iResearch
艾 瑞 咨 询

中国人工智能产业研究报告（VI）

山高泽长，AI鼎新自显于时

部门：企业服务三组　　署名：王祺 李冬露

©2024 iResearch Inc.

PREFACE

前言

研究背景：

在问到如何平衡ChatGPT和大学录取的时候，斯坦福大学终身教授李飞飞老师给到了这样的回答，“录取最会使用ChatGPT的前2000名学生是个很有意思的答案。“

能够制造并使用工具成为人类进化史上一道显著的分水岭，而当下如何更好的使用AI工具已然成为人类在产业应用、生产生活与学习工作中的热门议题。随着大模型、生成式AI技术的到来，其强大的数据处理、学习泛化与内容生成能力，高质效加速了各行各业人工智能技术的赋能进程，为AI可赋能的场景领域、扮演角色提供更多创新性与可能性。人工智能应用正加速扩散，渗透到办公、设计、传媒、法律、游戏、教育、汽车等多领域。

艾瑞人工智能研究团队延续既往5年对人工智能行业的市场研究，于第六年聚焦人工智能产业发展进程、发展征程、发展旅程的各个发展阶段，集中探讨中国人工智能产业的发展环境、市场动态、产业机会、发展监管等核心要点，为市场提供有公信力、受到广泛认可的数据与观点。

研究方法：

本报告通过业内资深的专家访谈、桌面研究、案例实证研究、行业对比研究、投融资数据统计输出相应研究成果。

ABSTRACT

摘要

发展进程

2023年，生成式AI为人工智能领域带来重大突破与新的希望，而**围绕生成式AI的政策布局**也迅速铺开，从数据资源和算力基础夯实，到快速对生成式AI规范化引导，再到产业扶持，形成一套强有力的组合拳；**从资本市场来看**，AIGC概念火爆，近40%的投资事件指向2023年新成立的AIGC公司；**从技术发展来看**，大模型基座催化AI工业化生产，加持各AI细分技术赛道的革新发展。而大模型以泛化推理能力见长，小模型以高成熟度、性价比优势仍存市场，大小模型是当下产业应用的核心落点。未来，多模态模型与MOE架构将共同拓展大模型产业空间；**从产业发展来看**，中国AI企业正积极抓住应用探索机会，获取新技术浪潮变现的先发优势，生成式AI进一步加速内容产业的渗透进程。国家也从基础设施角度积极开展智算中心建设，推动AI数据标准建立，鼓励开源数据集发展。2023年中国人工智能产业规模已达到**2137亿元**，预计到2028年，中国人工智能产业规模将达到**8110亿元**，五年复合增长率达到**30.6%**。对比原本大模型未出现涌现能力的人工智能产业规模值，艾瑞测算，大模型带来的产业加成比例在2028年或达到**32.9%**，

发展征程

1）生成式AI产业洞察：2023年，全球进入AI驱动的生产革命，**生成式技术是时代际遇。**预训练大模型的技术架构在多模态路径下优化底层模型训推与理解产出，让决策式AI与生成式AI共筑AI产业发展。国内大模型落地声量加大，行业大模型进入爆发期，其中，医疗与金融为典型落地领域。**从模型模态来看**，生成式AI应用的文本模态达高应用成熟度，代码、语音、图像具备商业化基础；**从商业应用来看**，国家对大模型上线监管采取“备案制”，已有40+家大模型持“证”上岗。艾瑞认为，B端场景出发需逐步渗透打磨，打通业务逻辑实现更多场景的落地应用闭环，呈延续性曲线融合赋能。C端场景需从供给侧满足硬件设备条件及大模型能力适配，在软硬件生态成熟后涌现阶梯式能量爆发。

2）AI产业边缘与端侧洞察：大模型的出现，**加速了AI能力由云向边端多智体的演化进程。**当前，大模型正在从**算力统管和场景优化**两个维度在边缘侧进行落地尝试，部分替代和接管原有云端计算中心的算力调度权限与能力，大大减少云端传输所带来的时间损耗，同时大模型能够取代原有边缘侧用于预测、决策、判别、生成等多类任务的小模型，提升场景泛化能力和使用效果，改善ROI，并正在对自动驾驶技术栈进行全方位升级与重构。AI与终端正在进行加速融合，端侧大模型率先落地于手机、智能座舱等场景。从硬件维度来看，AI重塑操作系统是释放大模型潜力的关键。

发展旅程

从社会层面来看，值得关注的主要风险在于人工智能技术对用户心智、用户隐私及安全伦理问题的潜在影响。**从企业应用来看**，AI技术的内生性缺陷对企业应用的影响更为明显，人工智能框架、数据、算法、模型任一环节都能给系统带来脆弱性。基于上述对人工智能发展风险的探讨，**未来人工智能产业发展需从技术研究、行业标准规范和法律监管三个层面持续完善和引导。**在技术研究方面，提高算法的准确性和透明度，防止偏见和不公平现象出现；在行业标准方面，建立统一的规范和伦理准则，确保人工智能应用符合道德和社会价值；而在法律监管方面，则需制定和修改相关法律法规，保护个人隐私，防止滥用和侵犯权利，由此保证中国人工智能产业稳定实现高质量正向发展。

CONTENTS

目录

01/ 中国人工智能产业进程

AI - ing

2023年人工智能产业活跃动态

人工智能产业进入高速发展期，创造多个技术、市场、监管里程碑

在人工智能发展历程中，2023年必将被载入史册。相比前代AI具备高可用性、高拟人化的预训练大模型跨越技术奇点，国内外技术公司、高校、研究院的语言、图像、视频、音频大模型在2023年以极快的速度相继推出和迭代，基于预训练大模型的应用在全球范围内产生了爆炸式的影响，从社会群众到AI从业者，对人工智能技术能够带来的生产生活变革，都实现了颠覆性的再认识。艾瑞通过技术本身、应用变体、算力支持、政策监管和国际局势五个维度，对2023年AI世界的发展进行全面梳理和俯瞰。

2023年人工智能产业大事记总览

大模型技术进程

大模型小型化

2月25日， *Meta AI在官网公开发布了LLaMA大型语言模型，包括7B、13B、33B和65B 4种参数规模，后续被国内外多款展垂类、端侧大模型作为技术底座。*

多模态模型下场

3月15日， *OpenAI发布了多模态预训练大模型GPT-4，直接开放API。*

文生图模型快速进化

7月27日， *Stability AI正式发布文生图模型——SDXL 1.0。*

大模型落地应用进程

垂直行业化

3月30日， *彭博社发布BloombergGPT论文，该模型专门针对各类金融数据进行训练。*

现有应用融合

2月7日， *微软发布ChatGPT版搜索引擎New Bing，上线48小时内获100万账户申请。*

Agent崛起

11月7日， *OpenAI发布了AI Agent初期形态产品GPTs，并推出了相应的制作工具GPT Builder。用户仅仅通过跟GPT Builder聊天，把想要的GPT功能描述一遍，就能生成专属GPT。*

智算实力加强

端侧算力准备

10月25日

高通推出新一代移动芯片骁龙8 Gen 3，其AI引擎支持多达100亿个参数的生成式AI模型。Hexagon NPU的性能提升了98%，持续的AI推理的每瓦性能提高了40%。

英伟达持续扩大领先优势

11月13日

英伟达发布了目前世界最强的AI芯片H200，性能较H100提升了60%到90%，还能和H100兼容。

国家动态监管

合成数据再度引发关注

6月20日， *国家互联网信息办公室发布关于发布深度合成服务算法备案信息的公告，公开发布境内深度合成服务算法备案信息。*

生成式AI进入监管体系

7月13日， *国家网信办等七部门联合公布的《生成式人工智能服务管理暂行办法》即将自2023年8月15日起施行。*

L3自动驾驶迎来曙光

11月17日， *工信部、公安部、住房和城乡建设部、交通运输部共同发布《关于开展智能网联汽车准入和上路通行试点工作的通知》。*

中美关系限制

网络安全对抗升级

3月2日， *美国拜登政府正式发布新版《国家网络安全战略》，提出中国是“对美国政府及私营部门网络最广泛、最活跃、最持久的威胁。*

AI芯片出口再禁

10月17日， *美国更新出口管制标准，英伟达针对中国市场的特供版H800、A800两款芯片也面临禁售。*

来源：艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

顶层设计驶入深水区，生成式AI成焦点

规范引导、基础夯实和产业扶持三管齐下，促进AI全方位深化发展

自2019年以来，我国人工智能相关政策始终紧随技术和产业发展步伐，历经广泛试点、建设框架、产业化发展、场景化落地四个阶段，切实推动人工智能从一项新兴技术走向规范应用。2023年，生成式AI为人工智能领域带来重大突破与新的希望，而围绕生成式AI的政策布局也迅速铺开，从数据和算力基础夯实，到快速对生成式AI规范化引导，再到产业扶持，形成一套强有力的组合拳。

近年人工智能技术及产业发展相关政策分析

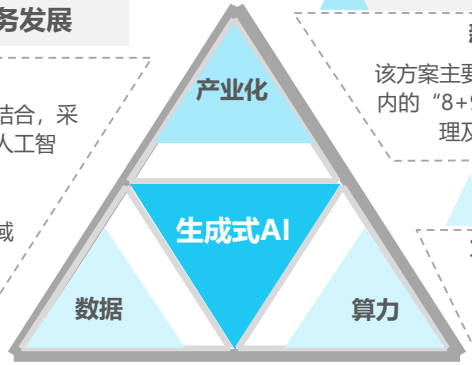
2022-2023年

规范引导——生成式人工智能产业与服务发展

生成式人工智能服务管理暂行办法

原则：坚持**发展和安全并重**、促进创新和依法治理相结合，采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行**包容审慎和分类分级监管**。

信号：明确**鼓励**生成式人工智能技术在各行业、各领域的创新应用，支持专业机构协作；明确要求提供者对**AI生成内容进行标识**；明确规定生成式AI训练全程使用**来源合法的数据**。



产业扶持——大力发展生成式AI及元宇宙产业

新产业标准化领航工程实施方案（2023—2035年）

该方案主要聚焦新兴产业与未来产业标准化工作，形成含生成式人工智能与元宇宙在内的“8+9”的新产业标准化重点领域，其重点任务是通过完善新产业全生命周期管理及各项标准确立，建立新产业标准化工作体系、强化新产业科技创新体系。

基础夯实——数据与算力

算力基础设施高质量发展行动计划

从计算力、运载力、存储力以及应用赋能四个方面提出了到2025年发展量化指标，引导算力基础设施高质量发展。

国务院关于构建数据基础制度更好发挥数据要素作用的意见

提出要建立数据产权制度、数据要素流通和交易制度、数据要素收益分配制度、建立数据要素治理制度。

2019-2022年

初试点

2019

国家新一代人工智能创新发展试验区建设工作指引

到2023年，布局建设20个左右试验区，产出一批重大原创科技成果，创新一批切实有效的政策工具，形成一批人工智能与经济社会发展深度融合的典型模式。

建体系

2020

国家新一代人工智能标准体系建设指南

建设涵盖基础共性、支撑技术与产品、基础软硬件平台、关键通用技术、关键领域技术、产品与服务、行业应用、安全/伦理八大模块构成的人工智能标准体系。

促产业

2021

“十四五”数字经济发展规划

提出要建立数字要素市场体系、提升产业数字化和数字产业化水平、普惠数字化公共服务、完善数字经济治理体系等目标。

推落地

2022

关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见

提出着力打造人工智能重大场景，提升场景创新能力，推动场景开放，加强场景创新要素供给。

2022

关于支持建设新一代人工智能示范应用场景的通知

提出打造形成一批可复制、可推广的标杆型示范应用场景，首批支持智慧农场、智能港口、智能矿山、智能工厂、智慧家居、智能教育、自动驾驶、智能诊疗等十大场景。

来源：中国政府网，艾瑞咨询研究院自主研究及绘制。

©2024.4 iResearch Inc.

AI示范先导区及产业集群初具规模

极点发挥示范引导作用，以点带面形成AI产业集群

我国人工智能重点区域与产业集群建设都取得显著成果。从重点区域看，北京、深圳、苏州等地人工智能产业发展迅猛，北京为其中之最，在中国新一代人工智能科技产业区域竞争力评价指数当中综合分位列第一，从企业扶持、技术探索、产业落地等多个方面建设取得显著成果。2023年，北京市显著加大对人工智能产业扶持力度，将重点通过扩大产业规模、提升链接能力，将人工智能产业打造成为北京市经济增长的新引擎。从产业集群看，我国产业集群发展总体呈现极化与扩散并存的特征，在地域关系上以北京、上海、广东为产业核心，通过区域内部和区域之间的紧密合作，打造出新一代人工智能产业网络空间；在企业簇群上以华为、腾讯、京东、阿里及三大运营商为核心节点，并通过与周边关系节点的交流合作，打造我国人工智能产业的复杂生态。

典型城市：北京市人工智能发展独具特色

发展成果

人工智能科技产业区域竞争力评价指数（2023）

企业实力强劲
截至2022年10月，北京拥有人工智能核心企业1048家，占我国人工智能核心企业总量的29%

落地场景丰富
海淀、朝阳等五区已开始或完成“智慧城市大脑”部署；无人出行从示范运营迈入商业试点；智能工厂加紧建设

学术能力领先
北京人工智能领域核心技术人才超4万人，占全国的60%。人工智能论文发表量居全国第一

产业体系完善
已经形成了全栈式的人工智能产业链

未来规划

扩大产业规模

加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）

提出“核心产业规模达到3000亿元，持续保持10%以上增长，辐射产业规模超过1万亿元”等具体工作目标

增强链接能力

北京市通用人工智能产业创新伙伴计划

聚焦汇聚从算力、数据、模型、应用到投资的产业链上下游合作伙伴，采用用户单位与大模型团队结对方式，构建政产学研用深度融合的协同联动产业体系

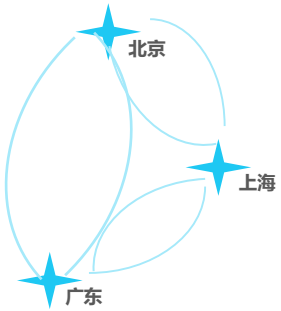
来源：中国新一代人工智能发展战略研究院，北京市人民政府，艾瑞咨询研究院绘制。

产业集群：极化与扩散并存

我国已经初步形成以京津冀、长三角、珠三角为代表的人工智能产业集群，其发展呈现出核心区域持续“极化”，从极点向外逐步“扩散”的整体趋势。

地域关系

从技术合作关系密度来看，北京市、广东省和上海市构成了我国人工智能产业集群价值网络的三个“极点”，这三个区域无论是对内对外，都是技术合作最密集的地区，且仍处于不断扩张的趋势。
虽然整体而言极点内部技术合作多于外部合作，但仅看外部合作，三大极点的对外技术赋能要高于从外部获得的技术输入，这说明从极点向外的技术扩散和带动效应十分显著。



企业簇群

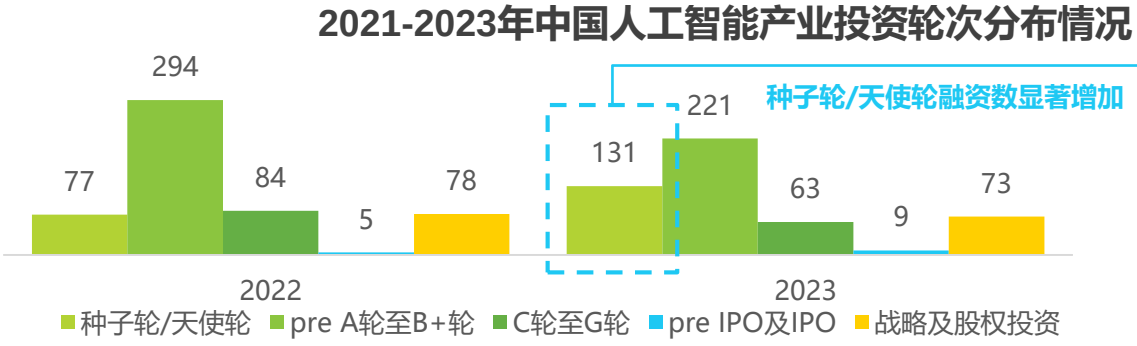
企业簇群是产业集群的关键构成要素，而人工智能企业簇群价值网络同样呈现极核状，少数核心节点是产业发展的主导。关键节点通过与研究型大学、国外基础软硬件供应商、产业智能化企业与地方政府的技术合作与交流，发挥辐射与带动作用，构成产业集群的灵魂所在。

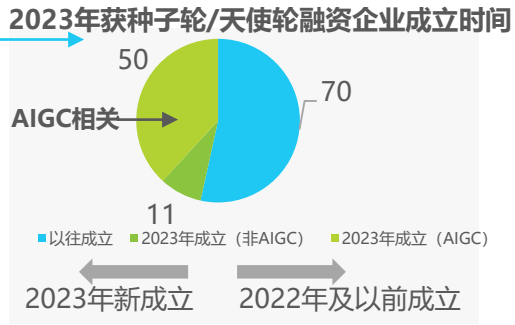
来源：中国新一代人工智能发展战略研究院，艾瑞咨询研究院绘制。

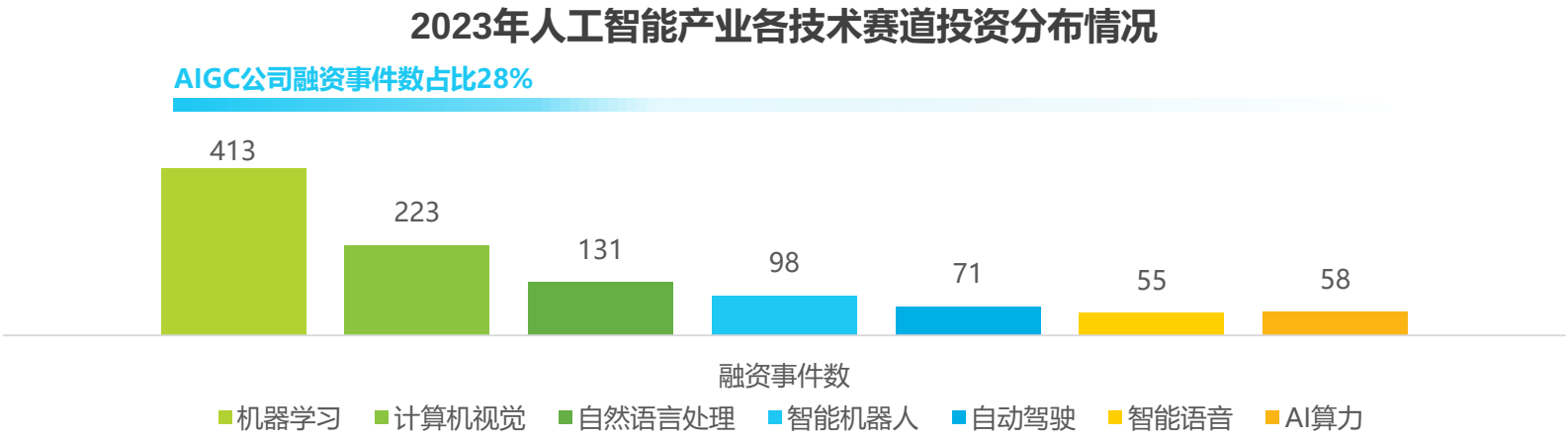
一级市场：AI产业投资风向转变

AI产业开启新一轮融资周期，新概念下原有赛道稳步跟进

从2017-2022年，随着人工智能产业成熟度不断提升，融资逐步向中后期过渡，而在AIGC概念火爆的2023年，种子轮与天使轮融资重返主力位，在这其中，近40%的投资事件指向2023年新成立的AIGC公司，这表明AIGC正在引领AI产业新一轮融资周期。与此同时，原有AI各技术赛道也依然保持活力，机器学习使用最广，存在感最为明显，计算机视觉、NLP依然紧随其后。值得注意的是，AIGC相关融资占2023年AI产业全部融资事件的28%，并且在除机器人外的各个赛道均有渗透，成为年度AI融资的最大关键词。







注：2023年共497条有效数据；因一家公司可同时具备多项AI技术，图表2不同技术赛道融资事件互有重合。
数据来源：IT桔子，艾瑞咨询研究院自主研究撰写

©2024.4 iResearch Inc.

一级市场：新的投资逻辑创造资本神话

资本抢注“有背景”的AIGC团队，构建新格局时代，企业资源决定成败

2023年中以前，许多投资人对AIGC持观望态度，但仍有不少资本势力躬身入局。从全年战绩来看，2023年资本缔造了5家中国AIGC独角兽和一家准独角兽，其中4家为2023年新开项目，“AIGC速度”充分证明这一轮AI技术爆发对传统AI赛道投资逻辑的改变。一方面，资本对AI所能提供的商业价值普遍产生新的认识，另一方面，从具体公司来看，资本明显看好名人创业+大模型团队的配置。筑造优秀大模型技术企业需要算力资源、数据资源及生态资源的多方加持，方能从技术研究走向商业生态的长久闭环。未来AIGC应用创业公司将成为未来赛道健康成长的关键支柱。

AIGC带来AI投资逻辑新变化

价值视角——纵观AI赛道，进行价值再认识

关键词：

可批量化复制 → 技术落地快 → 适用范围广

传统AI赛道的典型模式为**AI技术+垂直场景+项目制开发**，在可持续经营和规模化扩张能力方面稍显不足，而大模型的技术特征和应用效果，让市场对于AI的商业价值产生了全新的认知与期待：

投资策略——落回具体标的，重视公司的资源整合能力

在看好AI赛道的前提下，对投资标的选取也呈现出与以往不同的显著特点：

关键词：

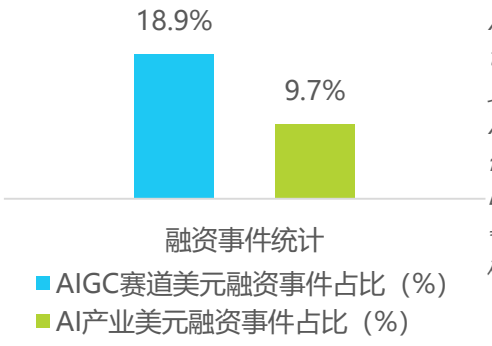
资源整合能力

大模型创业需要大量高端技术人才与算力储备，在如今市场当中属于顶尖稀缺资源，有行业影响力的创业者更具有资源凝聚力。在国内5家AIGC独角兽中，有3家为国内科技圈“大佬”牵头打造。

动态的投资视角

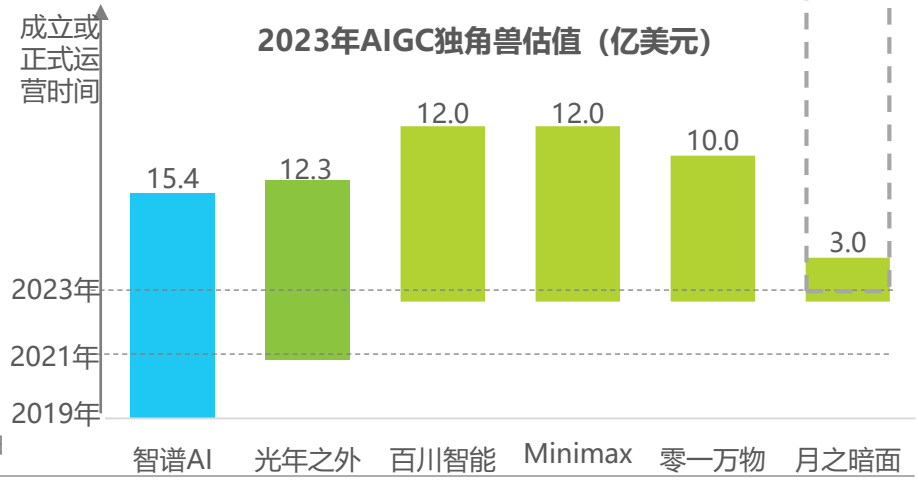
AIGC作为新兴赛道，又生长于紧张的国际局势当中，其影响因素诸多且实时变化，资本会以长期动态眼光看待AIGC公司的发展，这也将使得资本对市场动作的反应速度进一步加快。

2023年AI产业及AIGC赛道美元融资事件占比（%）



AIGC技术的上述特性也使得AIGC公司成为更加国际化的投资标的，AIGC赛道美元融资事件占比近20%，大大超出整个AI产业美元融资事件比例，冲刺美股将趋势已显现。

2023年AIGC独角兽估值（亿美元）



注：月之暗面于2024年2月估值已达到25亿美元，本报告数据截止2023年12月31日
数据来源：IT桔子，艾瑞咨询研究院自主研究绘制

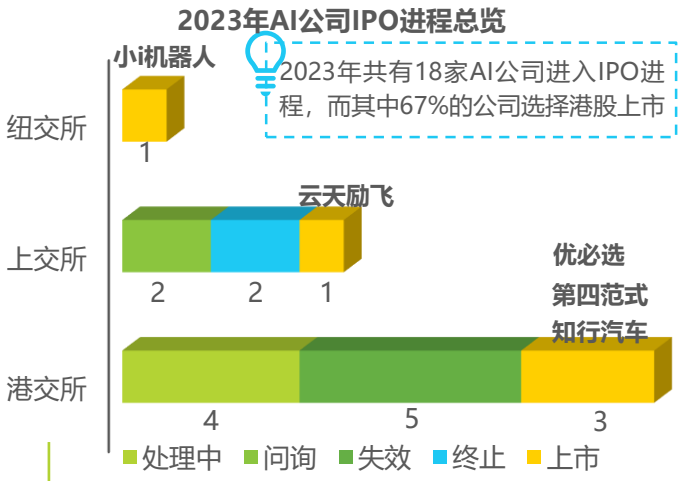
©2024.4 iResearch Inc.

