024年在百模大战背景下，部分行业场景率先落地；2025 – 2028年通用大模型格局逐渐趋于稳定，垂类赛道会涌入一批通用领域转型的玩家，生态持续繁荣，各行业均开始落地大模型场景；2029年以后初步市场出清，各细分行业呈现碎片化竞争格局，应用层获取的市场价值日趋增长。



资料来源：《关于支持垂直大模型开发与应用的建议》、专家访谈、公开资料、亿欧智库

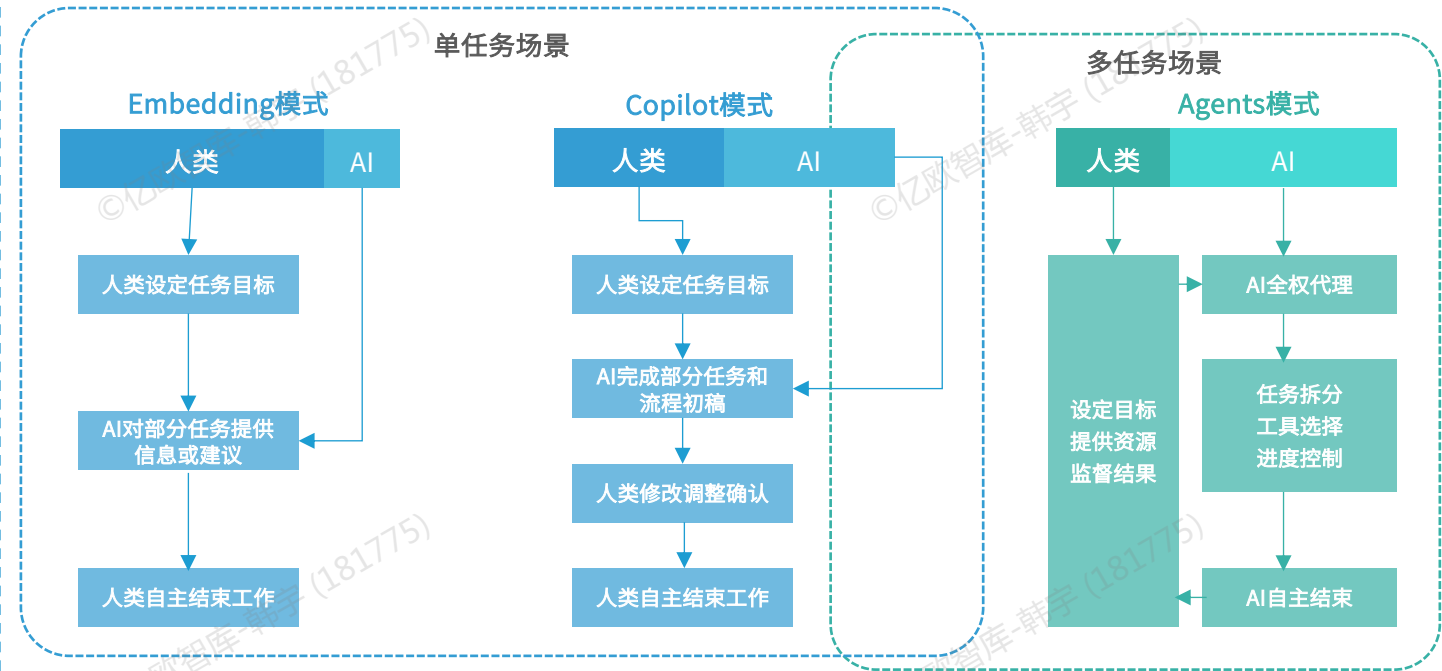
◆ 在通用大模型和垂类大模型领域，开源和闭源都将持续维持双线发展路径。在通用大模型领域，开源模型和闭源模型的性能差异在逐步缩小，在开源模型得到持续资源支持的情况下，其性能将有望接近闭源大模型。在垂类大模型领域，厂商将根据模型训练数据的性质选择走开源或闭源道路，维护企业竞争优势的同时促进产业生态繁荣。