二级市场：AI公司IPO机遇与挑战

港股将为AI公司开放更大窗口，长期盈利能力证明是过审关键

2023年，共有18家AI公司进入IPO阶段，而其中12家选择港股上市，除了全球化募资的野心外，更大原因在于港股对高新技术企业上市的财务指标要求表现出明显的宽松倾向，预计未来会有更多AI公司选择冲刺港股。与此同时，AI公司上市过程仍旧相对反复曲折，许多公司多次重复交表，而进入问询阶段后，企业面临对长期盈利目标与战略更加严格细致的拷问。

2023年AI公司IPO进程分析



赴港IPO的12家AI公司中，5家处于失效状态，有2家曾在失效后重新提交资料；在上交所IPO的5家AI公司中，有两家上会被否，AI公司上市依然面临不小的挑战。

交易所关注问题排名：

1、长期盈利能力

对于无账面亏损企业不会问询这一问题，对于账面亏损企业必须提出扭亏为盈的措施与依据

2、核心竞争壁垒

需证明公司技术先进性

来源：艾瑞咨询研究院自主研究绘制

上市路径选择——港股成为AI公司首选

2023年3月31日，港交所18C条款生效：

香港联合交易所在主板《上市规则》中新加入第18C章，针对**特专科技公司**特征，对其*净利润、营业收入、现金流考察标准均有放宽，甚至在研发费用率和估值符合一定条件下，允许未商业化公司上市融资*。而黑芝麻也成为国内第一家采用18C条款进入港股IPO程序的AI公司。如内陆交易所没有更多利政策释放，未来AI公司港股上市优势将进一步扩大。

特专科技行业	细分领域
新一代信息技术	云端服务
	人工智能
先进硬件及软件	机器人和自动化
	电动及自动驾驶汽车
先进材料	元宇宙
新能源及节能环保	……
新食品及农业技术	……

上市过程卡点——交易所对AI公司长期盈利能力判断更加审慎

公司名称	赛道	进展	问询内容
节卡机器人	智能机器人	已问询	首先关注核心技术先进性及技术来源问题；其次关注可持续经营能力（客户交易、收入增长、毛利、费用、现金流可持续性）
百奥赛图	AI医疗	已问询	主要围绕公司盈利能力展开全方位提问，首要关注现在亏原因，改善亏损措施，其次关注公司核心竞争壁垒，如技术先进性、团队、收入与客户等。同时关注成本、费用等
思必驰	智能语音 NLP	上会被否	第一轮：对思必驰的“持续经营能力”提出疑问，要求说明是否具备扭亏为盈的基础条件和经营环境，说明扭亏为盈的测算依据及合理性，审慎论证是否具有客观性和可行性。 第二轮：要求思必驰说明预测的未来年度营业收入及复合增长率是否合理、审慎，是否具备实现的基础。 第三轮：再次要求思必驰说明相关营收增长预测是否合理、谨慎，扭亏为盈的测算依据及合理性，审慎论证是否具有客观性和可行性。

©2024.4 iResearch Inc.

二级市场：AI上市公司表现接近大盘

AIGC概念并非万金油，投资者预期收缩

2023年AI上市公司未显现逆势上行，投资者买单AIGC概念

2023年，A股大盘指数在跌宕中呈下降态势，国内AI上市公司共计**36家**（含美股与港股），其中2023年末收盘价相比年初下跌的有24家，总体与A股大盘全年态势基本符吻合。

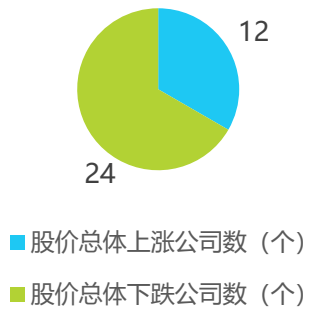
同时，也有部分AI公司及AI概念股出现明显上涨趋势，AI领域如计算机视觉、NLP、智能语音与AI数据领域，基本符合AIGC概念从算力+数据+相关算法全链条对相关公司的利好逻辑，但由于国内大部分AI芯片厂商未上市，二级市场反馈并不明显。AI概念股中，与AIGC相关的云基础设施公司如浪潮云，属于AIGC最直接应用场景的网文、影视公司如掌阅、天娱数科等，也受到投资者青睐。

AI上市公司股价变动分析

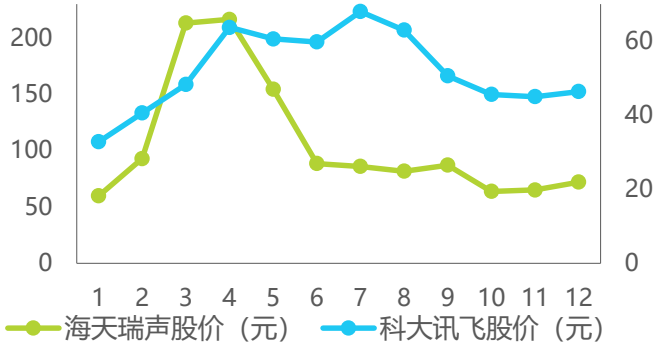
市场对AI“故事”正在脱敏，AI公司需要尽快交出盈利答卷

2023年新上市AI公司共5家，分别为第四范式、小i机器人、云天励飞、知行汽车与优必选，涵盖了AI领域大部分赛道，*上市后股价均不容乐观*，而其中无论是作为为传统机器学习头部厂商如第四范式，还是典型以NLP、智能语音技术提供智能坐席等服务的厂商小i机器人，都有明确拥抱AIGC的动作和规划，但从实际结果而言，资本市场并未买单。*AIGC概念股在2023年也普遍经历了一轮猛烈上涨后回落。*从交易机构审查到真正进入投资者视野，AI公司讲好故事仅仅是第一步，找到切实可行的盈利方向迫在眉睫。

2023年中国AI上市公司股价变动情况



主打AIGC概念的AI上市公司2023年股价走势



AI上市公司一览

赛道	公司名称	上市时间	上市板块
机器学习	百融云创	2021	港股
	商汤科技	2021	港股
	创新奇智	2022	港股
	云从科技	2022	科创板
	易点天下	2022	创业板
	第四范式	2023	港股
计算机视觉	海康威视	2010	中小板
	中科信息	2017	创业板
	虹软科技	2019	科创板
	天准科技	2019	科创板
	医渡科技	2021	港股
	罗普特	2021	科创板
	鹰瞳科技	2021	科创板
	格灵深瞳	2022	科创板
	凌云光	2022	科创板
	联影医疗	2022	科创板
NLP	拓尔思	2011	创业板
	万兴科技	2018	创业板
	小i机器人	2023	美股
智能语音	科大讯飞	2008	中小板
	汉王科技	2010	中小板
AI数据	美林数据	2014	新三板
	数据堂	2014	新三板
	海天瑞声	2021	科创板
AI芯片	紫光国微	2005	中小板
	北京君正	2011	创业板
	中基国威	2018	新三板
	瑞芯微	2020	主板
	寒武纪	2020	科创板
	复旦微电	2021	科创板
	安路科技	2021	科创板
	云天励飞	2023	科创板
自动驾驶	四维图新	2010	深交所
	知行汽车	2023	港股
智能机器人	科沃斯	2018	主板
	优必选	2023	港股

注：本报告定义AI上市公司为以提供AI产品或服务为主营业务的公司，非AI概念股。

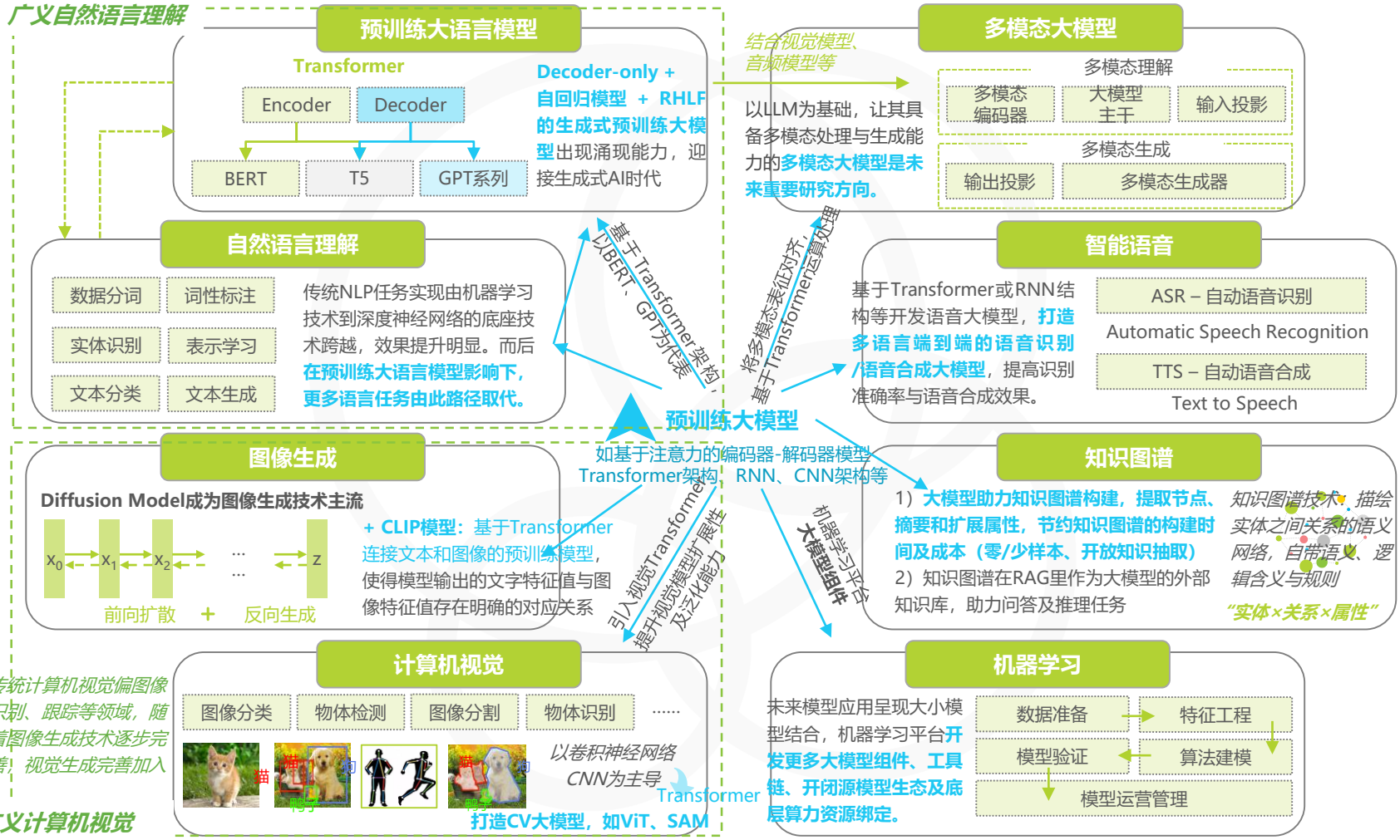
数据来源：同花顺、艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

大模型加持AI技术赛道革新发展

大模型基座催化AI工业化生产，Decoder only路径引领生成式AI产业变革

基于模型底座的AI技术迭代图



来源：《MM-LLMs: Recent Advances in MultiModal Large Language Models》，艾瑞咨询研究院根据公开资料、专家访谈自主研究绘制。

©2024.4 iResearch Inc.

在自然语言处理能力上不断突破创新

突破语言理解能力、文本处理长度、知识增强等技术，缓解LLM幻觉问题

大模型文本能力提升路径

01 大模型的上下文本长度不断增加，提升文本处理能力

理解Token的概念

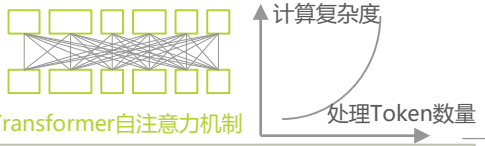
"模型可以理解和生成的最小文本单位"
"能够被编码的最小单元"

1 token ~= 英文中的4个字符

1 token ~= ¾个单词　　100 tokens ~= 75个单词

且相较于英文，中文语义需要更多token表示

大模型上下文支持更多token数的难度



- 计算复杂性与token数量的平方成正比，**增加计算复杂度、消耗资源及响应时间。**
- 上下文支持需存储输入和输出token，**对内存需求有更高要求。**

国内外大模型在上下文本长度的突破进展

OpenAI
- 2024.4 GPT-3.5 Turbo 4K
- 2023.6 GPT-3.5 Turbo 16K
- 2023.7 GPT-4 8K/32K
- 2023.11 GPT-4 Turbo 128K

Google
- 2023.12 Gemini 32K
- 2023.4 Gemini 1.5 Pro 1000K

百川智能 BAICHUAN AI
- 2023.10 Baichuan2- 192K

零一万物 01.AI
- 2024.3 Yi-34B-Chat-200K

Moonshot AI
- 2023.10 Kimichat- 200K 20万汉字
- 2023.10 Kimichat- 2000K 200万汉字

大模型厂商在上下文长度PK火热，**然而部分模型测试时发现文本长度增加时，准确率呈现下降的趋势，**需回归本心，上下文长度是为了更好的应用。

02 检索生成增强RAG成为大模型外挂知识库的有力帮手

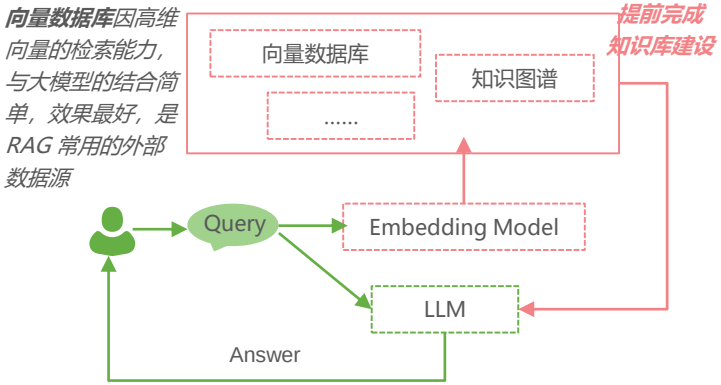
大模型的应用痛点

◆ 存在幻觉问题　◆ 信息时效性问题　◆ 专业知识需微调定制投入

检索生成增强RAG技术

检索增强生成技术（Retrieval-Augmented Generation，RAG），用从其他地方检索到的附加信息来补充用户输入到大型语言模型 (LLM)。

***向量数据库**因高维向量的检索能力，与大模型的结合简单，效果最好，是RAG 常用的外部数据源*



◆ 通过更新知识库具备时效性　◆ 结合用户知识、数据实现个性化与扩展性　◆ 满足专业领域回答，提升回复质效

大模型幻觉问题

1）通过延长上下文长度泛化模型能力，拓宽认知边界，缓解幻觉问题

大模型幻觉问题：输出内容看上去合理、有逻辑，甚至可能与真实信息交织在一起，但实际上却存在错误的内容、引用来源或陈述，是影响大模型规模化应用的核心问题。**业界提出诸多解决办法，如延长上下文、外挂知识库、微调、提示工程等。**

3）**微调：**基于特定数据集重新训练微调模型，需时间与资源投入，若保证时效性则需定期更新
4）**提示工程：**通过优化Prompt缓解幻觉问题，对提示生成要求较高，且专业化领域可用度低

2）外挂知识库，结合知识检索完成人机交互，缓解幻觉问题

来源：艾瑞咨询研究院根据专家访谈、公开资料自主研究绘制。

©2024.4 iResearch Inc.

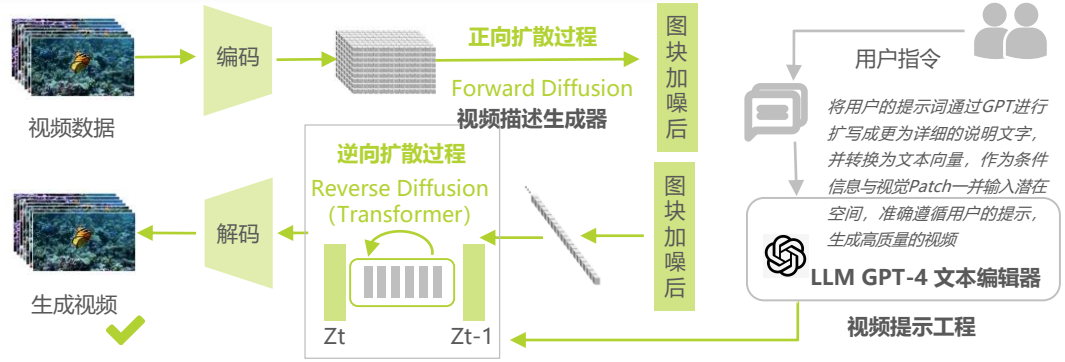
在计算机视觉赛道优化补全生成能力

AI初具对世界的三维理解与创造能力，Sora模型为全球带来更多想象

2024年2月，OpenAI发布Sora模型，在全球范围内引起剧烈反响。Sora是一个以视频生成为核心的空间模型。它的出现，标志了DiT（Diffusion Transformer）架构的融合成功，且在视觉领域同样可以出现涌现能力，未来持续迭代有望进一步提升视觉模型的生成效果，可喻为视觉生成领域的“GPT 3”时代。此外，随着模型计算规模逐渐增大，模型在物理世界的关键特征、数字和内容的标注理解下，成功建立相应的物理特征与数字关联，具备模拟现实世界中人类、动物和环境，甚至事件生成的能力。因此未来Sora模型不仅可以在影视、医疗、教育等领域提供生产力角色，还能基于对世界空间的认知理解，服务于空间模拟、视频计算、数字孪生等深层需求，成为一款更具通用能力的世界模型。

引爆全球的“世界模拟器” Sora

Sora的技术原理



Sora的技术局限性

- ***不能准确地模拟许多基本交互的物理过程***，如玻璃破碎等。
- ***物体交互并不能总是产生正确的物体状态变化***，如吃东西等。
- ***描绘生成复杂动作或捕捉微妙面部表情方面仍然有待提升***
- 由于形状、大小、位置及与其他手指的关系极其复杂多变，***手指生成有待提升***
- ***生成视频没有声音***，后续结合应用场景可进一步生成声音配合

.....

Sora的技术优势及战略意义

Sora 功能实现：60s视频长度

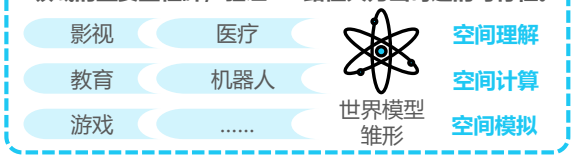
文本生成图像	文本生成视频	视频延长
图像生成视频	视频连接	视频编辑

Sora 技术优势：

- ✓ ***连贯性、一致性、稳定性：****人物和背景的连贯性、前后主题的一致性与稳定性及物体恒存性等*
- ✓ ***视频内容生成的灵活性：****可生成不同尺寸、像素、画幅的图像视频，并完成图生视频及视频延长、视频连接、视频编辑等功能*
- ✓ ***视频内容对物理规律的理解：****没有人为约束下视频内容满足物理学规则，但仍存在局限性*

Sora 战略意义：视觉的“Scaling LAW”

第一个确认展示出涌现能力的视觉模型，是计算机视觉领域的重要里程碑，验证DiT路径大力出奇迹的可行性。



来源：Open AI官网、Open AI官方文件，艾瑞咨询研究院根据公开资料、专家访谈自主研究绘制。

©2024.4 iResearch Inc.

“大小模型融合赋能”是当下核心应用落点

大模型以泛化推理能力见长，小模型以高成熟度、性价比优势仍存市场

“ChatGPT爆火后，NLP技术不存在了”，这类说法在2023年讨论的如火如荼。而艾瑞与人工智能产学研厂商深度交流后认为，NLP小模型仍在被广泛应用，为供给侧厂商完成意图识别、检索匹配等任务。NLP技术不存在更多是从前瞻性学术角度来看，而从产业应用角度，大小模型结合仍是人工智能产业的当下核心应用落点。而随着智算规模扩张、大模型能力提升及应用成本降低之后，大模型的确会对小模型的更多应用场景展开替代趋势，尤其是在大模型擅长的归纳推理、内容生成等语言语音应用场景。

中国人工智能产业应用落地大小模型应用逻辑

大小模型结合应用：从需求侧角度出发，客户并不会核心目的不是对于大小模型的选择，而是AI产品方案的实现与应用，因此供给侧厂商目前普遍采用大小模型结合的办法达到成本效益的最优化。

大模型具备泛化与深层推理理解力

视觉场景：从图像识别角度来看，CV大模型具备场景泛化与更深层理解推理能力，可应用在工业、自动驾驶、安防园区等复杂场景；从图像生成角度，Diffusion大模型为技术底层架构。

语音场景：语音大模型架构，模型层面有效解决小语种、方言等小样本问题，交互层面提升内容理解识别能力与拟人化生成能力，衍生音乐生成、音乐创作等场景

数据分析、语言应用场景：大语言模型擅长知识归纳、理解推理、总结生成等语言类任务，在搜索问答、内容创作、知识助手、角色扮演等领域率先得到应用，其归纳推理能力也应用于企业数据中，以对话交互形式优化BI分析呈现。

小模型具备高成熟度与性价比

视觉场景：如安防 、人脸识别等场景，CNN、RNN等小模型成熟度高，CV产品应用具有高实时性与高性价比

语音场景：人机对话采用的ASR、TTS等小模型已发展成熟，基于性价比、时延等要求，多数简单对话场景仍应用小模型

数据分析、语言应用场景：在与行业知识、业务数据紧密结合的时候，当下供需两侧出于性价比、小模型专业度等原因仍会采用小模型或者搭配使用。

大模型剪裁剪支

企业实际应用时，会对千亿级大模型进行剪裁优化，缩减至数百亿、数十亿参数的模型，此时艾瑞仍认定为本页大模型讨论范围。

来源：艾瑞咨询研究院根据专家访谈、桌研资料自主研究绘制。

“集大一统”的多模态模型是未来发展要点

多模态与MOE共同拓展大模型产业空间

从产业发展视角，当前大模型明显的痛点一是适配场景有待发掘，二是落地成本偏高。现实世界当中的数据往往是散乱且混合多模态类型，尤其对于自动驾驶、安防等人工智能产业的主战场，多模态模型相比单一模态，其适用场景将有数倍增长。另一方面，从落地成本出发，大模型剪枝虽然能够有效缩减参数，但也面临应用效果的折扣。MOE架构通过专家模型之间的合作和调用，在降低模型应用成本的同时，还能提升应用效果，将成为未来大模型技术拓展的重要方向。

单模态、款模态向多模态：开启大量潜在应用场景

为什么需要多模态

- 信息的理解能力增强，泛化性好
- 可在更小参数量级实现“涌现”
- 可用更少的数据量实现较好的微调效果

多模态模型的行业应用

自动驾驶

汽车需实时采集理解路况、车况，涉及图像、语音、文字数据，多模态模型是自动驾驶必须攻克的难题

泛安防

原有视觉检测模型只能执行单独任务，而叠加语言模态能实现模型对场景的理解

工业互联

工业生产数据有温度、湿度、触觉、图像等多维度，多模态模型能真正打通不同模态数据，实现对工业生产环境完整的理解和控制

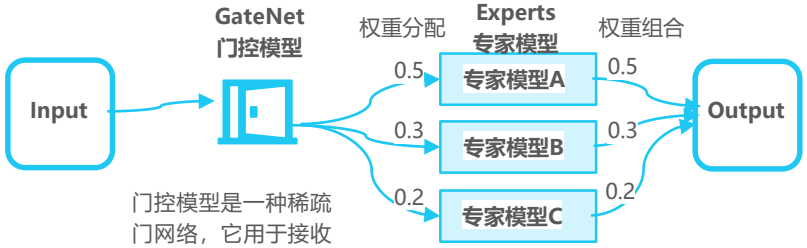
多模态模型的技术路径

多模态大模型领域的技术栈尚未收敛，但当前主流方式基本以大语言模型为核心，主要手段将其他模态数据统一转化为LLM能够理解的向量表征，从而实现语言与其他模态数据在理解和输出方面的对齐（详见下图）。其余方式还有通过提高大语言模型对其他模态数据感知能力、采用工具辅助或是数据驱动的方法。

资料来源：《NExT-GPT: Any-to-Any Multimodal LLM》，艾瑞咨询研究院自主研究绘制。

单一架构向MOE架构转变：改善大模型落地成本

MOE架构(混合专家模型）：技术原理



门控模型是一种稀疏门网络，它用于接收输入信息并判断该任务由哪个专家模型处理，并分配每个专家模型的处理权重。

训练时，门控模型将任务分配到不同的专家模型；在推理的过程中，这些专家模型会针对输入的数据，产生相应的输出。这些输出最后会按初始权重进行加权组合形成最终输出结果。

MOE架构优势

计算效率提升 推理成本降低	任务处理精准性提升 可解释性提升	处理大规模数据和复杂任务效果提升
在执行具体任务时只有少数专家模型被激活，无需全盘调用，因此能够大量节约算力，同时计算效率也得到提升。	通过将特定任务指派特定模型处理，能提升输出内容精准度，同时由于分配到特定模型，也提升了模型的可解释性。	门控模型能够有效进行复杂任务拆解，将大规模数据分解为小模块进行处理，因此其任务复杂度和输入数据上限也会增长。

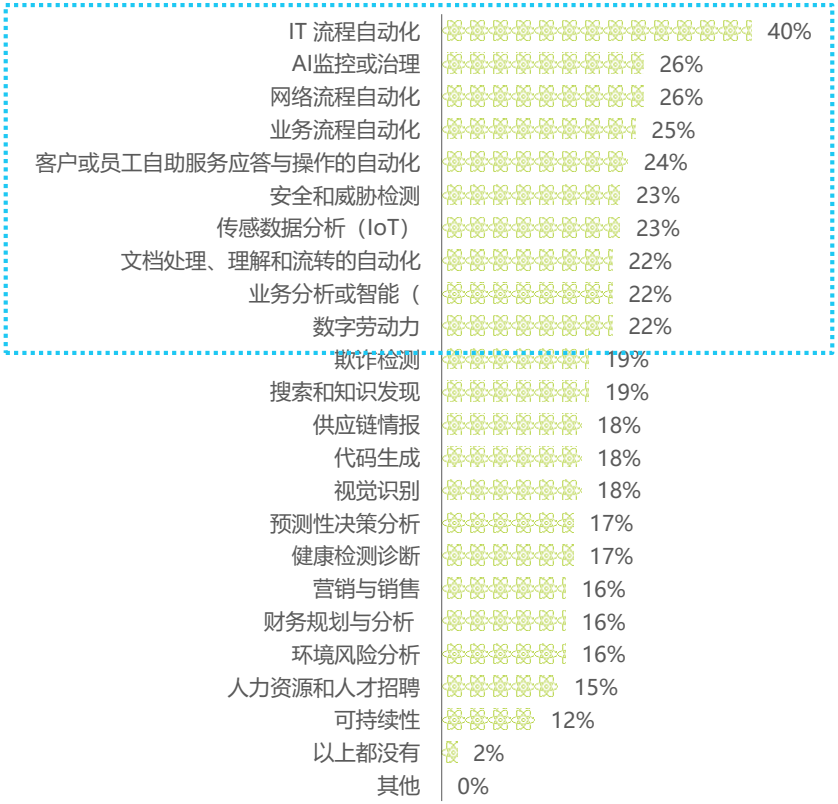
来源：艾瑞咨询研究院自主研究绘制。

人工智能产品实现有序应用

AI及自动化技术有序应用在中国企业的IT网络流程与业务职能部门

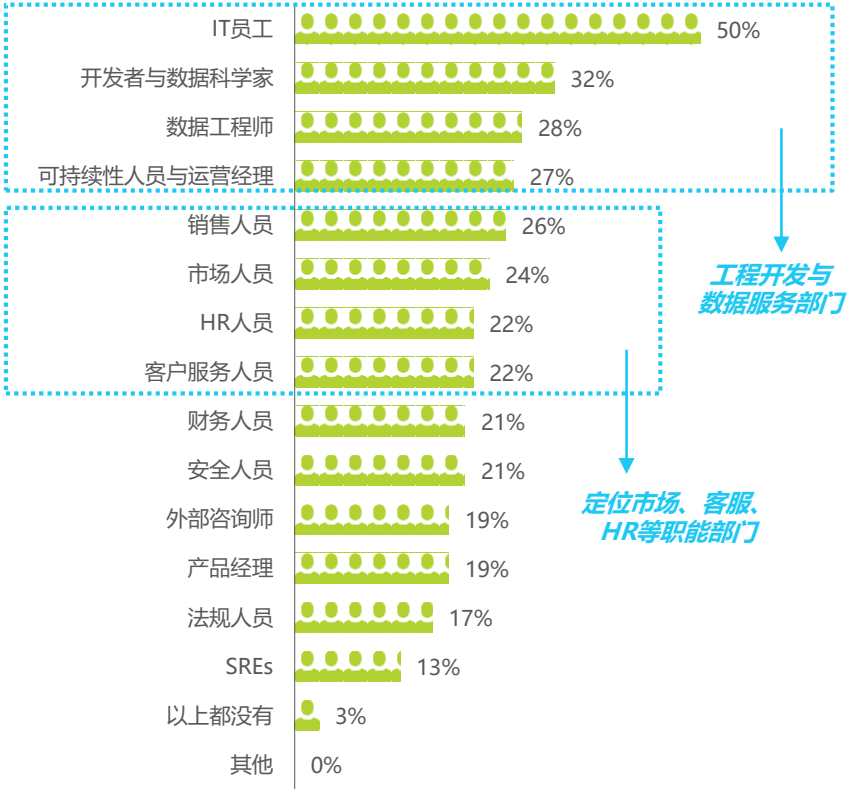
全球的自动化及人工智能浪潮正以前所未有的速度推进，深刻重塑着各行各业的运作方式。AI技术在数据分析、机器学习、自然语言处理等方面取得了显著进步，不仅极大地提高了生产效率和服务质量，还显著推动了新产业的诞生和旧产业的转型升级。根据IBM发布的《2023年全球AI采用指数》数据，中国已将AI及自动化技术运用在IT、网络流程、业务流程等业务领域，并将AI技术广泛服务于IT开发、人员运营、销售市场、客户服务等部门人员。

中国企业应用AI与自动化技术的业务比例



来源：《2023年全球AI采用指数》，IBM，艾瑞咨询研究院研究绘制。

中国企业应用AI技术的人员比例



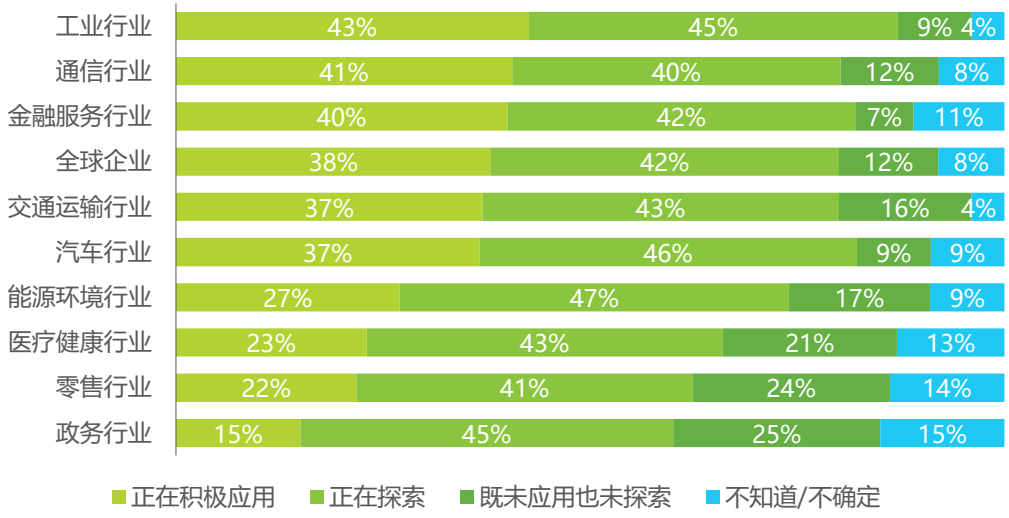
来源：《2023年全球AI采用指数》，IBM，艾瑞咨询研究院研究绘制。

生成式AI产品初衷更在价值提升

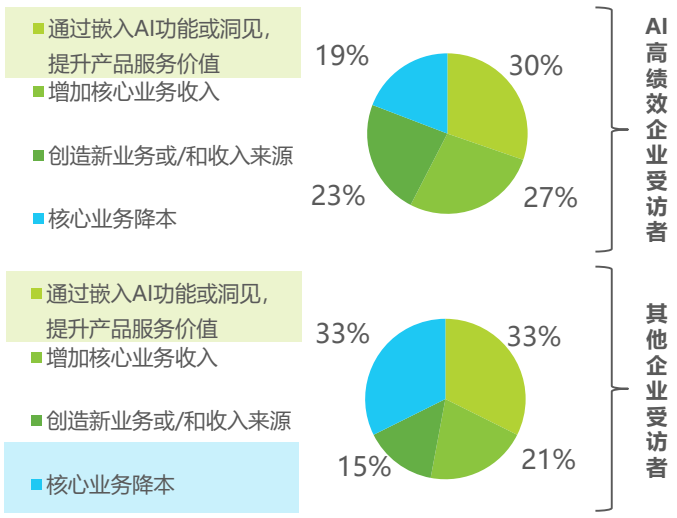
生成式AI产品率先落地于营销销售与产品开发等场景

生成式AI的产品价值在于其强大的内容生成能力，能够为用户提供高度个性化的内容生产，满足企业内外服务的多样化需求。根据IBM发布的《2023年全球AI采用指数》数据，以工业、通信、金融为代表的行业企业是拥抱生成式AI产品的领域先行者，且相较于原本对标“降本增效”的AI产品，生成式AI产品的首要目标更多在于“提升产品服务价值”，尤其是AI高绩效企业表现更为明显。

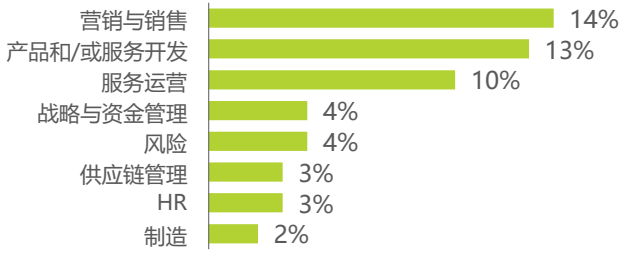
各行业企业对于生成式AI产品的应用现状



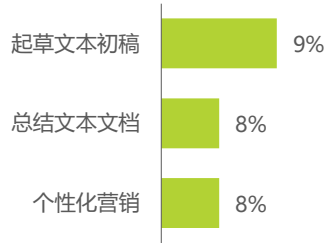
组织生成式AI产品的首要目标



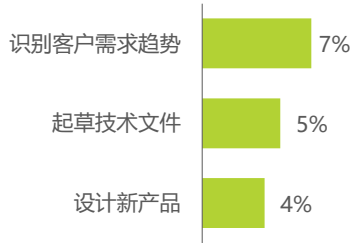
生成式AI产品的常用受访者占比



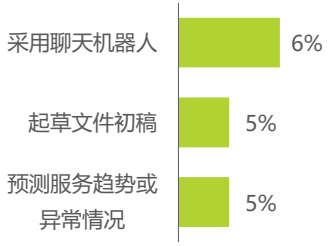
营销与销售常见用例



产品服务开发常见用例



服务运营常见用例



来源：《2023年全球AI采用指数》，IBM；《生成式AI的突破之年》，麦肯锡，艾瑞咨询研究院研究绘制。

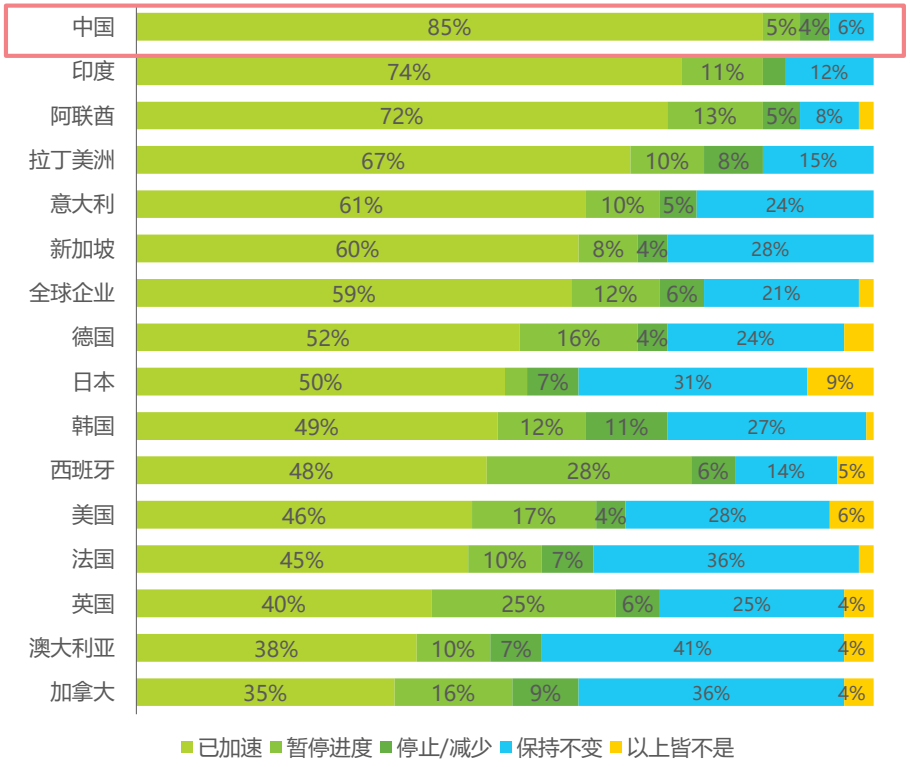
©2024.4 iResearch Inc.

中国对AI的关注与应用位于全球前列

中国AI企业正积极抓住应用探索机会，获取新技术浪潮变现的先发优势

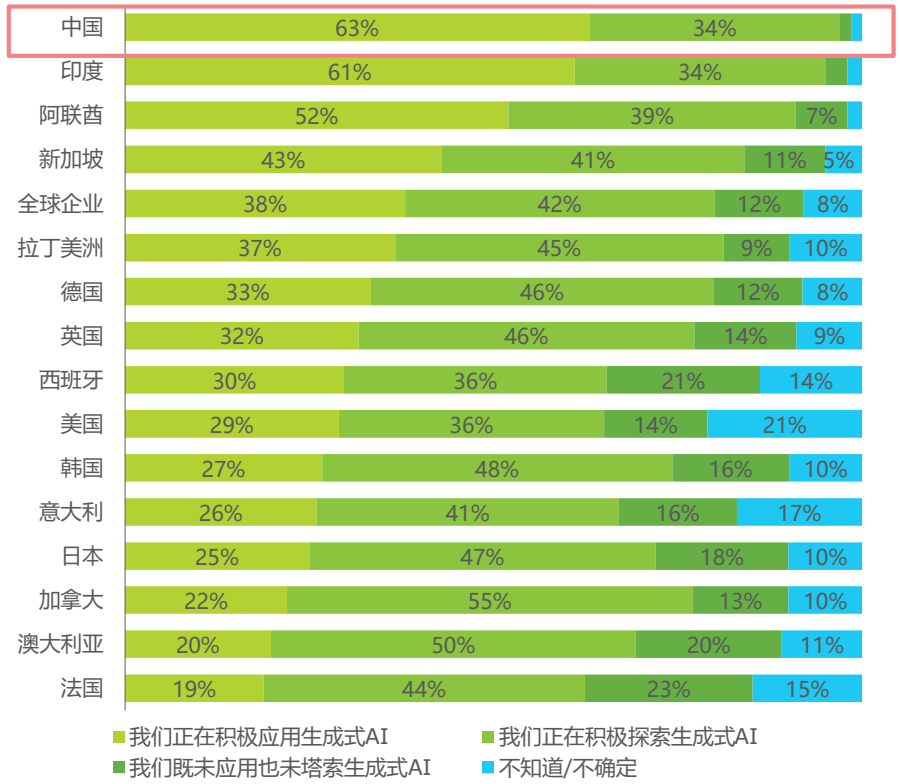
根据IBM发布的《2023年全球AI采用指数》的数据显示，2023年，有高达85%的中国企业表示在过去的一段时间里加快了对AI的投入应用，63%的中国企业表示正在积极应用生成式AI，34%的中国企业正在积极探索生成式AI。全球范围内，中国展示出了对AI应用的超前积极姿态，不仅关注投入AI技术的前沿动态，更致力于AI落地探索的实际应用，以获取新技术浪潮下的新一轮竞争性优势。

全球不同国家已探索/应用 AI的IT公司对AI产品的投入变化



来源：《2023年全球AI采用指数》，IBM，艾瑞咨询研究院研究绘制。

全球不同国家对于生成式AI的应用现状



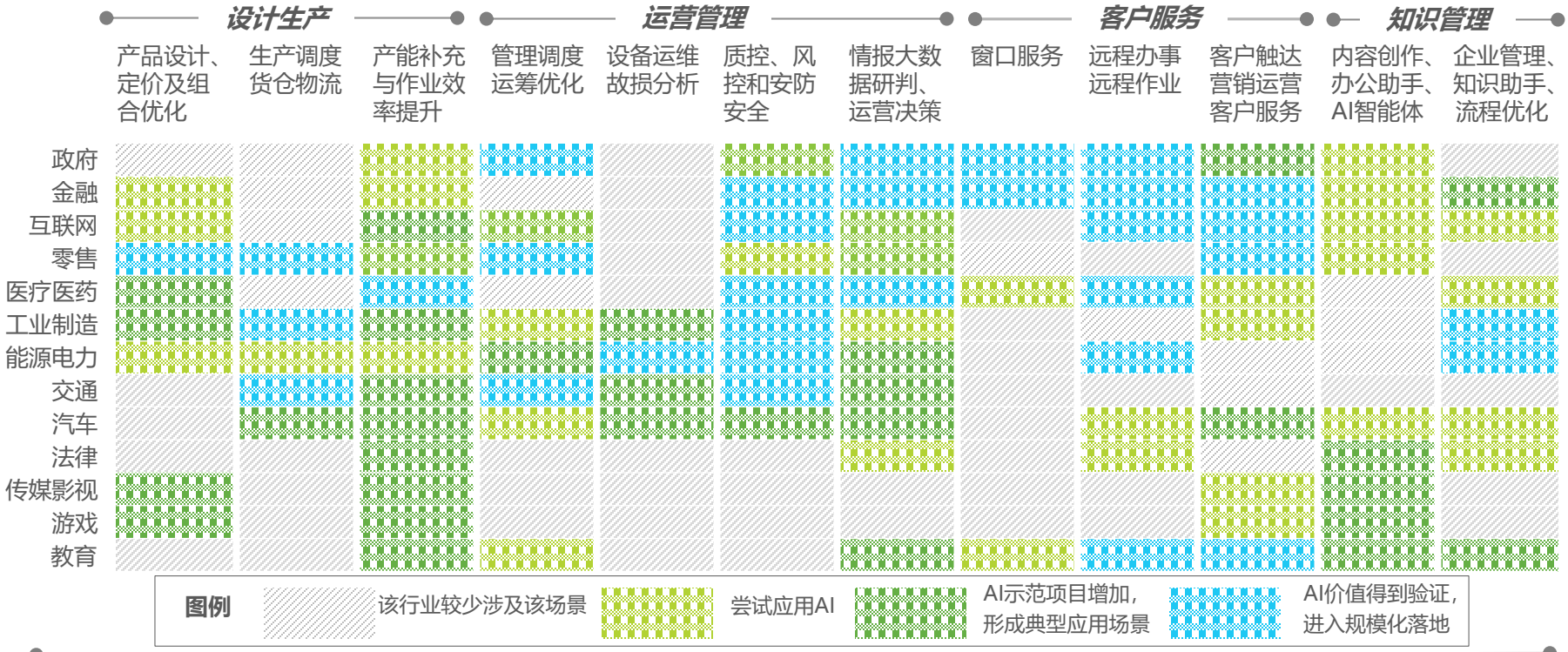
来源：《2023年全球AI采用指数》，IBM，艾瑞咨询研究院研究绘制。

中国AI+行业进程加速渗透

决策式AI与生成式AI的共同赋能，生成式AI加速内容产业的渗透进程

人工智能技术正与人类经济生产活动的主要环节达成紧密结合，提供生产办公效率提升、运营管理优化、服务体验增强等效果实现。而随着大模型、生成式AI技术的到来，其强大的数据处理、学习泛化与内容生成能力，高质效加速了各行各业人工智能技术的赋能进程，为AI可赋能的场景领域、扮演角色提供更多创新性与可能性。在对原本计算机视觉产品、对话式AI产品、决策智能产品完成能力优化外，衍生出更多文本生成、代码生成、图像生成等产品功能，由技术底层实现内容生产效率的飞越，并有望进一步变革人机交互方式，以对话形式降低人机交互门槛，高维度优化用户交互体验。

人工智能技术广泛渗透进经济生产活动主要环节



来源：艾瑞根据公开资料自主研究绘制。

©2024.4 iResearch Inc.

中国人工智能产业图谱

艾瑞咨询

2023年中国人工智能产业图谱

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

©2024.4 iResearch Inc.