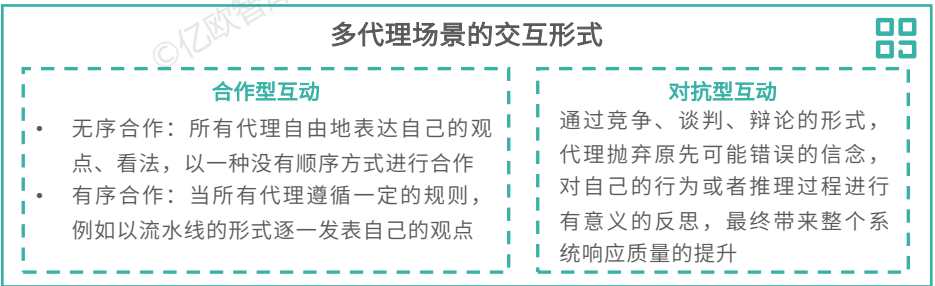
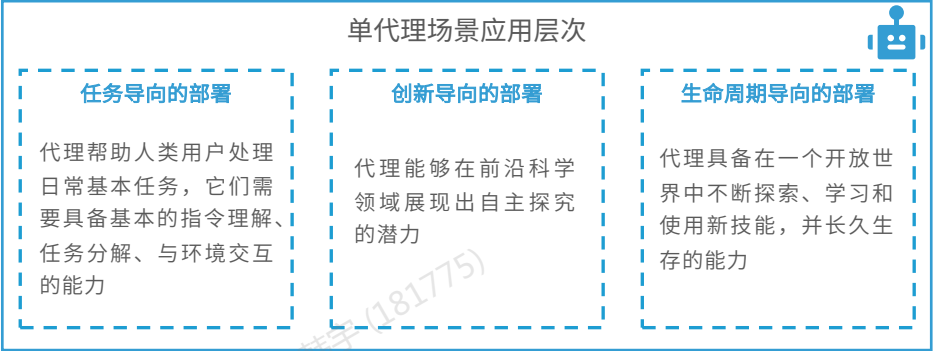
- ◆ AI Agent（人工智能助手）是在人工智能领域具有自主决策能力、环境感知能力和反应能力的智能体。AI Agent可分为单任务场景代理和多任务场景代理，单一场景的Agent聚焦于垂类简单问题的处理，多场景Agent以“高级秘书”的身份帮助用户调动多种资源解决复杂场景问题。
- ◆ 目前AI Agent处于单代理场景落地阶段，现有产品大部分从已有场景出发进行赋能。通用化和智能化的多代理和多任务场景AI Agent为长期发展趋势，一方面能帮助用户解决复杂问题，另一方面和终端硬件的结合能够实现更深层次的人机交互。

亿欧智库：AI Agnet技术架构及未来趋势

AI Agent分为单一任务场景应用和多任务场景应用，单任务场景通常聚焦于某个相对垂类的专业领域，用于解决相对垂类且简单的问题，初级Agent(Copilot)的使用即可实现。多任务场景任务复杂度提升，大模型需要具备拆解、解决任务问题的能力。



人与AI协同三种方式



线下场景中AI Agent的想象空间

机器人Agent是代理在线下场景中的落地尝试，它能够在一定程度上模拟人类智能，执行任务、做出决策，并**与用户或其他软件系统交互**。一方面依托代理的决策链路承担个人助理的线下工作场景（比如家务），另一方面长期随着大模型学习与自适应能力的提升，机器人的行为及反馈可以给予其“大模型”大脑增强交互和任务处理能力。

- ◆ 随着大模型技术的成熟，模型生成能力将从单模态生成向多模态生成演变。具备“3D一致性、物体持久性、模拟数字世界”等核心能力的Sora大模型的发布引发业内对于文生视频领域的广泛关注。由于初期国内对技术趋势预测过于保守叠加缺乏承担试错风险的厂商，国内目前虽已有产品落地，但和Open AI仍存在较大差距，需要持续投入追赶差距。
- ◆ 文本、代码和图像生成领域逐渐成熟，视频生成处于初步尝试阶段，未来文生3D有望成为新的落地热点。