中国人工智能产业规模

2028年中国人工智能产业规模将超8000亿元，五年复合增长率达到30.6%

根据艾瑞咨询研究院测算，2023年中国人工智能产业规模已达到2137亿元，大模型带来的底层技术革新将为中国人工智能产业的规模增长带来更多存量扩张与增量空间。2028年，中国人工智能产业规模将达到8110亿元。对比原本大模型未出现涌现能力的人工智能产业规模值，艾瑞测算，大模型带来的产业加成比例在2028年或达到32.9%，在语言语音模态规模加成最为显著，未来大语言模型、语音大模型的产品门槛与应用成本将逐步降低，带来更多API能力调用与产品解决方案的AI能力融入发展，尤其在2024年以后，更多AI产品逐步变现、AI能力下放至边缘侧与端侧之后的影响将更为明显；原本以图像识别为主的计算机视觉市场增长变缓，受政策及政府预算影响，泛安防类的业务增长更多被医疗、工业等CV产品取代，且图像生成市场将在未来3-5年迎来更多商业变现机会，进一步填充计算机视觉模态的市场空间驱力。

中国人工智能产业规模盘点

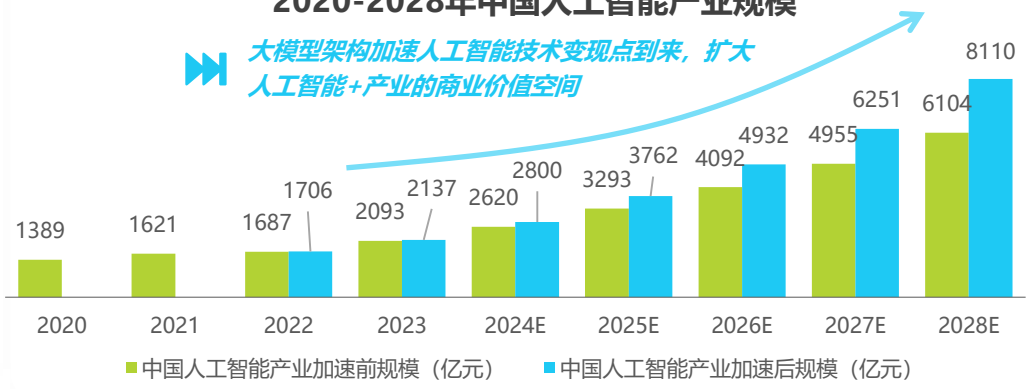
2028年大模型产业规模加成

30+%

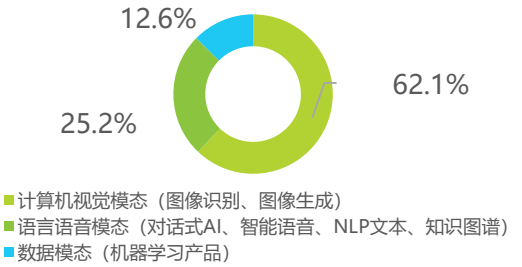
1）存量逻辑：大模型架构对现有人工智能产业带来重构加成，更多小模型方案被大模型产品替代，实现产业存量扩张。

2）增量逻辑：大模型架构为人工智能产业应用带来更多落地可能与场景机会，带动更多生成式AI应用的产业增量。艾瑞测算增量规模仅考虑现有产业与技术架构的带动加成，未来AI技术与VR、泛互联网、游戏等产业空间的TAM规模将更具想象空间。

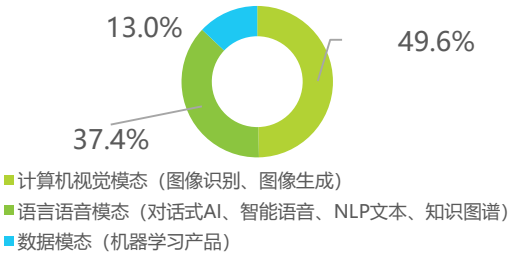
2020-2028年中国人工智能产业规模



2023年中国人工智能产业模态分布



2028年中国人工智能产业模态分布



注：中国人工智能产业规模口径包括中国AI芯片市场规模、AI基础数据服务规模、计算机视觉市场规模、对话式AI与智能语音市场规模、NLP市场规模、知识图谱市场规模。

来源：艾瑞咨询研究院根据专家访谈、桌面研究自主研究绘制。

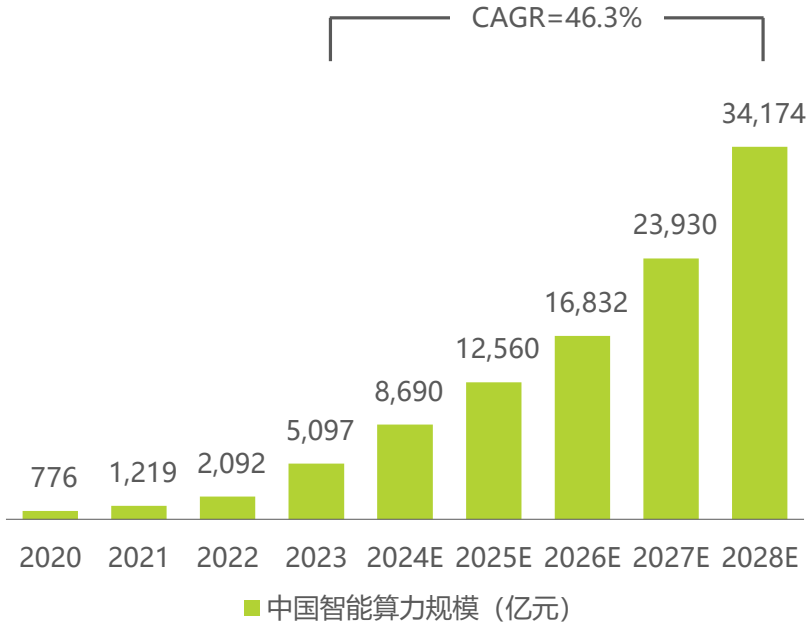
©2024.4 iResearch Inc.

中国AI产业基础设施规模

智能算力产业规模高速增长，上层模型应用需求带动AI基础数据服务市场

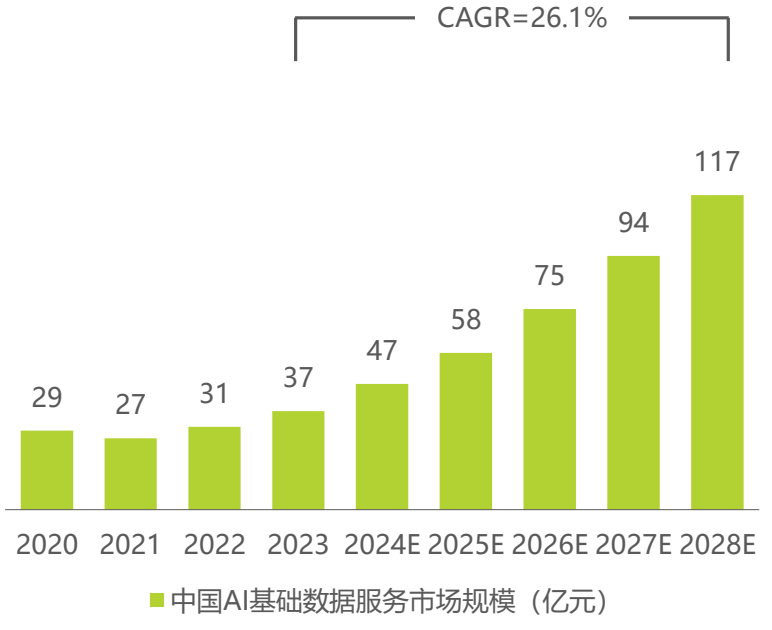
根据艾瑞咨询研究院测算，2023年中国智能算力市场规模已达到5097亿元。在大模型训推需求影响下，2023年中国智算市场规模相较于2022年完成大比例跃升，一方面AI芯片的单卡算力呈倍数增加，另一方面，以训练场景为主的服务器载卡数量从原来的2-4卡逐步升级到4-8卡的普遍配置。随着中国各地智算产业的投入建设、大模型在边缘侧及端侧的算力释放，2028年，中国智能算力市场规模或将达到3.4万亿元，五年复合增长率达到46.3%。2023年的中国AI基础数据服务市场规模为37亿元，由上层大模型应用带来的数据需求正改变着AI基础数据服务的工作结构。以传统NLP任务为主的分词、词性标注等工作被慢慢取代，且数据服务厂商更加拥抱融合大模型范式的全流程、自动化工作流，将业务重心开拓到大模型训练数据集、RLHF微调、提示词生产、偏见数据库、评测服务等采集生产工作中去，2028年中国AI基础数据服务市场规模将达到117亿元，五年复合增长率达到26.1%。

2020-2028年中国智能算力市场规模（金额口径）



来源：艾瑞咨询研究院自主研究绘制。

2020-2028年中国AI基础数据服务市场规模



来源：艾瑞咨询研究院自主研究绘制。

中国AI产业基础设施呈现倒三角特征

国家直面AI产业发展的“中国式困境”

中国人工智能产业发展呈现“倒三角”特征，数据层的规模质量、算力层的规模性能制约着模型层及应用层的飞速发展。2023年3月，中共中央、国务院印发的《党和国家机构改革方案》对外公布，组建国家数据局。2023年10月25日，国家数据局挂牌成立。国家从政策监管角度积极引导数据要素市场建设，鼓励开展数据确权授权，构建数据流通体系，旨在为大模型训推提供高质量数据市场资源。另外，Sora模型的横空出现点爆产学研界对于视频生成的研究热情，而对图像、视频、多模态模型的训推将指数级加大对AI算力的底层需求。国家及地方政府鼓励并积极开展智算设施建设，同时以“算力券”等形式降低企业的训练成本、提高算力对接效率，更多支持中小企业购买算力服务。

中国AI产业发展洞察

完善AI技术底座 & 打造AI应用生态

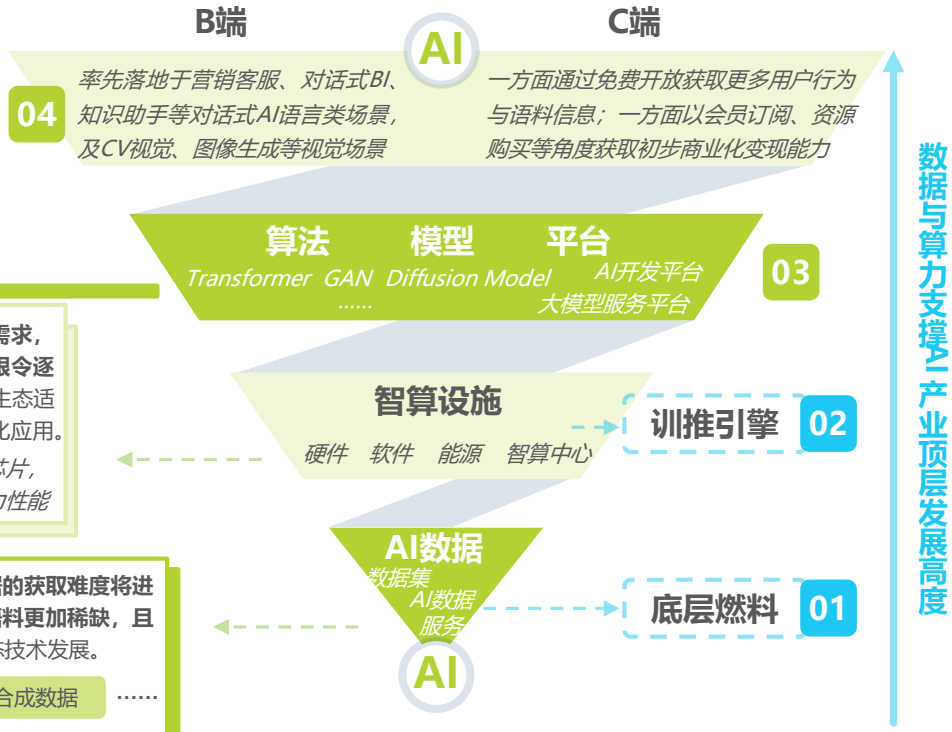
- **应用层**：受底层资源与模型技术等限制，**预训练大模型带来的产业变革尚未带来颠覆性生态改变。大模型仍然缺乏透明度与可解释性，且在内容生成角度缺乏一致性**，是未来大模型应用实现规模化联动应用需要解决的核心问题之一。
- **模型层**：**无论从大模型 v.s. 小模型，还是决策式AI v.s. 生成式模型角度出发，未来一段时间内都将处于并存状态。**从场景需求、业务适配、性价比等角度出发，选择对应模型。

大模型应用对训推算力需求持续加大。此外，**多模态数据将进一步抬升智算需求，**未来中国将持续智算中心建设弥补产业算力缺口。**随着美国对中国的半导体限令逐步收紧。中国企业普遍走上以NV卡为存量增量，拥抱AI芯片的国产化道路**，生态适配历经阵痛期，以华为昇腾、百度昆仑、海光等为代表的芯片产品陆续规模化应用。

中美限令导致版本更迭： NVIDIA 以英伟达为例 A100 H100 → A800 H800 → H20 L20 L2 *中国特供版芯片，持续阉割算力性能*

当下**高质量中文语料资源仍然处于短缺状态，且随着时间推移，优质数据的获取难度将进一步加大。**从数据维度来看，除了文本语料外，**图像、视频方向的数据语料更加稀缺，且亟需发展是带有时间和空间维度信息的3D数据，**助力AI视频生成及多模态技术发展。

自有数据集 开源数据集 商务授权 数据服务 合成数据 ……



来源：艾瑞咨询研究院自主研究绘制。

中国注重数据资源能力提升

AI赋能下，数据生产能力将逐步提升，数据标准及生态将逐步建立

根据IDC预测，到2025年中国有望成为全球最大的数据圈。从国内AI基础数据供需角度，当前大模型的迅速发展大大提升了对AI训练数据质量、数量和生产效率的要求，传统以人工标注为主的数据生产从成本和效率上都难以满足需要。一方面，AI训练数据生产正在向AI数据标注、全流程自动化方向演进，同时在部分仍需人工标注的场景，配备的人员素质也有了明显提升。另一方面，AI数据的供需问题也使得AI数据合规方面的问题更加突出，我国正在通过行业协会、研究院的力量，推动AI数据标准建立，鼓励开源数据集发展。

中国AI基础数据工程发展现状及趋势

数据质量标准、更新速度提升，数据生产自动化

需要更高效快捷的数据准备 ← 更新迭代速度加快

需要更高质量的训练和微调数据 ← 任务更加复杂精细

大模型

数据质量和生产效率的高要求导致数据成本增加，一方面，企业开始尝试使用AI进行数据标注，并取得了良好的效果，另一方面，数据供应商也在进行数据工具链的进化和完善，推动数据工程向标准化和自动化发展

AI数据标注

- 特斯拉使用大模型进行数据标注，1万个60秒内的视频，大模型只需要运行一周，而人工标注需要数月。
- 理想汽车训练大模型进行自动化标定，实现人工标注1000倍效率

自动化数据工程平台

数据采集 → 数据标注 → 数据管理

数据结果调优　实体关系命名

数据排序　数据分类

.....　.....

数据量遇瓶颈，合成数据成为有力补充

大模型 → **B端：**中型以上企业均大概率部署企业专属模型，需大量垂直领域数据

大模型 → **C端：**大量C端应用将使用大模型作为支撑，需要海量用户数据

B端、C端 → 高质量的真实数据已不能满足大模型规模和数量增长

大模型训练数据通常来自企业自有数据、网络爬取数据、外部付费/开源数据集以及AI合成数据。大模型训练和微调所需数据量快速增长，真实世界数据将在数年内被用尽，根据Gartner预测，到2024年，用于训练AI的数据中有60%将是合成数据

合成数据优势

可定向提升模型某方面能力	可避免数据隐私纠纷
可降低模型安全风险	可降低模型训练成本

数据供给能力与合规问题推动相关标准及开源数据集的发展

当前数据合规存在严重风险与漏洞

对于国内大模型厂商而言，当前主要使用的各类数据都存在一定的合规风险。如互联网公司收集用户数据，或互联网爬取数据，甚至调用其他厂商大模型大量生产合成数据用于自家模型训练，已产生数起纠纷。AI数据需要更加开放合规的生态建设

数据标准与规范的建立

建立AI数据标准对于提升数据质量、标准化及后续监管和权责划分都有重要意义。欧盟于2023年5月发布促进人工智能标准化相关文件，其中涉及机器学习数据质量管理要求和分析过程框架。信通院也着手编写《人工智能数据集质量管理能力评估方法》

开源数据集

通过开源开放的生态，有利于带动高质量数据集的利用效率，缓解数据资源紧缺问题。当前国内数据开源意识不足，上海人工智能实验室、浪潮云、蚂蚁集团等已经推出不同领域的开源数据集，并被多个大模型采用。

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

©2024.4 iResearch Inc.

中国智算中心发展多维关注点

软硬耦合、规模互联、能源提供、清洁环保、集约高效

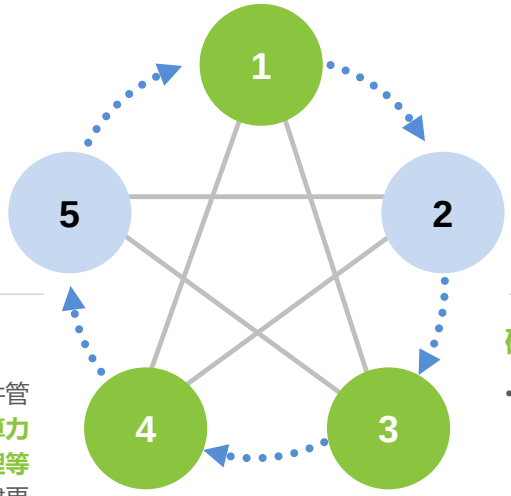
算力是集信息计算力、网络运载力、数据存储力于一体的新型生产力，作为新型信息基础设施的重要组成部分，算力基础设施发展呈现多元泛在、智能敏捷、安全可靠、绿色低碳等特征。在IDC时代，以CPU为主的计算任务场景相对单一或者标准化。而在智算中心时代，以GPU并行计算及AI加速卡为主的异构计算任务变得更加多元多样，面向安防园区、自动驾驶、互联网推荐、人机交互等不同下游场景提供智算资源及训推服务应用。所以艾瑞认为，中国智算中心的发展建设，不仅仅要注重硬件性能、网络带宽、算力规模等硬实力特征，更是要注重与智算中心的硬件设施能力融合，结合产业需求及地域业务，以低成本、高效能、定制适配的软件平台完成客户场景需求的软实力发展。

中国智算中心发展关键要素

*1）**发展散热液冷技术,优化中心运维能力**；2）加强AI算力集群建设，**面对不同AI加速卡的异构体系，提供高适配强扩展的算力集群支撑**，由粗放式扩张到精细化管理。*

能源能耗

- “限制AI发展的将是电力和降压变压器的短缺。” “AI的尽头是光伏和储能。”
- **国外超大规模集群的算力供需方都越发关注大模型在能源方面的需求及能耗问题。**寻找高效清洁能源的获取，同样也在注意节能环保技术与设计，通过软件平台管理更好的优化PUE能效等指标。



网络互联

- 随着数据量与计算量飞涨，数据中心需**优化网络带宽、计算总线协议，实现数据在节点内与节点间的高吞吐低延迟的传输与连接，并进一步优化计算集群的架构与设计**，保证数据中心的高效利用率，打造高带宽、高吞吐、低延迟、自动化的新型智算中心网络设施。

软件平台

- 搭建智算中心软件平台，通过平台操作系统的软件管理，优化AI算力供给，**为下游客户侧提供大规模算力资源的资源纳管、算力调度、优化监控、运营管理等智算服务，并有望进一步屏蔽底层硬件差异**，构建更完整全栈的智算应用软件生态。

硬件能力

- 受中美关系影响，国产芯片实现自主创新迫在眉睫，**中国算力层也会进一步尝试脱离对头部厂商英伟达的依赖，以“云巨头自研自用+独立/创业公司服务于信创、运营商等To G与To B市场”**为两条主线发展，实现国产“算力+应用”的正循环

来源：艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

数实产业期待AI驱动的原生生态

AI仍非无所不能，AI原生应用重塑人类生产生活的质变节点尚未到来

生成式AI产品的出现，让人工智能这一关键词重回人类关注焦点。基于扩散模型的AI电商设计图、基于大语言模型的对话式AI解决方案等热门B端产品率先出圈变现，C端应用也涌现了一批AI搜索对话、图像生成等APP产品。2023年7月，妙鸭相机小程序横空出世，成为AIGC时代下的第一个C端爆款。人们沉浸于生成式AI技术带来的巨大改变，同时也对时代技术的飞速发展产生FOMO情绪，恐惧因新技术的诞生而失去现有平台产品或错失机会。然而与第一轮计算机视觉带动的人工智能浪潮类似的是，AI的技术跃升并未达到无所不能的地步，需在用户预期上做合理规划引导，填补由于市场宣传与落地应用带来的认知差距。互联网时代下，诞生以阿里为代表电商厂商、以腾讯为代表的通讯厂商；长短视频时代下，诞生以爱优腾为代表的长视频厂商、以抖音快手为代表的短视频厂商；社区经济时代下，诞生以小红书为代表的平台厂商。人工智能时代下，以AI驱动的原生生态尚未出现，AI原生应用带来的B端生产重构，及C端的流量洗牌值得期待。

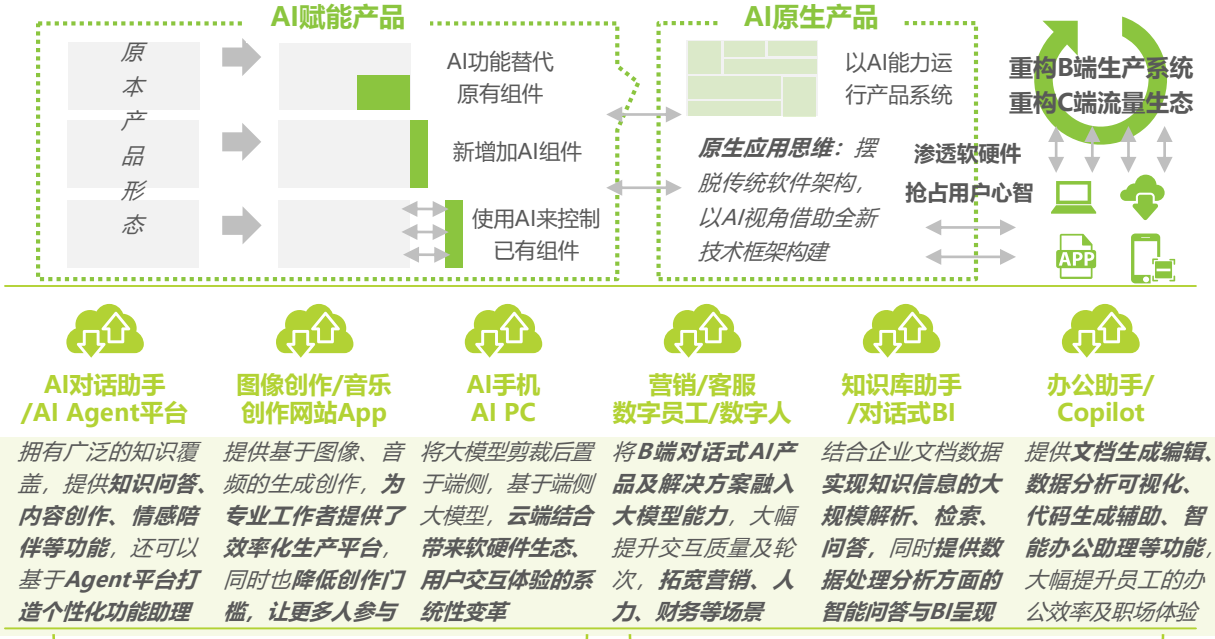
2023年中国AI热门产品及演进形态分析

AI带来的恐慌焦虑

AI飞速发展带来FOMO情绪，即Fear of missing out，可被译为错失恐惧症。人们普遍恐惧新技术的诞生会淘汰现有的产品平台，从而引发恐惧感。

AI带来的新一轮认知GAP

生成式AI的技术变革，在媒体、资本、厂商的影响宣传下，仿佛被打造成“无所不能”的形象，然而从B端和C端实际落脚点出发，AI产品方案落地仍需严谨方法论及场景结合。



来源：艾瑞咨询研究院根据专家访谈、公开资料自主研究绘制。

©2024.4 iResearch Inc.

理性探寻通往AGI产品之路

五年之内AGI成为现实？未可望未可及，量化AGI的实现过程提高预见性

产学研界对AGI的讨论热度持续不减，但艾瑞认为，技术奇点到来的一刻具备不可预见性，当下关于AGI的时点预测更多是思考大于实际。然而AI的确在朝着范围愈发广泛的普遍性发展，且欲实现接近并超越人类的普遍性，其演进节点是对全人类世界运行逻辑的挑战更新，对全球劳动力结构、产业经济发展、道德伦理制度、社会规则运行，以及技术经济优势带来的的地缘政治以及国际军事都将带来底层逻辑的巨大变数。需量化人工智能的通用性与自主性，理解并定位AGI的演进道路节点，提升AGI实现过程的可预见性及社会稳健性。另外，随着AI能力的通用化提升，具身智能实现了成本更低廉、能力更泛化的产品形态，国内外企业也纷纷加大人形机器人等产品投入，寻路实体社会的“AGI”世界。

AGI的发展路径探讨


马斯克：“到2025年，AI可能比任何人类都聪明。到2029年，AI可能比所有人类加起来还要聪明。”


黄仁勋：“AI会在5年内通过人类测试，未来10年算力将再提高100万倍。”


奥尔特曼：“AGI将在5年内实现。”

Alex Irpan：“到2025年，AGI就有10%的概率出现。”

Logan.GPT：“超级AI，将在10年内出现。”

AI角色：AI工具、AI顾问、AI智能体、AI协作者、AI专家

Google DeepMind 论文划分的AGI等级

	专用（Narrow）	AGI角色	通用（General）
Level 0：无AI	计算机软件；编译器	无AI	人机回圈计算，例：亚马逊土耳其机器人
等于或略优于非熟练的人类 Level 1：新兴	GOFAI，简单的基于规则系统。例：SHRDLU	AI工具	新兴AGI：ChatGPT，Llama 2，Bard AGI雏形，当下AGI节点
至少50%的熟练成人 Level 2：熟练	恶意评估检测器，如Jigsaw；智能扬声器，如Siri，Alexa或Google助手；VQA系统，如PaLI，Watson；SOTA LLMs的子任务（例如，短文写作、简单编码等）	AI顾问	熟练AGI：尚未达到
至少90%的熟练成人 Level 3：专业	拼写与语法检查，如Grammarly；图像生成如Imagen或Dall-E 2	AI协作者	专业AGI：尚未达到
至少99%的熟练成人 Level 4：大师	Deep Blue，AlphaGo	AI专家	大师AGI：尚未达到
优于100%人类 Level 5：超越人类	AlphaFold，AlphaZero，StockFish	AI智能体	超越人类AGI/ASI：尚未达到

将AI能力赋予实体

关注具身智能发展

具身智能与AGI：无处不在 硅基时代 软硬能力的究极结合 无所不能

具身智能 = **具身** + **智能**

- 具身：对真实世界中物理本体的控制，具备感知、计算执行能力的硬件
- 智能：具备认识世界、理解世界、主动影响世界、持续学习迭代的能力

基于预训练大模型的分层实现：为机器人赋予强大能力，提升任务泛化性，让机器人可以更好理解指令，理解知识世界，对应执行完成任务或做出更优质回复

◆ **人形机器人，具身智能的究极形态？**

未来，具身智能在不同需求场景下会有各自成本效益适配的硬件载体，长远来看，**人形作为与人最适配的外在装置，在融入替代人类生产生活的基建设施搭配上具备更多泛化优势。**

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

©2024.4 iResearch Inc.

02/ 中国人工智能产业征程

AI - Ongoing

2023，全球的生成式AI元年
—— AIGC产业洞察

◆ 时代背景

① 全球进入AI驱动的生产革命，生成式技术是时代际遇。
② 中美在生成式AI产业展开科技竞争，全栈组合拳拉锯发展。

◆ 技术变革

① Transformer架构优化模型泛化的训推能力与理解生成的内容能力。
② 文本模态达高应用成熟度，代码、语音、图像具备商业化基础。

◆ 商业应用

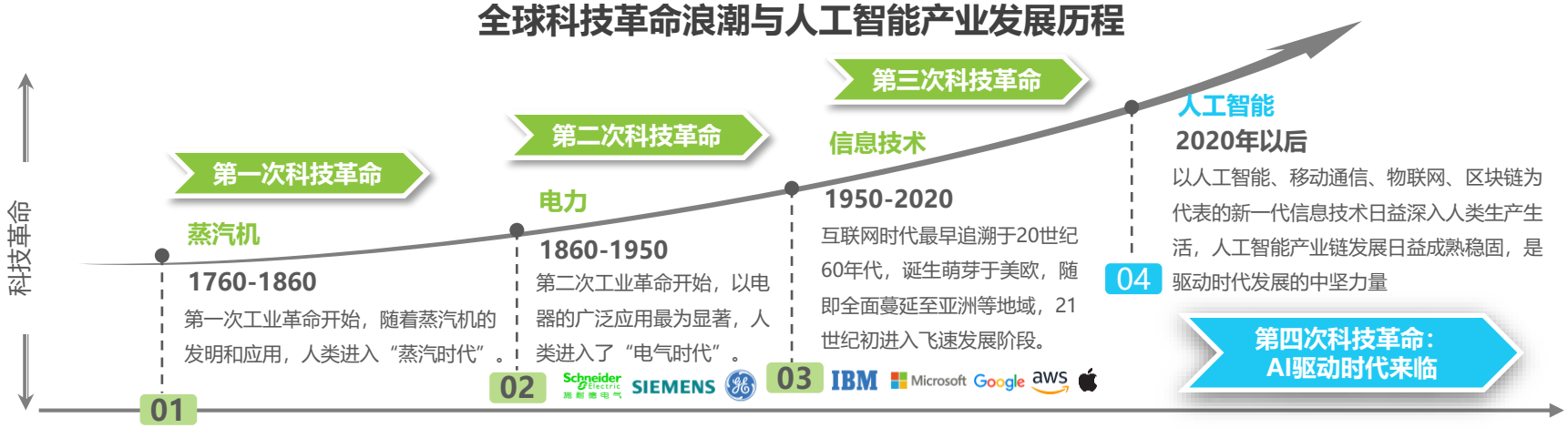
① 国家对大模型上线监管采取“备案制”，40+家大模型持“证”上岗。
② B端：B端场景出发需逐步渗透打磨，打通业务逻辑实现更多场景的落地应用闭环，呈延续性曲线融合赋能。当下B端产品方案融合更多大模型技术，以API调用、SaaS产品及定制方案等方式加速企业智能化赋能。
③ C端：C端场景应用需从供给侧满足硬件设备条件及大模型能力适配，在软硬件生态成熟后涌现阶梯式能量爆发。当下C端商业模式普遍以免费为底提供会员订阅与资源购买。

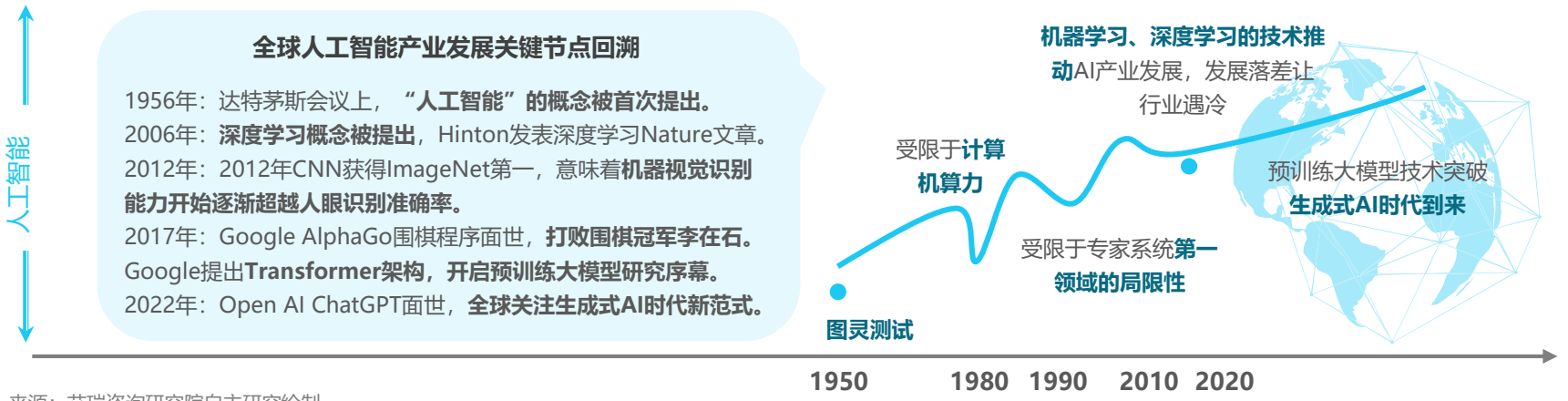
2023年迎来生成式AI元年

全球进入AI驱动的生产革命，生成式技术是时代际遇

人工智能伴生于信息技术时代，经过数十年的研究积累及经验沉淀，已逐步跨越科学与应用之间的技术鸿沟，迎来新一轮的红利爆发与创新机遇。21世纪以来，全球技术创新进入空前活跃期，生成式AI技术的到来被誉为“最具革命性”的技术进步，未来产业发展是抢占全球创新高地、重构全球创新版图、重塑全球经济结构的关键节点。

全球科技革命浪潮与人工智能产业发展历程





来源：艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

生成式AI产业发展定位

大模型做底，生成式AI与决策式AI共筑产业发展

来源：艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

生成式AI产业厂商占位

中美在生成式AI产业展开科技竞争，全栈组合拳拉锯发展

根据《全球人工智能创新指数报告》，美国的人工智能创新指数已连续四年位居全球第一，中国连续三年保持全球第二水平，均位于人工智能产业发展的第一梯队，随后为英国、德国、新加坡、加拿大等国家，整体来看欧洲大多数国家位于第二梯队；印度以23名位于第三梯队。因此本页产业洞察选取美国、中国、欧洲、印度四个区域为代表展开梳理分析。

全球生成式AI产业洞察

美国

起步早，重视人工智能技术发展，**走在生成式AI产业浪潮前列，全球范围内占位基础层以英伟达为代表与模型层以Open AI为代表的头部厂商**，不断丰富工具层及应用层的落地发展。

中国

着重技术创新追赶，**由下到上打造从基础层到应用层的全栈自主能力**，应用层及工具层生态活跃，**模型层及基础层的技术能力相较于美国仍有代际差距。**

欧洲

欧洲地域分散，基座模型研发集中在英国、法国等国家。总体来看，在AI大模型方面，**欧洲或更多扮演一个应用者角色**，即通过接入各国大模型基座的API能力来开发应用。**法国、德国、英国等国家在支出和采用方面处于领先地位**

印度

印度对生成式AI充满期待，**人工智能相关课程需求及社区开发者数量大幅增长**，与海外英伟达等厂商展开合作建设算力资源及基础设施，**重点发力模型层与应用层产业发展。**

资源限制
渠道限制
头部竞争关系
基础资源提供
基础资源提供
资金提供

应用层	Microsoft, amazon, salesforce, Meta, Adobe, Chegg	小红书, 阿里云, 金山办公, 京东云, 百融云创, wondershare 万兴科技	DeepL, quantexa, Contentsquare, Pixis	haptik, BLEND, INMOBI
工具层	OpenAI, aws, FIXIE AI	阿里云, 百度智能云, 澜码科技, 火山引擎, 京东云, DataCanvas 九章云极	DUST AI	Webkul, HEXO, Arya.ai
模型型	OpenAI, Google, Meta, ANTHROP\C, Pika Labs, Midjourney, runway	阿里云, 百度智能云, 华为云, 腾讯云, 京东云, 百川智能 BAICHUAN AI, MINIMAX	stability.ai, IIElevenLabs, Synthesia, MISTRAL AI_, ALEPH ALPHA	Krutrim SI Designs, sarvam.ai 印度信实工业和印度顶尖工程院校共同组建的研究联盟 BharatGPT
芯片层	NVIDIA, intel, QUALCOMM, AMD, SambaNova SYSTEMS	NVIDIA, HISILICON, Cambricon 寒武纪, Enflame 燧原科技, 天数智芯 Iluvatar CoreX, 壁仞科技 BIREN TECHNOLOGY	NVIDIA, GRAPHCORE 2023年已大量退出中国业务，正寻求出售资方	AI算力资源：与英伟达 NVIDIA 达成战略合作，购买数亿美元订单，共建印度智算基础设施建设

来源：艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

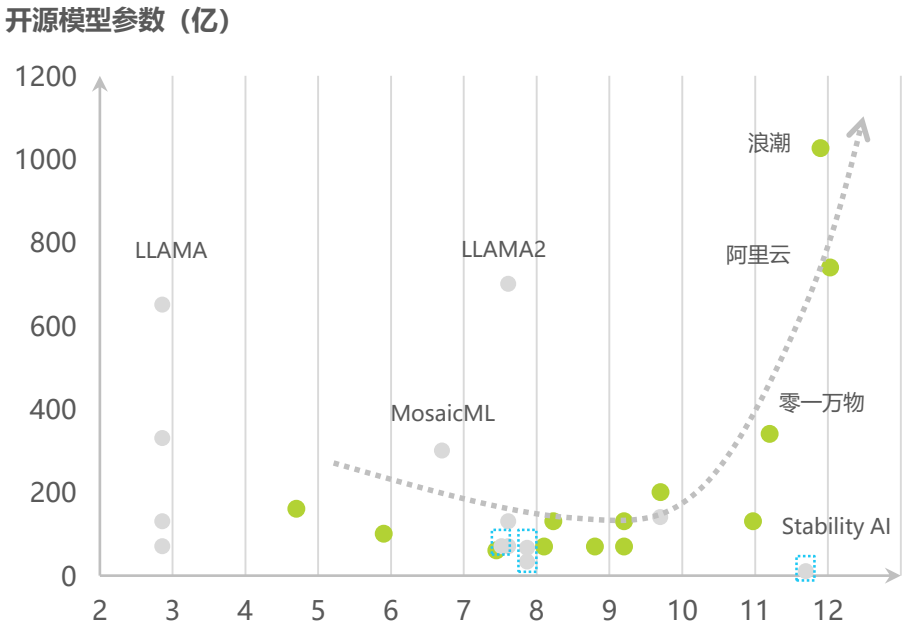
全球开源量级及参与度陆续提升

开源模型生态不断丰富，模型厂商在开闭源路径下适应调整市场策略

2023年，海内外厂商在开源模型的参与度不断提升，Meta陆续开源Llama系列模型，Google于2024年初开源Gemma模型。更多高质量开源模型避免了生成式AI产业“重复造轮子”的基础设施问题，基于开源模型的定制微调成为众多厂商低成本试错、高质效落地的商业化手段。由此，通用基础大模型厂商从0到1的训练投入动作变得更为谨慎，在通用基础大模型的巨额训练成本及模型产出结果的投入产出比考虑基础上，一些厂商开始重新定位市场策略，避免置于头部闭源模型与优质开源模型的中间尴尬地位。此外，模型的开源参数量级不断提升，进一步优化模型输出效果，但模型参数同样也存在一定边界效应临界点，过大参数量级反而增加模型定制微调成本，影响到开源应用的性价比。

2023年全球厂商开源动态洞察

2023年全球厂商开源动作部分列举散点图



注：绿色为中国厂商开源动作；灰色为海外厂商开源动作。

来源：艾瑞咨询研究院自主研究绘制。

国内大语言模型的参与度及开源参数不断提升

- 国内高校，**以复旦、清华为代表的高校率先开启中国的开源模型浪潮**。
- **从2023年7月起，中国进入开源模型密集建设期**，诸多厂商，如阿里云、百川智能、面壁智能、零一万物、昆仑万维等公司加入开源社区，**且开源参数不断提升。**

图像模型参数以低量参数开源，视频技术日益成熟

- **以Meta、Stability AI为代表的海外企业**积极建设图像生成模型的开源生态，陆续开源文生图、视频基础模型。相较于语言模型，**图像参数模型开源一般在数十亿量级以内**，参数量小，便于部署于本地及做定制化开发。

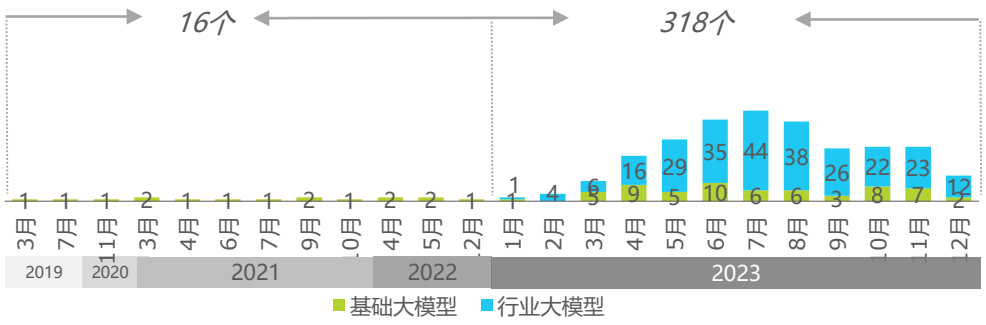
©2024.4 iResearch Inc.

中国通用与垂类大模型落地声量加大

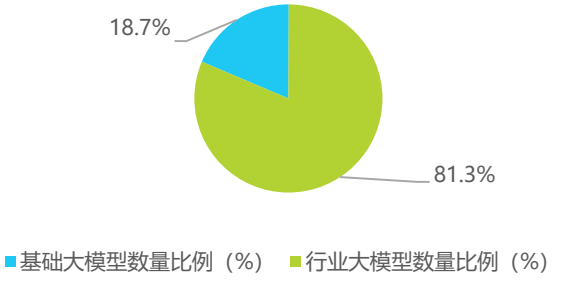
2023年行业大模型进入爆发期，医疗、金融及科研教育为集中落地领域

2017年起，国内互联网巨头厂商相继投身预训练大模型的产品研发，百度、华为等产业基因强的巨头厂商已在2022年左右发布过行业大模型矩阵，探索金融、制造、能源、媒体等场景落地。2022年末，ChatGPT掀起生成式AI浪潮，预训练大模型再度成为AI产业焦点。2023年，“大模型热”愈演愈烈。其中，国内互联网厂商持续更新迭代技术底座及模型能力，更多高校与大模型创企加入。2023年5月起，行业大模型发布数量显著增加，互联网巨头达成进一步行业分化及产业伙伴合作，同样受益于开源生态的建设，更多垂类厂商结合开源模型研发契合自身业务的行业大模型产品。根据艾瑞不完全统计，截止2023年底，中国行业大模型的个数占比已经超过8成。以医疗与金融为首要落地领域，分别占比达到21.9%与12.8%。

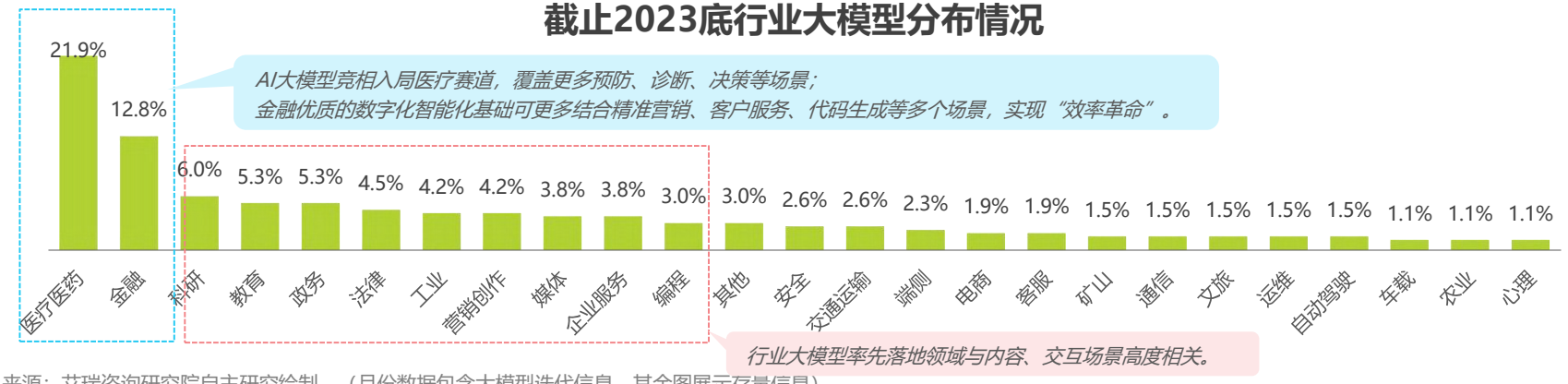
2019-2023年大模型发布数按月份分布情况



截止2023年底，中国通用基础大模型与行业大模型分布情况（%）



截止2023底行业大模型分布情况



来源：艾瑞咨询研究院自主研究绘制。（月份数据包含大模型迭代信息，其余图展示存量信息）

中国巨头厂商补全生成式AI产业技术栈

大模型服务平台设施日益完善，整合工具链条支撑生成式AI技术落地

通用基础大模型落地因需求差异展开产业路径分化，在面对业务复杂的垂域环境时，需进一步结合业务数据定制微调成行业级、企业级大模型支撑上层应用。此外，随着模型开源生态的繁荣，开发者及中小型企业等客户更需要具备生成式AI生产及应用全流程开发工具链的平台设施，为其提供数据集服务、模型工具链、微调SFT、模型能力评估等服务。大模型服务平台与底层云资源、算力资源及数据资源呈强绑定关系，国内巨头纷纷搭建平台设施进一步补全生态技术栈。如今，模型能力评估仍是需求侧痛点，虽然各家平台提供一定模型能力评估工具，但是业内缺少统一权威量化标准，如何选择成本能力适配的模型产品是需求侧选型的核心痛点。

解构大模型服务平台

1 百度-千帆大模型平台

千帆AI原生应用工作台

模型广场	通用大模型 行业大模型	百度文心大模型	第三方大模型
大模型工具链	数据管理	模型调优	模型评估&优化
	推理服务部署	Prompt工具	插件库

***2023年3月**，百度发布千帆大模型平台，**9月**再度发布千帆大模型平台2.0版本，覆盖金融、制造、能源、政务、交通等行业的400余个场景。*

2 字节-方舟大模型平台

场景应用	泛互联网 金融 汽车 大消费 .....
业务支持	模型提供方 行业ISV 火山引擎交付团队
产品功能	训练推理 数据干预 展示体验
模型能力	语言大模型 视觉大模型 语音大模型
基础设施	机器学习平台 云服务器 .....

***2023年6月**，火山引擎推出大模型服务平台火山方舟，面向企业提供模型精调、评测、推理等平台服务。*

3 京东-言犀大模型平台

数据准备 → 模型定制 → 模型评测 → 模型部署

AI资产	模型	数据集	镜像
资源管理			
Frameworks			
集成云产品	云主机	高性能计算引擎	CPU GPU .....

***2023年7月**，京东推出京东言犀大模型，言犀AI开发计算平台，能够为客户的大模型开发和行业应用开发，提供定制化解决方案。*

4 阿里-百炼大模型平台

生态支持	模型工具	应用范式	安全服务
模型及应用广场	预训练/微调	预制插件	数据安全屋
开发者工具	评测工具	大模型应用编排	内置安全
开发者服务	多行业模型	微应用服务	安全工具

行业大模型 通义大模型 三方大模型

***2023年10月**，阿里发布百炼大模型服务平台，提供模型选型、微调训练、安全套件、模型部署等全链路服务和应用开发工具.*

来源：艾瑞咨询研究院自主研究绘制。

全球生成式AI产品模态日益成熟

文本模态达高应用成熟度，代码、语音、图像具备商业化基础

目前生成式大语言模型已经成为当前NLP能力的主要技术依托，以文本为核心的大语言模型率先落地于生产办公、客服营销、内容创作等应用场景，未来将继续攻关内容可控性、模型可解释性、数据隐私安全等技术问题。此外，代码生成、语音生成、图像生成的技术成熟度排序呈现从高到低发展，且日益完善提升，具备商业化技术基础。而视频生成技术相较于图像生成技术仍有1-2年的代际差距。OpenAI于2024年2月发布的Sora模型为视频生成领域带来标杆性产品，短时间内可为文娱、影视、内容创作等领域带来新生产力，长远来看AI理解和模拟运动中的物理世界之后，可进一步解决现实世界的模拟运行、数据训练及理解交互问题，赋能工业应用、自动驾驶、数字孪生等领域。而Sora模型仍是以视觉生成为核心，未来产学研界也将更期待以文本为桥梁，融合语音、视觉能力，具备更深层次推理能力的多模态模型出现，离通往AGI的道路更进一步。

AIGC的生成路径与对应场景

	文本生成	代码生成	图像生成	语音生成	视频生成	3D模型生成	其他生成
技术模型	Open AI GPT 4	Open AI GPT模型	Stability AI Stable Diffusion	Open AI Jukebox	Open AI Sora模型	Google Dream Fusion	策略生成
	Meta Llama2	Meta Code Llama	Open AI DALL-E	Meta VoiceBox	Runway Gen -2	NVIDIA GET3D	科学计算
	阿里 通义千问	Stability AI Stablecode	阿里 通义万相	Minimax abab-speech-01	Pika Labs Pika 1.0	Meshy AI Meshy-2	研究研发
技术成熟度	单模态 — 文本 多模态	在文本大模型基础上基于代码库做定向调优训练，**成熟度高**	以扩散模型SD为开源主流，**图像生成技术日益成熟**，细节不断优化	大模型提升语音识别及语音生成能力，**缓解方言、小语种等问题**	视频生成技术**在帧率、连续性、可控性、长度等方面仍需提升**	3D资产质量**在内容、可控性仍待提升，需优化耗时、成本问题**	策略生成**依赖多模态能力、判别式与生成式AI融合应用**
应用场景	个人助理	代码生成	设计工具	语音合成	视频创作	3D建模	决策建议
	文案助手	数据库管理工具	电商商品图	语音助手	社交娱乐	数字人生成	自动驾驶
	客服营销	编程助手	社交娱乐	音乐创作	空间理解计算	数字孪生模型	工业模型

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

©2024.4 iResearch Inc.