亿欧智库：Sora引爆文生视频领域，国内尚存在较大差距



国内文生AI视频产品和Sora在视频时长、分辨率、帧率、文字理解和物理世界还原等方面都存在显著差异，国内厂商和Sora的差距在1-2年

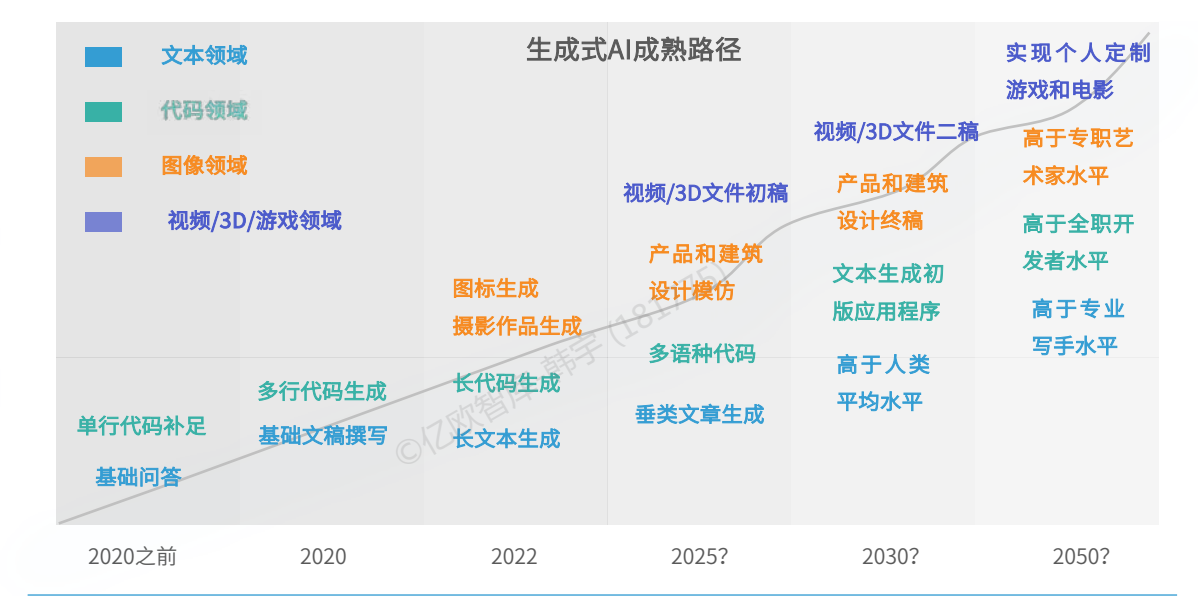
国内部分厂商在Sora问世前推出了产品，但由于对文生视频的预测过于保守和缺乏承担先发风险的厂商，导致目前差距扩大的现状

国内部分AI视频生成产品 vs Sora

公司	产品	视频时长	分辨率	帧率/FPS
OpenAI	Sora	60s	1920x1080和1080x1920之间任意尺寸	高
字节跳动	CapCut AI Video	3s	960x540	24
Morph Studio	Morph Studio	3-7s	1920x1080	8-30
NeverEnds	NeverEnds	2s	1024x576	10
智象未来	Pixeling	4s	1024x576	24
爱诗科技	PixVerse	4s	1024x576	18
右脑科技	Vega AI	4s	1024x576	12
腾讯	Video-Crafter2	2s	1024x640	8

资料来源：红杉资本、国金证券、浙商证券、专家访谈、亿欧智库、公开资料

亿欧智库：多模态生成未来趋势



文生3D发展趋势及赋能行业

- 未来随着数字化的持续发展，3D资产数量的快速增长，3D建模的自动化生成有望成为新的风口。
- AI+3D建模是利用人工智能技术，自动化地生成高质量的3D模型的过程。用户只需要输入关键词或者上传一张2D图片，这些工具就能在短时间内直接生成多个比较初步的3D模型，而用户满意的话，则可以选择进一步生成精度更高的3D模型。