中国生成式AI应用商业步伐稳健

40+家大模型持“证”上岗，监管与需求对大模型厂商提出更高要求

2023年8月15日，《生成式人工智能服务管理暂行办法》 正式施行，国家对大模型上线监管采取“备案制”，确保内容安全和意识形态准确性。之前大模型产品多采用邀请注册制，仅开放给企业伙伴或有限名额使用。通过“备案制”的企业则可以正式面向公众开放注册、提供服务，相较于“审核制”，“备案制”监管为生成式人工智能技术的产业发展提供良好培育环境，以动态角度应对管理技术发展的不确定性。2023年8月31日，第一批AI大模型企业获批。截止2024年1月，中国已获备案审批大模型40余家，其中通用和垂直大模型的占比分别为55.8%与44.2%。

大模型备案流程及要求

2023年8月15日正式施行的《生成式人工智能服务管理暂行办法》，要求所有想“持证上岗”的大模型企业的生成式AI产品，应按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续。 即：AI大模型的厂商，在互联网信息服务算法备案之外，还需进行大模型备案。

算法备案

*以企业邮箱注册“互联网信息服务算法备案系统”，分**三个步骤**完成深度合成备案填报：*

1）填报主体信息

2）填报算法信息

3）关联产品及功能信息或填报技术服务方式

大模型备案

- 安全评估报告
- 拦截关键词列表
- 应急处置预案和记录材料
- 个人信息安全保护制度
- 网络安全管理制度及操作规程
- 评估测试题库拒答测试题库
- 用户管理制度
- 安全培训制度
- 模型服务协议
- 语料标准规则
- 用户投诉举报处理制度
- 网络安全负责人任命书

中国生成式大模型产品名单

2024.4.31 第一批

公司	模型
百度	文心一言
百川智能	百川
抖音	云雀
智谱AI	智谱清言
商汤科技	日日新SenseNova
中科院自动化研究所	紫东太初
Minimax	ABAB
上海人工智能实验室	书生
阿里云	通义千问

2023.11.4 第二批

公司	模型
网易有道	子曰
面壁智能	面壁露卡Luca
出门问问	序列猴子
昆仑万维	天工
美团	未公布
知乎	知海图AI
月之暗面	Moonshot
金山办公	WPS AI
好未来	MathGPT
360	奇元
蚂蚁集团	百灵

2023.12.28 第三批

公司	模型
京东	言犀
云知声	山海
聆心智能	CharacterGLM
澜舟科技	孟子GPT
中科闻歌	雅意
深言科技	语鲸
抖音	福禄瓜
快手	快意
红棉小冰科技	小冰

2024.1 第四批

公司	模型
小米	小爱同学
第四范式	式说
零一万物	Yi
创思远达	魔方
步刻科技	微步情报智脑XGPT
BOSS直聘	南北阁
衔远科技	品商
衔远科技	摹小仙
新壹科技	新壹视频
识因智能	一叶轻舟
什么值得买	AI问答机器人
脉脉	智能问答
智联招聘	AI改简历
掌阅	阅爱聊

来源：综合微软研究院的《Sparks of Artificial General Intelligence》等公开资料研究绘制。

©2024.4 iResearch Inc.

全球生成式AI应用场景探讨

发力B端还是C端？企业出海挑战与机遇并存

场景产品侧作为预训练大模型及生成式AI技术的核心落点，B端与C端的市场发展成为产业界关注的焦点。从B端场景出发，大模型与企业业务离不开经营数据，需逐步渗透打磨，打通业务逻辑实现更多场景的落地应用闭环，呈延续性曲线融合赋能。因此To B企业需打造足够垂类企业合作基础或良好产业合作生态满足前期场景探索；从C端场景出发，大模型落地应用需从供给侧满足硬件设备条件及大模型能力适配，在软硬件生态成熟后涌现阶梯式能量爆发，C端企业需准备足够资金储备及短期变现方式以应对前期的发展沉寂期。

中国生成式AI落地场景多维分析

To B & 国内

- 国内B端企业**在B端SaaS软件付费意愿及支付能力有待提升**；基于行业属性、客户数据隐私等考量，**更多企业会要求定制及私有化部署**，关注“投入成本高、规模化变现难”AI盈利怪圈的风险

To B & 出海

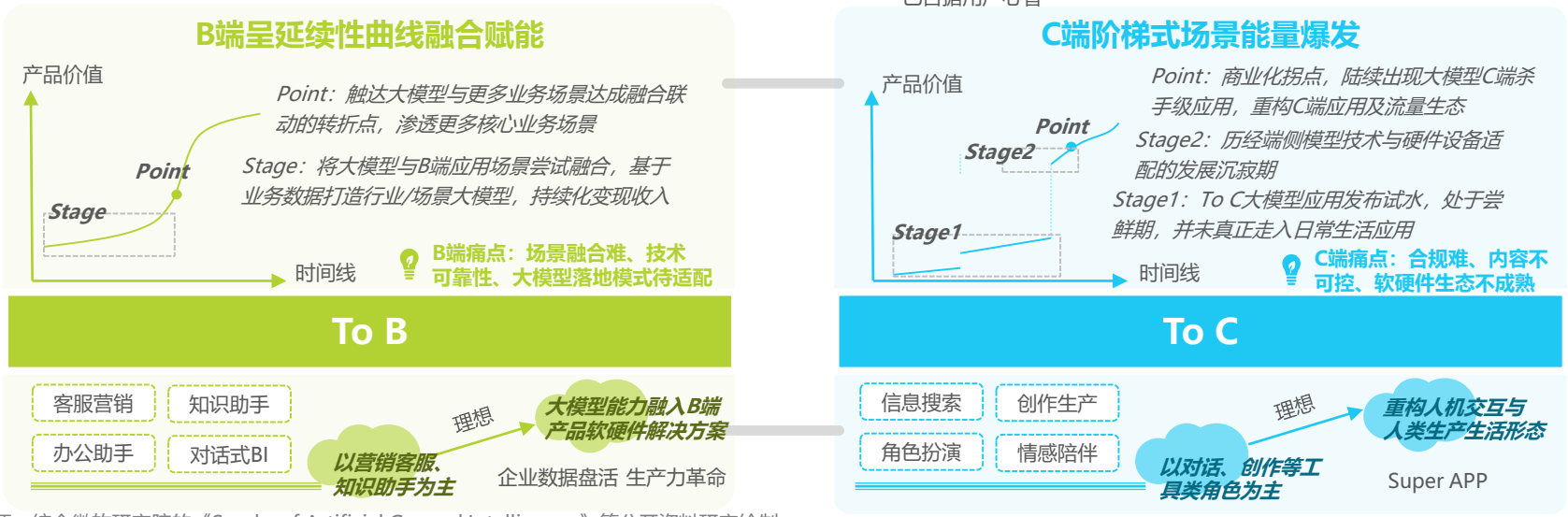
- 海外B端企业**具备更高付费意愿及支付能力**；需了解海外B端目标业务，**结合海外业务数据及应用场景提供解决方案，对企业产品逻辑理解要求高**

To C & 国内

- 国内**To C市场产品表现尚未表现明显差异性**，在流量获取、抢占用户新站、数据飞轮等方面有更多机遇与不确定性
- 国内C**端生态与互联网巨头厂商**绑定明显

To C & 出海

- 海外To C APP**技术底座选择更加开放**，拥有更多模型能力选择，在应用表现上占据优势；**竞争激烈，在品类与数量都更加丰富**，部分初创产品已占据用户心智



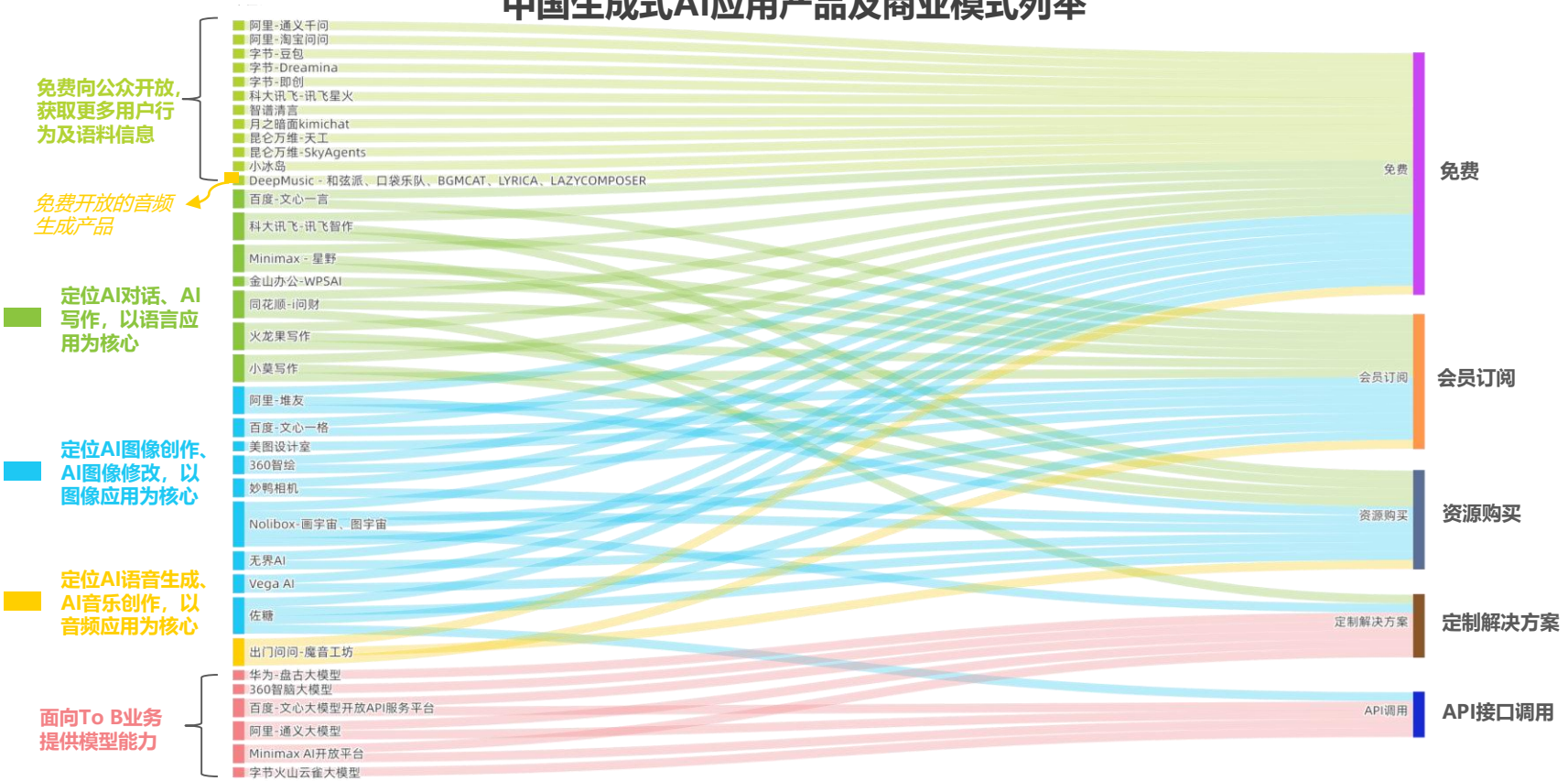
来源：综合微软研究院的《Sparks of Artificial General Intelligence》等公开资料研究绘制。

中国生成式AI应用变现初探

B端带动云资源及产品方案售卖，C端以免费为底提供会员订阅与资源购买

大模型可以更好带动云资源（MaaS模式）及行业产品方案的售卖，在B端产品方案融合更多大模型技术，以API调用、SaaS产品及定制方案等方式加速企业智能化赋能。目前C端产品主要以工具定位为主，页面广告、流量变现（如占领用户更多流量及使用时长实现引流花费）等方式尚未进入当下产品，前者主要原因为发展初期以获取语料及用户行为角度为首要，页面广告会影响用户体验与心智占领，后者主要原因在生成式AI产品在软硬件生态能力尚未达到。

中国生成式AI应用产品及商业模式列举



来源：艾瑞根据公开资料自主研究绘制。

注：B端 SaaS产品结合大模型能力后，仍保持原来会员订阅或资源购买等商业模式，受篇幅限制未列举

由云到边端的AI产业协同
—— 边缘与端侧洞察

◆ 技术背景

① 大模型增加边缘侧计算量，推进边缘智能算力部署。
② 大模型加速终端硬件多模态感知和推理能力的升级。

◆ 边缘发展

① 大模型正在从算力统管和场景优化两个维度在边缘侧进行落地尝试。
② 作为边缘应用前沿，大模型正对自动驾驶技术栈进行全方位升级与重构。

◆ 端侧动态

① 终端模型需与云端模型协同提供服务，存、算、网同步升级。
② AI重塑操作系统是释放大模型潜力的关键，旨在颠覆产品体验及生态。

大模型渗透下，由云向边端多智体进化

AI与新基建相辅相成，形成正向循环，共促智慧物联产业扩展与升级

多智体系统是指由多个智能体构成的系统，智能单体具备感知、存、算、通信能力，智能体之间通过协作交互AI相关信息，实现智能在网络内的流动，从而提升各节点及网络平台的智能水平，是未来物联网发展的目标。大模型在各层融合应用对原有云、边、端的算力及调度、通信、感知能力都提出了全新的要求，物联网的智能进化也为大模型落地铺开更广阔的场景与空间。

大模型在云、边、端落地对物联网技术体系的影响与塑造

来源：《通算一体网络十大基础问题白皮书》、艾瑞咨询研究院自主研究绘制。

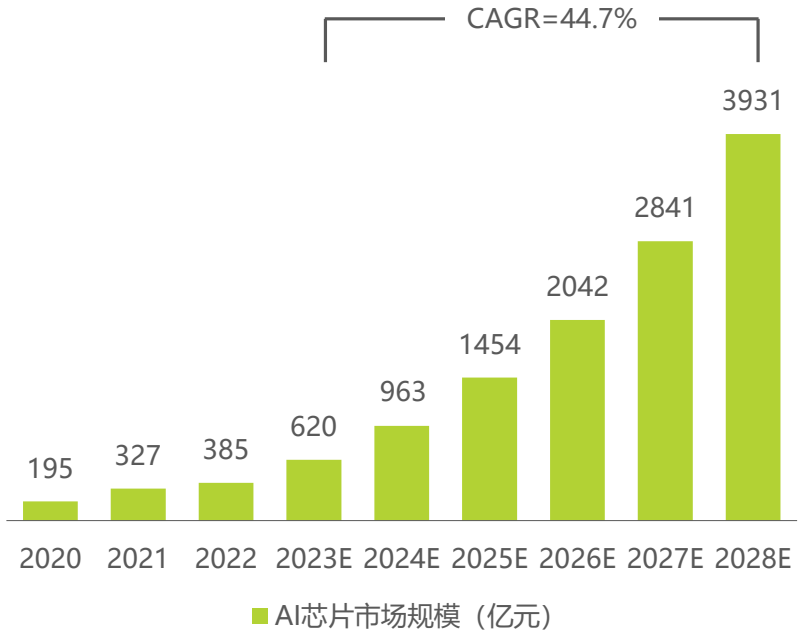
©2024.4 iResearch Inc.

智能算力加速下放至边缘侧与端侧

国产AI芯片积极进行生态适配，边缘大模型与端侧AI应用态势趋显

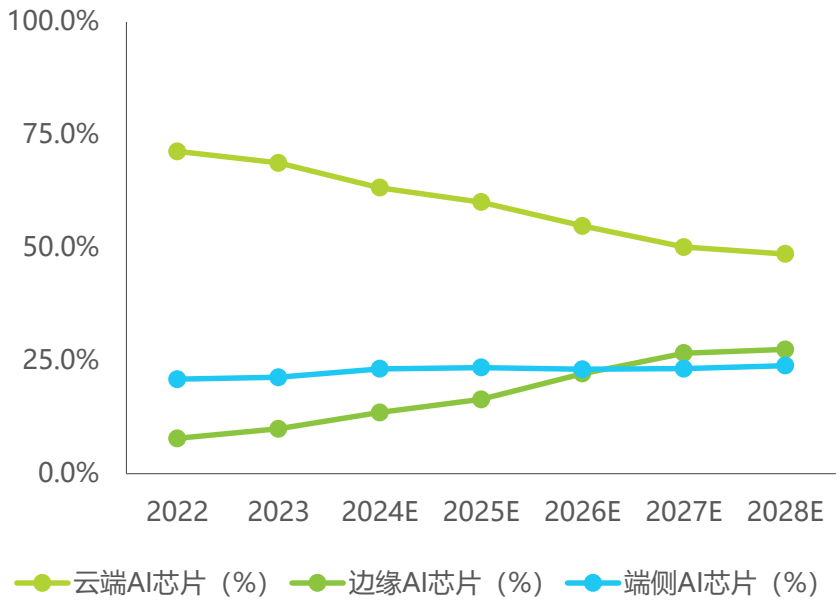
2023年，中国AI芯片市场规模约为620亿元。近年来，国央企加大算力基础设施建设，由更多人工智能产业服务商作为生态伙伴参与共建，强化需求牵引与行业赋能，2028年中国AI芯片市场规模将达到3931亿元，五年复合增长率达到44.7%。在中美关系越发紧张的时代背景下，中国AI芯片厂商坚守自主可控道路，期望早日摆脱对国外厂商以英伟达为代表的高性能卡依赖。以百度昆仑、华为昇腾、中科海光等厂商为代表的芯片产品陆续完成软件生态移植，进入规模化应用阶段。伴随自动驾驶、智慧医疗、智能家居、工业互联网等场景，AI算法与算力调度将从云端逐步下放到边缘侧和端侧。预计2028年云端、边缘侧、端侧AI芯片的比例将分别达到48.6%、27.5%与23.9%，面对大量、复杂任务如何进行与云端的无缝高效的分工配合，是当前边端模型需要解决的关键问题。

2020-2028年中国AI芯片市场规模



来源：艾瑞咨询研究院自主研究绘制。

2022-2028年中国AI芯片应用场景比例变化



来源：艾瑞咨询研究院自主研究绘制。

边缘侧：大模型延展边缘智能空间

实际应用能力与场景仍处于初探阶段

当前，大模型正在从算力统管和场景优化两个维度在边缘侧进行落地尝试。在算力统管方面，大模型能够部分替代和接管原有云端计算中心的算力调度权限与能力，大大减少云—端传输所带来的时间损耗，对边缘侧算力使用效率带来改进。同时，大模型可取代原有边缘侧用于预测、决策、判别、生成等多类任务的小模型，提升场景泛化能力和使用效果，改善ROI。

大模型边缘侧落地应用分析

算力统管
提升边缘算力调度与响应能力

Before:

使用传统的运筹优化算法或深度神经网络模型，由云端计算中心统一对边缘节点进行调度和优化，边缘节点之间能够互相通信，能够单独或协同进行计算存储任务，但仍然缺乏灵活性，智能程度低。

After

大模型实现边缘算力调度及运维

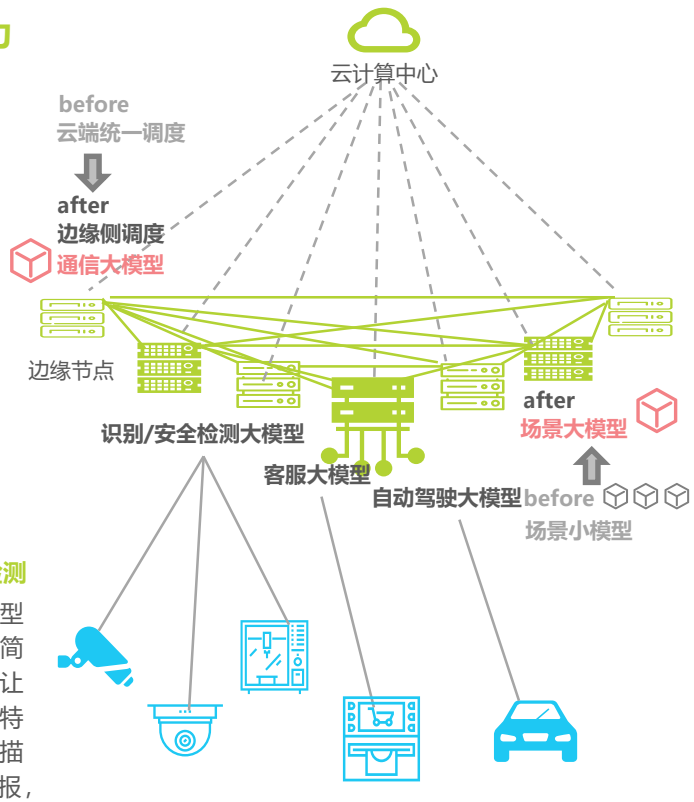
在边缘侧算力扩展以及网络通信能力提升共同作用下，可以支持直接将预训练大模型部署在边缘侧。

01 边缘算力调度优化

相比云端调度，时延能够从数百毫秒降低到数十毫秒，大大提升算力对终端计算需求的响应能力，避免或弱化卡顿、网络崩溃等问题。

02 数据异常检测

利用大模型Agent，通过简单指令设置，让大模型自动对特定数据进行扫描并发现问题上报，提升检测效率。



场景优化
替代小模型，拓展应用能力

由于端侧算力部署成本过高，因此AI模型在智慧城市、智能制造等领域应用一直以来普遍采用云端训练、边缘部署和推理的方式实现。Transformer架构大模型具有较强的内容理解和泛化能力，在产业端替代原有小模型，*训练周期大大缩短，可移植性提升，应用效果明显增强，将大幅提升AI边缘侧场景ROI。*

01 预测+决策类

设备预测性维护/自动驾驶

如电力行业的电量预测、工业设备维护预测等典型时间序列预测场景，使用大模型取得更好的预测效果。自动驾驶当中系统对下一时间即将发生的路况预测及对应决策的生成，能够提升面对长尾场景的响应能力。

02 视觉判别类

安全检测、工业质检

在矿山、工厂等环境的安全检测，以及工业流水线质检、城市道路及车站人流监测等场景，大模型能够提升同时监控对象数量，还能够通过对画面内容的理解进一步提升识别准确率。

03 语言/语音生成类

金融/医疗+数字人

客户接待压力较大的强交互场景，使用大模型+数字人方式构建交互式对话机器人，解答客户问题同时还能够精准导流，未来还会逐步拓展至医疗诊断、金融业务办理、方案咨询等复杂场景。

专家访谈，艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

边缘侧：大模型应用前沿—自动驾驶

大模型正在对自动驾驶技术栈进行全方位升级与重构

生成式AI在自动驾驶软件技术栈应用分析

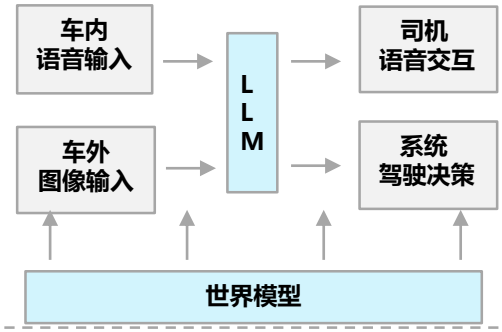
未来新方向

▲**DriveGPT4：多模态形式理解和输出驾驶任务**

在车辆行为描述与辩护、问题回答和控制信号预测等任务上表现良好，且展现出稳健的零样本泛化能力。

▲**世界模型：用多模态大构建自动驾驶的基座模型**

通过对驾驶环境动态建模，预测未来驾驶环境将如何演变，相应做出驾驶决策，其理念与预训练大模型完全贴合，将具备强大的泛化能力，从而有望成为自动驾驶中的基座模型，赋能下游各类具体任务。



系统拟人化

嫁接大语言模型已涌现的上下文学习、零样本学习、逻辑推理、常识判断等能力，提高智能驾驶面对复杂场景的泛化性与可解释性，同时能够将车内与车外信息融合决策，更加智能化实现自动驾驶系统与司机关于路况与驾驶决策的实时交互，

当前自动驾驶技术栈

车端

●**BEV+Transformer：**

不依赖高精地图判断车辆位置和环境轮廓，进行纵向距离测算和补全，实现目标检测、跟踪、3D分割等任务。

▲**集成预测、决策和运动规划的BEVGPT：**

以BEV图像作为唯一输入源，并根据周围交通场景做出驾驶决策，最后通过优化运动规划方法实现驾驶轨迹的可行性和平滑性。

"偷梁换柱"

Transformer架构在各模块中替换任务小模型和规则代码，提升系统简洁性和整体任务效果，为端到端打好基础。感知层应用已相对成熟，头部车企均已实现量产，规控处于起步阶段。

云端

数据采集与回传 ⇩　⇧ 模型蒸馏与更新

▲**SAM：可提示的分割系统**

经过预训练获得强大泛化能力，能辅助数据预标注，以及生成感知、预测和规划的特征输入。

▲**NeRF：2D图像合成3D**

能通过2D数据素材生成3D场景，实现高真实性场景重建，对于长尾场景模拟仿真有重要意义。

降本增效

一方面，利用大模型的生成能力，能够高质量生成训练所需特征和环境，对模拟仿真进行有效数据补充，另一方面，用大模型实现自动化数据标注，降低训练成本。

▲ 初步探索中　● 已成熟应用

来源：艾瑞咨询研究院自主研究绘制。

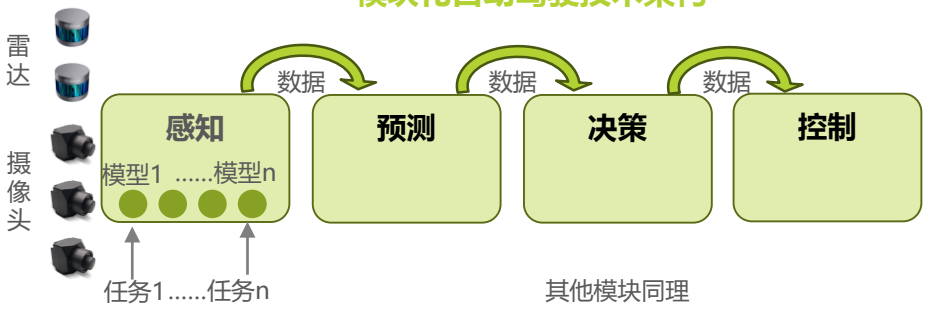
©2024.4 iResearch Inc.

统一的大模型架构是自动驾驶明确演进方向

驾驶全局优化和落地成本改善，端到端正成为领先自动驾驶的技术标杆

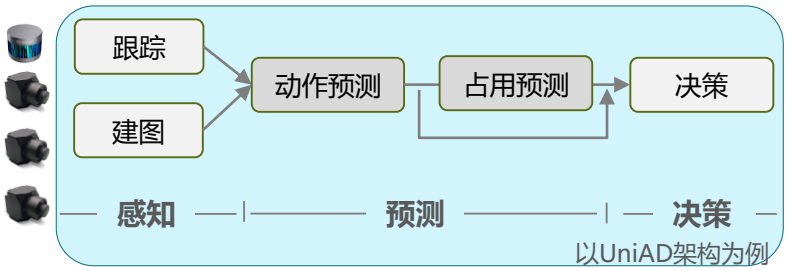
端到端方案架构与优势分析

模块化自动驾驶技术架构



针对独立任务的若干小模型和大量基于规则的指令构成了每个模块，不同模块之间通过接口传递数据，能够实现完整决策链条但算法网络之间彼此断开。

端到端自动驾驶技术架构

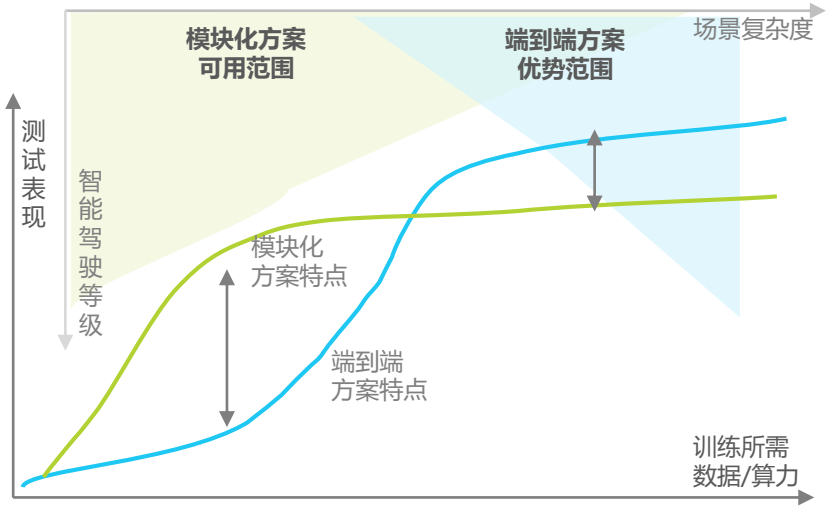


将感知、预测、决策各模块全部用神经网络模型替换，形成统一不间断的算法框架，能够实现以全局优化为目标的计算和决策。

模块化方案特点

量产成本高，技术天花板低，在环境简单的场景能更快应用。

采用大量手写代码实现车辆规控，且不同模块独立优化，这意味着随环境复杂度提高，一方面需要的代码会非常庞大，维护升级困难，且会消耗更大功率，不利于车端部署，另一方面系统设计很难穷尽所有情景，面对新情况，仍需驾驶员接管，很难迈出辅助驾驶的范围。同时，这种方案依赖多颗激光雷达的信息辅助，导致成本居高不下。



端到端方案特点

数据驱动，智能化程度高，响应速度快，驾驶体验好，有望在复杂环境实现高阶自动驾驶。

各模块代码都用神经网络编写，通过大量输入图像视频数据进行训练，“学习”各种情境下的驾驶动作，这使得其能够处理多样化的驾驶场景任务，对复杂长尾问题响应能力提升，同时系统更加简洁，还能够以整体最优为目标实现各模块联合优化。最后，端到端方案普遍采用BEV+Transformer技术框架，能够大幅减少对激光雷达的使用，从而降低单车自动驾驶前装成本。

来源：2023年CVPR、中金公司、艾瑞咨询研究院自主研究绘制

©2024.4 iResearch Inc.

大模型深化赋能是高级别自动驾驶落地关键

L3利好政策信号释放，数据、算力、算法等方面仍需长期积累

L3级级别自动驾驶发展要素分析

1 政策引导L3试点，车企拿到“准考证”但“毕业” 还有很长距离

2023.07 **《北京市智能网联汽车政策先行区自动驾驶出行服务商业化试点细则（试行）》**

正式开放智能网联乘用车“车内无人”商业化试点，企业达到相应要求后可在示范区内面向公众提供常态化自动驾驶付费出行服务

2023.11 **《关于开展智能网联汽车准入和上路通行试点工作的通知》**

遴选达到L3、L4量产条件的车企发放准入测试牌照

明确高阶智驾事故责任归属

虽然上述政策明确支持和推进自动驾驶商业化发展，但《准入通知》中也明确指出，试点实施目的是引导加强能力建设，完善相关法规修订和完善，距离真正上路还有很长距离。

3 无论自用还是商业运营，都需要从各维度进一步降低成本

私家车：功能升级引发的前装溢价值得关注

相比L2，L3级别自动驾驶系统需接管更多复杂场景，激光雷达、摄像头等感知设备需要增加及升级，车端算力也要提升，单车软件+硬件服务成本可能在10万元甚至更高，考验消费者对自动驾驶价值的认知与判断。

出行服务：精细化运营填补商业化最后一环

除整车成本外，robotaxi还面临高昂的安全和运力运营成本。安全方面，在云代驾基础上，还需要逐步提升一个安全员监控的车辆数，摊薄成本，在运力方面，有效的能源成本优化、场站、售后等环节的精细化管理手段也是未来各家运营方的发力方向。

2 头部厂商横向对比，国内玩家仍需追赶

	技术栈	数据	算力（模型训练）
特斯拉	端到端架构，感知侧采用BEV+Transformer，纯视觉无图方案	FSD Beta累计行驶里程12.9亿公里	10E FLOPS（2023.08）
华为	非端到端架构，感知侧采用BEV+Transformer，多感知融合无图方案	阿维塔智能驾驶功能累计行驶里程4400万公里（2024.02）	2.8E FLOPS（2023.11）
小鹏	非端到端架构，感知侧采用BEV+Transformer，纯视觉无图方案	智能辅助驾驶累计行驶里程4.88亿公里（2023.10）	0.6E FLOPS（2022.08）
理想	端到端架初步构，感知侧采用BEV+ Occupancy，规划采用时空联合规划算法，MPC预测控制，无图方案	高速+城市NOA累计行驶里程5.6亿公里（2024.02）	1.2E FLOPS（2023.06）

从技术来看，特斯拉率先完成端到端落地，领先国内厂商1-2个身位，其背后是提前多年的前瞻性技术实践。当自动驾驶与Transformer融合越紧密，其所表现出的数据和算力驱动特征会愈发明显。来自真实驾驶环境的数据是模型训练最重要的数据来源之一，从数据与算力储备看，国内头部车企与特斯拉仍存在较大差距。

来源：公开数据、艾瑞咨询研究院自主研究绘制

©2024.4 iResearch Inc.

端侧：大模型加速AI与终端融合

终端模型需与云端模型协同提供服务，存、算、网同步升级

当前阶段，大模型已经率先在手机和汽车座舱中得到初步应用，其带来的计算存储需求也在催化终端硬件和网络性能迭代。同时，在大模型裁剪技术以及终端算力制约下，端侧部署大模型参数量小，功能相对有限，部分时刻借助云端大模型能力可以为用户提供更丰富的场景体验。未来随着场景复杂化和用户、设备协同等需求，对端侧和云端模型能力及算力需求也将同步提升。

端侧大模型应用能力及软硬件技术分析

终端硬件配置及网络性能升级

硬件（以手机为例）

SoC及内存
1）AI算力和图像处理能力不足，核数需降低，主频需提升
2）存储空间需增加

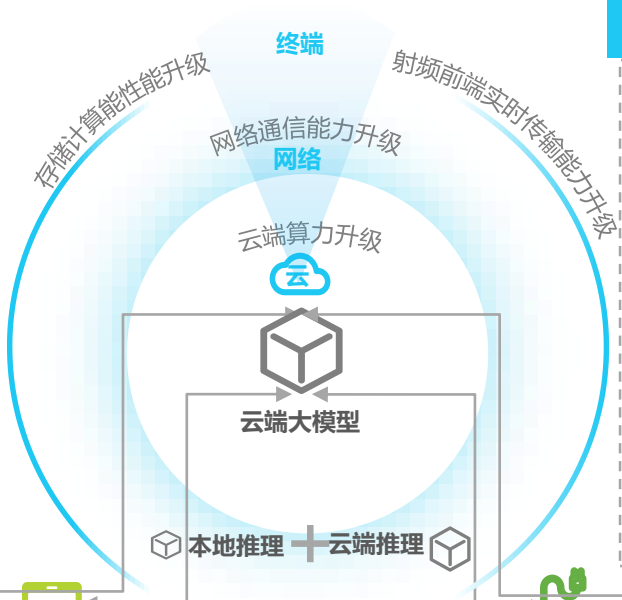
通信模块
需要更大带宽及更多通信频段，通信模块集成度需提升

8GB+256GB ⇩（2年内实现） 16/24GB+1TB
非阵列天线 ⇩ 阵列天线
射频分立器件 ⇩ 射频模组

网络与调度

1）智能传感器增加，数据维度及数据量提升，需要更大带宽及流畅性

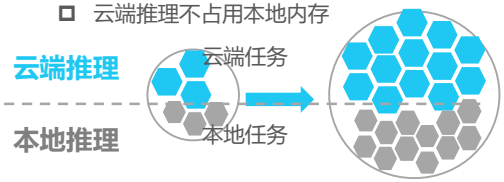
2）在不同环境及针对不同任务，对云端与终端算力调度能力需提升



端侧模型不会无限扩大，终端与云端推理任务量长期都将增长

云端与本地推理各具优劣势，厂商将根据实际应用场景和任务要求，综合调度云端与终端模型能力

- 云端模型参数量大，能够处理更复杂的任务，但实时性不好，且对网络有较高要求
- 云端推理不占用本地内存



- 本地推理速度快，成本低，但模型参数小，能力上限较低
- 本地推理用户数据隐私保护更好
- 本地使用更便捷，交互体验更好

PC
端侧参数量10-100B

功能实现：
- 通用办公类：文档、表格、幻灯片编辑、语音转写
- 专业工具类：代码生成、图像视频编辑、……

手机
端侧参数量1-7B

功能实现：
- 参考oppo、vivo、华为等
- 语言类：语音摘要、文字摘要
- 图像类：图像生成、图片编辑

座舱
端侧参数量1-10B

- 图像类：AI绘画、……
- 语言类：指令交互、出行规划、通识解答

功能实现：参考理想MindGPT

机器人
端侧参数量10-100B

功能实现：
- 工业生产：流水线装配、分拣等、……
- 医疗护理：瘫痪护理、重症监护、……

来源：艾瑞咨询研究院自主研究绘制

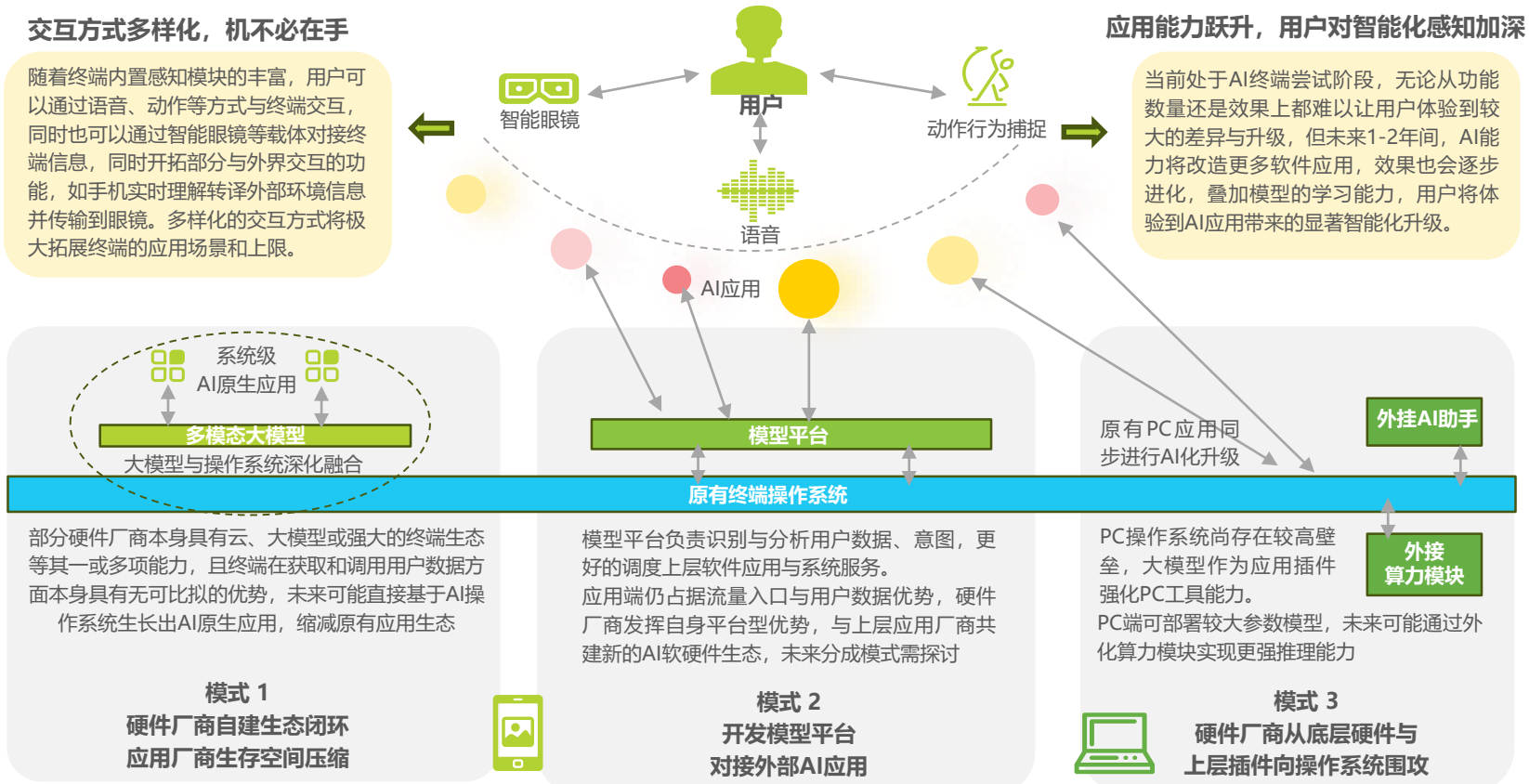
©2024.4 iResearch Inc.

AI原生硬件将颠覆产品体验及生态

AI重塑操作系统是释放大模型潜力的关键，硬件厂商更有机会建立完整生态

从用户感知视角，多模态的人机交互将解放用户双手，AI终端将从存储—应用—交互一体的娱乐/工具机，逐渐演化为用户随身携带的智能BOX。作为智能算力和应用的载体，终端应用的范围和能力也将得到极大拓展。从技术栈层面，操作系统作为全机能力调度的核心将发挥更显著的主体性作用，硬件厂商将以AI操作系统为核心重塑自身生态，原有软件厂商的用户数据与流量入口优势或将被削弱。

AI终端用户生态及软硬件生态变革分析



来源：专家访谈，艾瑞咨询研究院自主研究绘制。

03/ 中国人工智能产业商业化实践标杆

AI - Case

京东云·言犀

提供客户全渠道全生命周期的营服销一体化智能服务

京东云·言犀依托于全栈自研的人工智能技术，基于京东集团广泛实体业务、庞大而又复杂的产业生态，从内部真实、复杂的海量业务场景实践中推出千亿级参数的言犀大模型，打造全新的智能交互与生成能力，从文本、音频、图像到多模态内容生成，技术上推动从感知智能和认知智能到决策智能的跨越，应用上推动服务、营销、运营一体化创新，目前已涵盖“在线咨询机器人”“语音应答”“AI外呼”“商家智能客服”“直播数字人”“客服数字人”“营销图文生成”等在内的营服销一体化产品矩阵，聚焦体验、效率与转化，精准理解客户意图、高效解决客户问题。言犀不仅为京东超6亿用户提供智能化咨询服务，还为零售、金融、教育、政务等行业超过36万家客户提供以用户为中心的、全渠道全生命周期的营服销一体化智能服务方案，智能客服·言犀通过前沿的智能技术与规模化的应用实践相结合，助力政企客户服务和营销数智化转型升级。

京东云·言犀平台全栈产品及服务能力

来源：艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

京东云·言犀

零售营服销一体化解决方案，长效经营客户全旅程价值

京东云·言犀依托于全栈自研生成式AI技术与20年零售领域场景Know-how，对零售行业客户全生命周期管理痛点深度剖析，以言犀京小智、言犀AI外呼、言犀数字人等产品为抓手，面向36万品牌客户、京东商家提供集种草引流、消费导购、直播转化、咨询服务、私域运营为一体的智能解决方案，构建以电商渠道客户为中心的“服务-营销-销售”一体化体系，基于言犀大模型在用户触达、用户服务、消费洞察、经营分析、营销内容生成等细分场景落地，帮助品牌降本增效，智能化、精细化的长效经营客户全生命周期价值，助力企业业务新增长。

京东云·言犀营销服全链路解决方案

用户层

C端：消费者

B端：营销人员/店铺运营/会员运营/客服/AI训练师

价值层

售前-售中-售后全场景覆盖、智能转化、高效接待

用户全域全生命周期个性化沟通

品牌增量与内容传播效率提升

应用层

“京小智”商家智能客服

售前		售中		售后
主动营销	高效接待	订单转化	自动化服务	售后运营
智能推荐 优惠推送 新品推送 活动推送	活动咨询 智能尺码应答 商品属性应答 智能截图回复 智能问答预测 服务诊断优化	咨询未下单催拍 加购未下单催拍 降价未下单催拍 到货未下单催拍 下单未付款催付 预售尾款催付	物流咨询 订单确认 发货提醒 配送提醒 订单改地址 自助开发票 安装预约 评价邀约	定期服务提醒 增值服务推送 商品保养关怀 专属客服推荐 .....

11类独有自动化特色场景

多端系统打通价保、订单、物流、发票..

全品类知识体系沉淀

个性化店铺AI训练

AI语音外呼

引入期	成长期	成熟期	衰退期	流失期
私域导流 引导注册用户加粉、到店	首单+复购转化 引导注册用户完成首次下单，后续复购且快速转化为成熟用户	价值挖掘 持续挖掘用户价值，延长成熟期	预防流失 预防用户流失，引导回归成熟期	流失召回 流失用户召回，维持用户关系
企微加粉	首单特价 新会员特惠 节点福利 邀约到店	复购折扣 限时优惠	生日关怀 积分兑换通知 节日问候 优惠推送	老客优惠 会员权益提醒	新款邀约 定向大额券 用户回访

语音/视频外呼

高价值人群圈选

对话数据挖掘

全链路数据回流

个性化外呼

营销自动化SOP

AI多模态数字人

直播数字人

轻部署	高表现	有灵魂	高效率	深融合
云端一键开播，对电脑配置无要求	更真实的大姿态数字人，多种行业形象风格	智能导购营销交互，知识库持续调优	直播素材智能生成一键开播流程，直播切片自动生成	场景与数据高度融合，营销数据全角度追踪

数字人短视频

2D仿真人/形象库
预制素材模板+灵活自定义操作零门槛、多场景应用

数字人分身

5分钟素材定制
1句话声音克隆
医生、律师、教授、企业家、大V

支持大姿态(侧脸/走动)

大模型赋能交互与内容生成

支持动态局部高清

支持语义驱动的肢体动作编排

来源：艾瑞咨询研究院自主研究绘制。

百融云创

百融云创 Bairong.Inc.