游戏开发



动画设计



建筑设计



增强现实



教育培训



产品设计



文化遗产保护



樊星
松鼠Ai
联合创始人
CTO

展望未来，大模型技术将继续攀登新的高峰，显著提升多模态理解和生成能力，并进一步强化系统的自主学习与推理能力，从而推动更为智能化、自主化的决策与行动的实现。作为一种通用性极强的技术，大模型正逐步与多个领域实现深度融合，其潜在价值正日益凸显。然而，如何将大模型无缝融入各个领域，并将技术与理论紧密结合，以产生实际效益，仍是我们面临的重要挑战。

在教育行业，这一挑战尤为突出。为了成功应对这一挑战，我们需要从多个关键维度进行深入探索，包括提升技术实力、积累丰富的数据资源以及深入研究应用场景等。这些维度将构成我们持续突破和创新的核心动力，为教育行业带来革命性的变革。为学习者和教育者提供更优质、更智能的教育体验。



田小鹏
澳鹏Appen
全球副总裁
大中华及北亚
区总经理

随着“数据二十条”等一系列政策措施相继出台，数据要素市场的探索与发展已步入高速增长阶段。据亿欧预计，2025年数据要素市场规模可达1990亿元，年复合增长率可达25%。尤其是在人工智能快速迭代、大模型与数据相得益彰的发展态势中，数据要素的战略地位进一步凸显。数据是决定机器学习模型性能的三大要素之一。随着各类大模型的智能涌现，数据，尤其是高质量的行业数据，正在成为决定大模型高速发展的关键因素。澳鹏自研的算法模型和核心技术，通过AI赋能数据全生命周期，更快更多地给予AI应用数据养料，为大规模的大模型场景落地提供支持。



余明川
成都智算中心
CTO

算力是人工智能产业发展的基石，大模型时代的算力需求发生了新的变化。一方面算力需求规模呈现倍数级增加，另一方面随着大模型技术的成熟，未来推理侧的算力需求占比将高于训练侧。为了应对大模型时代算力需求的新变化，算力中心需持续加大算力建设，通过集约化扩展、底层架构优化和丰富算力供给类型的方式应对新时代的挑战。

此外大模型的高算力需求使得传统算力中心面临来自能源消耗和散热等方面的巨大压力，绿色算力的发展已成为必然趋势。使用一体化融合供电储能系统、引入液冷技术和提升清洁能源利用率等方式都是未来加速算力中心绿色化转型的有效路径。

- ◆ 亿欧智库经过桌面研究，结合相关公开报道及对相关企业、专家访谈后作出此份报告。报告重点对大模型产业最新发展现状和未来发展进行研究分析，在此，亿欧智库感谢相关企业及业内专家的鼎力支持。
- ◆ 未来，亿欧智库将持续密切关注人工智能领域，通过对于行业的深度观察，持续输出更多有价值的研究成果，助力产业可持续创新发展。欢迎报道读者与我们交流联系，提出报告建议。
- ◆ 特别鸣谢



成都智算中心
Chengdu Intelligent Computing Center



医渡科技
YIDUTECH



INSIGHT 因赛集团

◆ 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：



杨雨然
亿欧智库分析师
Email: yangyuran@iyiou.com



何舒阳
亿欧智库助理分析师
Email: heshuyang@iyiou.com

◆ 报告审核：



王辉
亿欧智库副院长
Email: wanghui@iyiou.com



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追求但不作任何保证。本报告所载的资料、意见及推测仅反映亿欧智库于发布本报告当日之前的判断，在不同时期，亿欧智库可发出与本报告所载资料、意见及推测不一致的报告。亿欧智库不保证本报告所含信息保持在最新状态。同时，亿欧智库对本报告所含信息可在不发出通知的情形下做出修改，读者可自行关注相应的更新或修改。

本报告版权归属于亿欧智库，欢迎因研究需要引用本报告内容，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。

◆ 关于我们：

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



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



网址: <https://www.iyiou.com/research>

邮箱: hezuo@iyiou.com

电话: 010-53321289

北京: 北京市朝阳区关庄路2号院中关村科技服务大厦C座4层 | 上海: 上海市闵行区申昆路1999号4幢806

深圳: 广东省深圳市南山区华润置地大厦 C 座 6 层 | 纽约: 4 World Trade Center, 29th Floor-Office 67, 150 Greenwich St, New York, NY 10006