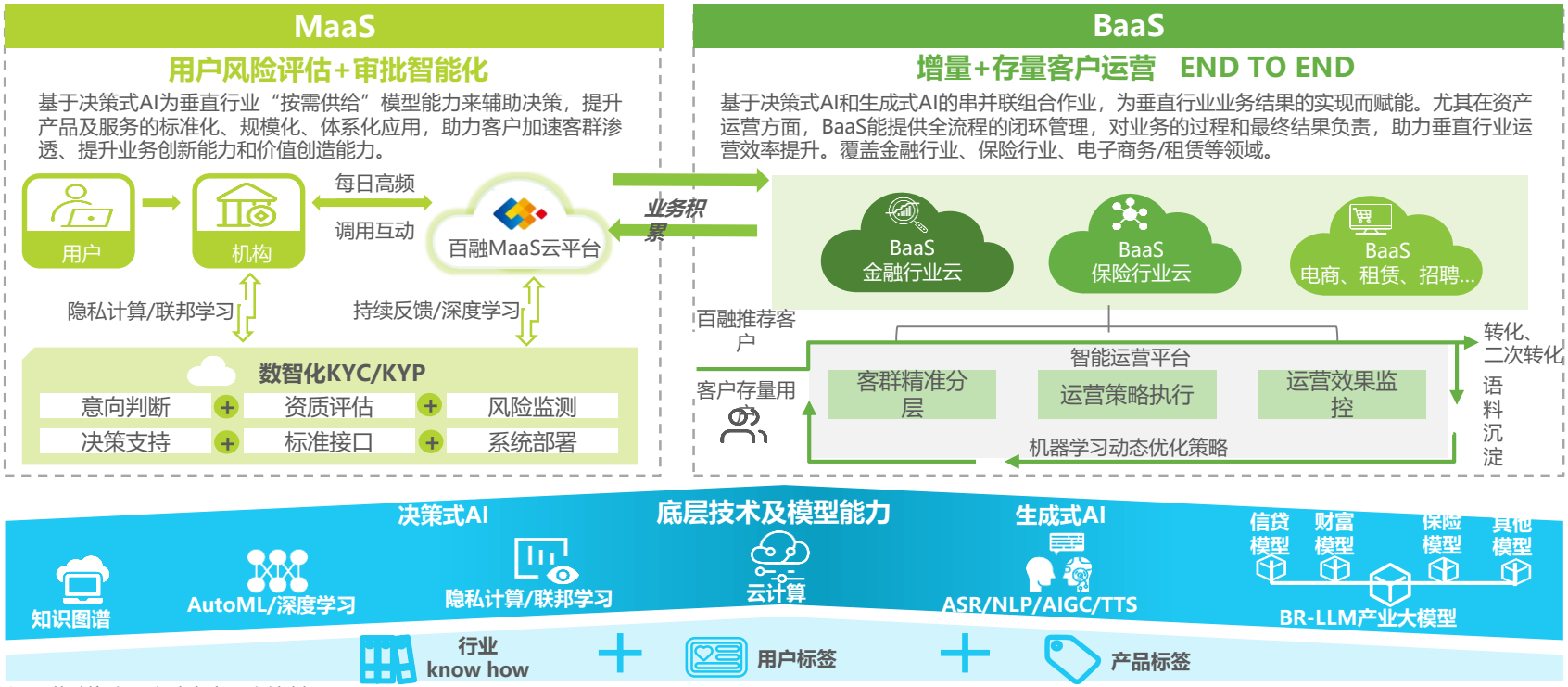
iResearch 艾瑞咨询

零售营服销一体化解决方案，长效经营客户全旅程价值

百融云创是一家人工智能（AI）技术服务公司，凭借全面的技术能力、长期行业Know-How和客户洞察，为金融、保险、零售等行业客户提供全流程数智化服务。其中，MaaS通过丰富AI模型与知识图谱进行用户风险评估和筛选，用户可直接基于云平台进行能力调用；BaaS遵照结果导向，为客户建设端到端的用户营销运营数智化体系，与客户收益共享，风险共担。同时，百融云创也具备自研的预训练大模型能力，并基于大模型形成多个标准化场景小模型，以及语音、数字人等标准化封装的模型能力，更高效赋能客户业务拓展。成立10年来，公司累计服务7000+机构客户，并已于2021年在香港上市。

百融云创产品技术架构

MaaS能力指标亮点： 1.模型平均每天**3亿次**以上调用，云平台稳定性达99.998%；2.BR-Coder将开发岗位中的自动生成代码渗透率提升至**10%**；3.ORCA-AutoML自动学习协助数据分析岗位将建模时间缩短了**30%**



来源：艾瑞咨询研究院自主研究绘制。

AI能力与业务场景无缝融合，多层次打造高增长高收益的数智企业

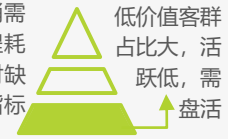
基于决策式+生成式AI技术底座与多年行业服务经验，百融云创为金融、保险、零售电商等多领域客户打造端到端的数智业务闭环，帮助客户快速实现营销目标。此外，百融还可利用自研大模型能力研发出企业数字员工平台，由平台生成的Bot以AI Copilot和AI Agent形式出现，相当于为企业装入一个“企业级助手团”，极大提升企业工作效率。

百融云创生成式+决策式AI深度服务案例

01 MaaS+BaaS打造财富管理业务闭环

客户痛点

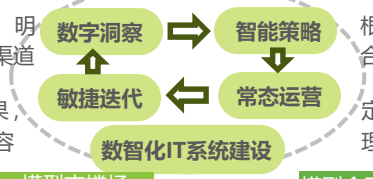
某城商行财富营销需手动策略执行流程耗费人力成本，同时缺乏数据量化监测指标和常态化运营机制。



解决方案：百融“4+1”财富管理数智化解决方案

定制分析客户画像、明确产品内容和营销渠道

持续收集营销结果，优化营销策略和内容



根据产品类型和客群特点，结合百融专家经验定制营销策略

定制人工话术和短信模板，梳理客户标签体系和营销策略库

提升效果

整体AUM增长达**15亿**

AUM < 1000元 > XXX户 **+8000户**

AUM < 10万元 > XXX户 **+4000户**

02 VoiceGPT提高金融营销转化率

模型支撑场景核心环节

- ✓ **响应速度：**低于500毫秒级的响应，基本接近真人
- ✓ **理解准确率：**客户语义理解准确率达95%以上
- ✓ **通话规模：** 支持日3000万通以上智能语音沟通

VoiceGPT 智能语音机器人

某国有银行案例：

客户痛点：客群数量大，传统的电话触达、现场走访等方式效率低，导致对大量尾部客户触达与管理不足，业务增长缓慢。

解决方案：以VoiceGPT为核心搭建智能营销-外呼系统

数字洞察	话术迭代	定制策略	内容运营
以画像标签和营销模型进行客户前筛，精准锁定潜在高价值客户	根据产品类型和客群特点，定制体系化外呼营销策略	定制话术，深挖营销内容，将AI语音服务嵌入零售业务条线	持续收集营销结果，并迭代优化营销策略和内容

提升效果：以人工服务标准90%-95%的效果，实现“IVR+人工”模式节省成本约50%，“IVR”模式节省成本约90%。

BR-LLM 大模型基座

03 大模型平台全面激发数字员工生产力

模型全面渗透各业务环节

Cybertron 赛博坦平台

行业语料

发布企业专属机器人

AI Agent 行业专家

企业知识库

AI Copilot 岗位助手

- 营销专家
- 理财专家
- 风控专家
- 数据分析助手
- 产品经理助手
- 人力助手
- 编程助手

企业级助手团

对内：辅助员工高效办公

对外：以智能数字人客服等方式对客户提供服务

提升效果：某商业机构接入编程大模型BR-Coder，辅助自动生成测试用例和单元测试、解答技术问题等，在保障企业数据资产安全的同时，将模型生成代码一次采用率提升20%，助力研发提质增效。

来源：艾瑞咨询研究院自主研究绘制

©2024.4 iResearch Inc.

九章云极DataCanvas

AIFS人工智能基础软件，赋予企业自主建设大+小模型能力

九章云极DataCanvas以“创造智能，探索未知”为使命，以“助力全球企业智能升级”为愿景，是中国人工智能基础软件领军者。公司致力通过自主研发的人工智能基础软件产品系列和解决方案为企业用户提供AI能力和人工智能基础服务，助力用户在数智化转型中轻松完成模型和数据的双向赋能，低成本高效率的提升企业决策能力，实现企业级AI规模化应用。AIFS作为一款行业领先的人工智能应用构建基础设施平台，覆盖了大模型的训练、精调、压缩、部署、推理和监控以及小模型的全生命周期过程，它为数据科学家、应用程序开发人员和业务专家提供了一套工具，使不同角色的人员可以相互协作，轻松地处理数据并使用这些数据来开发、训练和部署任何规模的模型。DataCanvas Alaya九章元识是九章云极DataCanvas自研的“通识+产业”白盒大模型矩阵，作为AI Foundation Software的核心能力之一，秉持开放友好的开源理念，为用户赋予更大自由度的AI创新能力，以求加速实现大模型在多元业务场景中的应用。DataCanvas Alaya提供了一系列不同配置和参数的，具备业界前沿能力和技术的预训练大模型，在文本对话，图像生成，重塑当前AI软件形态。

九章云极DataCanvas大小模型产品布局图

来源：艾瑞咨询研究院自主研究绘制。

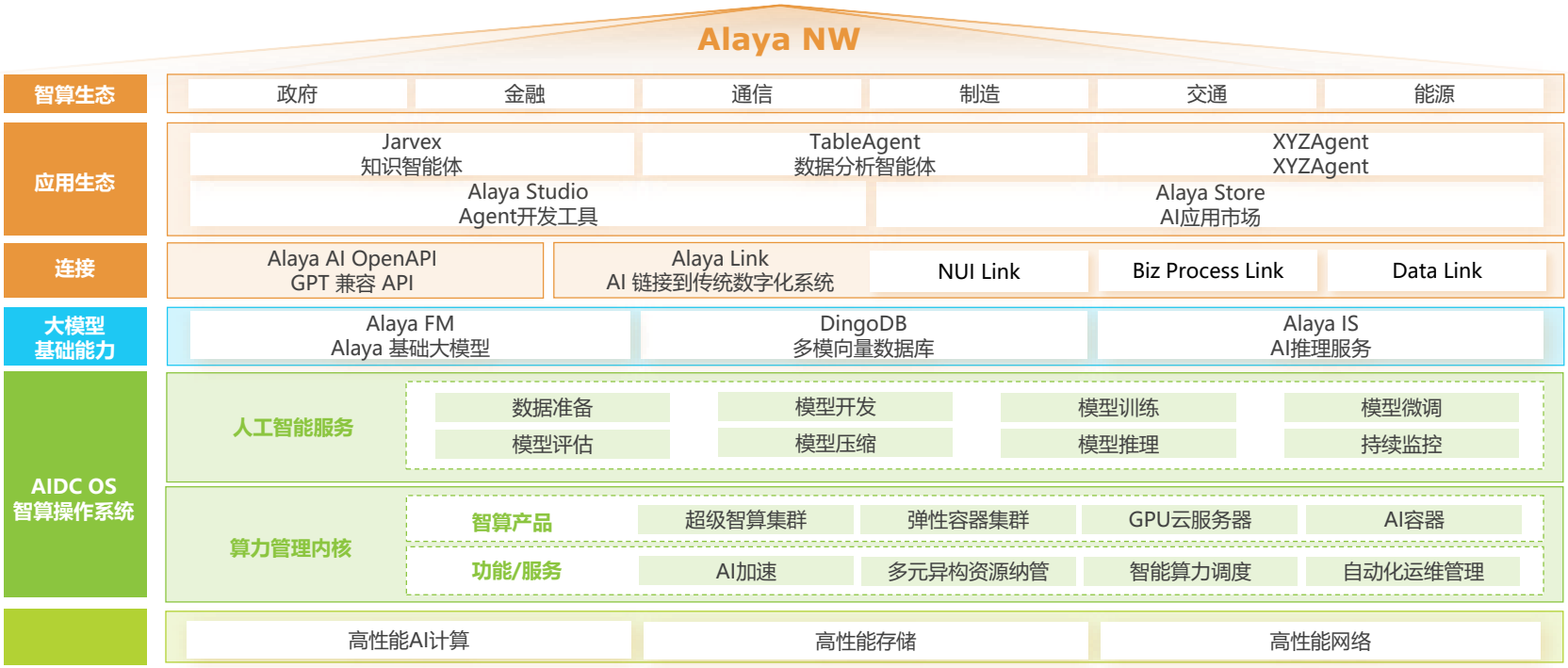
©2024.4 iResearch Inc.

九章云极DataCanvas

九章云极智算操作系统为AI而生，构建智算生态

九章云极DataCanvas智算操作系统（DATACANVAS AIDC OS，简称AIDC OS）面向智算中心、智算中心算力运行与业务运营，可以支持大中型企业内部智算集群的运行。其核心能力是智算资源的纳管、统一调度，智算业务的业务运营支撑，以及AI模型的构建、训练和推理。其特点包括：异构算力纳管、丰富的算力服务产品形态的支持、多策略统一调度、AI大模型+小模型低门槛的训练和推理、AI模型训练全过程监控与容错能力等。 AIDC OS能有效提升智算中心资产的附加价值，将运营方的运维能力从裸算力设备运维提升到AI大模型运维能力；同时便于不同类型的终端用户快速上手使用智算算力开发和运行AI应用；此外，它也是智算中心开放生态环境的基础支持设施。DATACANVAS AIDC OS为算力中心提供高效的算力操作系统，提供更多、更便捷、更高效的智算服务。

九章云极DataCanvas智算生态布局



来源：艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

04/中国人工智能产业旅程

AI - Coming

社会公众侧：AI应用衍生多种问题

公众接受度与容忍度考验AI技术进一步成熟、全流程可控、可监管

社会层面，人工智能技术值得关注的主要风险在于对用户心智、用户隐私及安全伦理问题的潜在影响。首先，人工智能训练需要大量互联网公开数据，包括图片、文字、音频等，可能包含大量私人数据，存在较高的泄露风险。其次，生成式AI可以用于创建虚假的图像、视频或声音，结合大数据分析，可以个性化地分发新闻，对目标受众的心理和行为产生影响，这可能导致人们被欺骗、误导，甚至被利用危害人的生命健康。

人工智能技术的社会公众影响分析

01 占领用户心智

AI能够生成虚假图像，并通过对用户心理和行为的分析定向产生内容，能够影响公众的价值观和政治立场，或诱导用户消费决策

之前，一段声称是乌克兰总统泽连斯基呼吁士兵放下武器的视频走红网络，但乌克兰国防情报部门在推特上澄清了这个视频，并解释了Deepfake技术。他们指出，Deepfake可以用来伪造政要的形象，例如美国总统拜登。

在数字直播当中，数字人所有互动皆为预设，缺少即时互动，消费者很难判断直播内容的真实性，同时也缺少产品评价、产品体验的相关信息，甚至没有人工客服的参与，在这种相对封闭信息来源的环境中，消费行为容易被诱导。

02 影响用户隐私

AI的训练可能吸纳部分涉及用户隐私的数据，同时这些数据可能伴随着AI的应用被二次泄露，并引发相应风险

人工智能技术具有信息关联的能力，一旦系统通过各种渠道获知了足够多与当事人相关的信息，如购物信息、订阅信息、旅行信息、认证信息、信用信息、位置信息等，通过人工智能技术就能够很容易地挖掘出人们的隐私，而且人们很难追踪这些个人数据和隐私信息的泄露途径与泄露程度。

03 引发生命安全与伦理问题

AI在部分场合取代人类引发权责问题，同时AI在人体和基因相关的应用可能导致伦理问题

自动驾驶系统的运转是通过对驾驶内外部环境进行感知，形成判断和决策并做出相应的驾驶行为，一旦感知出现错误，如没有感知到对面的障碍物，其决策就是错误的。2020年6月在台北仙桃，特斯拉的自动驾驶系统把白色翻倒的卡车误认为没有障碍物，导致了车辆在开启自动驾驶的状态下毫无减速地撞上卡车，随着理想、华为的NOA功能逐步推广，自动驾驶/辅助驾驶事故也开始增加。

上述维度交织共同引发AI犯罪（AI诈骗，AI谋杀等）

通过AI换脸技术冒充熟人进行诈骗案件已发生多起，随着技术发展，AI犯罪的形式可能还会增加

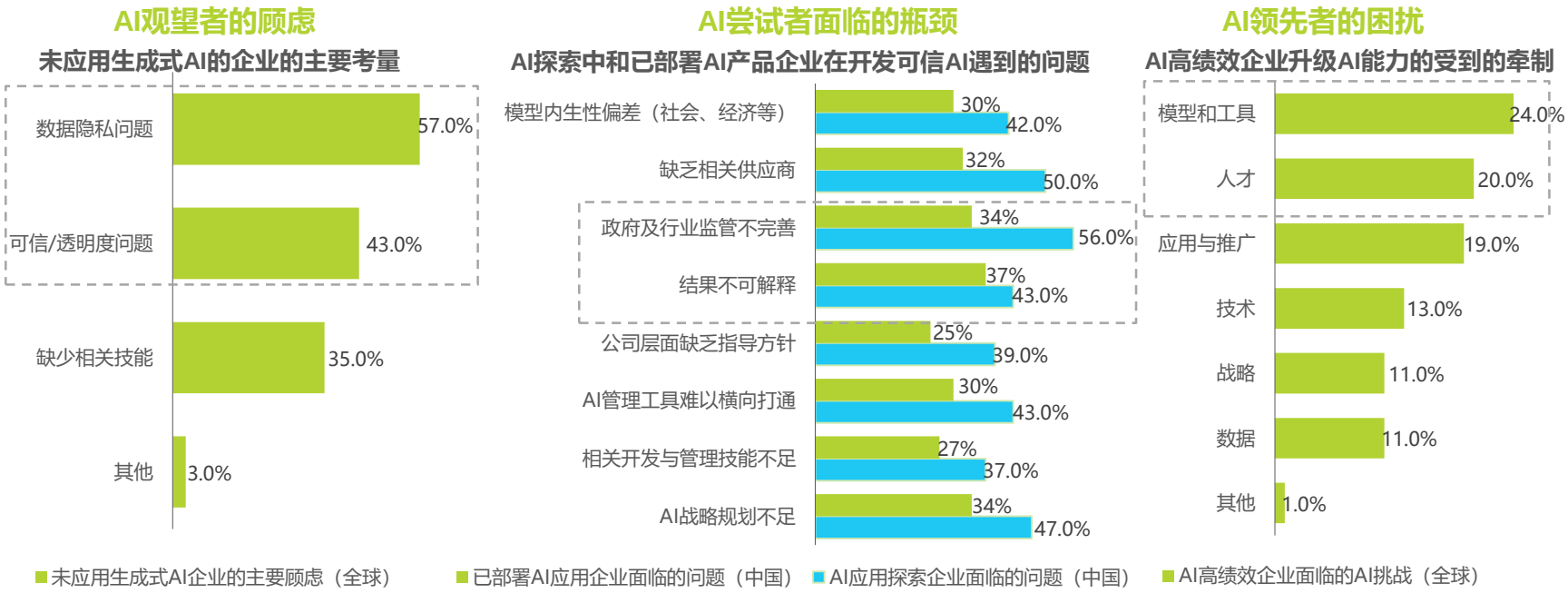
来源：艾瑞咨询研究院根据公开资料自主研究绘制。

企业应用端：AI可用性与易用性仍遭受挑战

新技术内生缺陷下，企业推广囿于数据安全、可信等问题

在企业端，AI技术的内生性缺陷对企业应用的影响更为明显，包含人工智能框架、数据、算法、模型任一环节都能给系统带来脆弱性。传统神经网络模型大多面临可解释性不足的问题，而同为神经网络结构的大语言模型作为近年AI领域重大突破，在可信、可解释性和生成内容安全性方面并未得到明显改善，甚至因其应用场景的扩大使得这一矛盾更加突出。根据IBM报告显示，全球范围内绝大部分公司对AI持积极态度，且从2023年3月至今，这种积极性还在持续发酵。但从不同AI成熟度企业关注问题可以看出，企业AI应用当中，数据安全、技术可信和工具迭代等各方面问题依然突出。处于观望或AI应用初级阶段的企业，致力于通过AI统管和AI技能补足，解决AI应用当中必将面临的数据和系统安全问题，实现AI初步可用可落地。同时，这些公司在AI应用过程中也十分需要公司内部顶层设计与政府、行业标准的支持。而少部分已经成熟应用AI并取得较好降本增效成果的企业则已经进入新阶段，其面临的主要难题在于对企业内外部AI模型和工具进行升级。

不同AI成熟度企业的AI落地难点分析



来源：《2023年全球AI采用指数》，IBM；《2023年人工智能发展现状：生成式AI的突破之年》，麦肯锡，艾瑞咨询研究院自主研究绘制。

©2024.4 iResearch Inc.

多措并举促进AI产业有序发展

从技术、法规与标准层面，坚持AI产业发展安全与效率并重

基于上述对人工智能发展风险的探讨，未来人工智能的发展需要在技术、行业标准规范和法律监管三个层面持续完善和引导。在技术研究方面，必须加强研究，提高算法的准确性和透明度，以防止偏见和不公平情况出现。在行业标准方面，需建立统一的规范和伦理准则，确保人工智能应用符合道德和社会价值。在法律监管方面，则需制定和修改相关法律法规，保护个人隐私，防止滥用和侵犯权利。

两会代表关于人工智能产业发展的建议

规范引导

全国政协常委、浙江省政协副主席陈小平：

1. **制定国家层面的人工智能发展战略规划**，优先扶持若干具有竞争力的地区、企业，给予“一地一策”“一事一议”重点支持。
2. **推动建立人工智能大数据开放创新平台**，鼓励企业和研究机构共享符合法律法规要求的高质量语料库资源。支持开展专门的数据标注清洗等预处理工作，为训练大模型提供优质数据来源。
3. **研究制定更适应人工智能产业发展特点的数据权益保障和流通规则**，降低企业合规成本。

全国人大代表、科大讯飞董事长刘庆峰：

制定国家《通用人工智能发展规划》，系统性加快推动我国通用人工智能发展。正视差距，聚焦自主可控的底座大模型“主战场”，从国家层面聚焦资源加快追赶。

技术升级

全国人大代表、农工党中央委员、南昌大学元宇宙研究院院长闵卫东：

要在研发及商业应用等阶段，**在数字技术层面加快识别AIGC内容的研究**，能够做出智能判断，加强技术手段来防范人工智能技术应用带来的内容风险；

全国政协委员、360集团创始人兼董事长周鸿祎：

鼓励企业在定制AI大模型前，做好知识管理，将企业大数据平台升级为企业知识平台。在此基础上，通过垂直训练，深入企业级场景，满足企业需求。鼓励和引导企业将大模型与数字化业务系统深度结合，同业务流程相结合，充分发挥大模型价值。

法律监管

全国政协委员、知乎创始人兼CEO周源：

1. **建立数据合规的监管机制，推动完善AIGC监管立法**，保护和规范人工智能领域的数据合规。
2. **对大模型的数据采集来源、处理方法、合规性等进行监督和审查。**加强对大模型的社会影响和风险评估，及时发现和解决可能存在的问题。
3. **加强数据安全和知识产权的保护措施**，并加快高质量中文数据集的开发与利用。

全国政协委员、金杜律师事务所高级合伙人张毅：

当前我国人工智能算法治理存在法律法规效力层级不高、监管机构监管竞合、监管真空或失焦、用户权利难以充分救济等问题，可从健全算法综合治理体系、优化监管工具、加强用户权益保护等方面着手完善。

产业发展措施

产业发展风险

企业及个人

数据隐私保护，技术可信、可解释性较差

系统自身

系统复杂性加剧的同时漏洞也会增加，安全性堪忧

社会

潜移默化影响社会公序良俗，甚至危及社会治安

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

艾瑞—专家之声

AI即将带来新时代变革，邀请专家共话人工智能产业未来

报告结尾，艾瑞特别邀请了百融云创、九章云极、云从科技、力维智联、拓尔思、中关村科金、活树科技等国内明星厂商及标杆企业，分享对我国人工智能产业技术变革、赛道动态及未来趋势展望的见解，共同探索中国人工智能产业的发展方向。

专家之声-精华总览

百融云创

只有在行业高度和广度上有超过通用大模型的表现，垂直大模型才能赢得生存空间，有机会形成业务和数据的飞轮效应，实现场景闭环。

——陈昀彰

九章云极

我们对于未来的愿景，即智算操作系统将定义的新的算力基础设施，让人类进入计算的新世界。

——方磊

云从科技

用技术创新赋予「躯干」灵魂，打造有竞争性的「神经」和「大脑」，让 AI 具备人机协同能力，真正成为各行业的专家。

——姚志强

力维智联

AI智能体将成为下一代平台，从Copilot副驾走向主驾，让数字员工成为新常态；具身智能加速进化，AI有望完成“感知-决策-行动”的闭环。

——员晓毅

拓尔思

随着不断推陈出新的开源大模型的快速迭代，国内自训的闭源大模型将逐渐收拢，面向企业级的大模型应用正在开花，基于智能体的服务模式将是主流。

——林松涛

中关村科金

随着技术的不断成熟，大模型的应用场景将变得更加多样化，尤其是数字员工或智能助手类的应用场景商业机会将会非常大。

——张杰

活树科技

在当今数据驱动的时代，高质量/多语种的数据不仅是大型AI模型开发的基础，也是推动整个AI行业向前发展的关键力量。

——张伟贤

来源：艾瑞咨询研究院自主研究绘制

©2024.4 iResearch Inc.

陈昀彰

百融云创AI创新负责人，大数据及机器学习专家，SaaS和云计算专家

只有在行业高度和广度上有超过通用大模型的表现，垂直大模型才能赢得生存空间，有机会形成业务和数据的飞轮效应，实现场景闭环。

大模型与生产力工具的有效结合，将在未来成为数字系统的基座和核心，例如在客服、SDR和运营等场景，基于大模型的语言理解生成能力可以自动编排作业流程。由于大模型基座对应用数据的虹吸效应，随着诸多基座模型的开源，以及一系列低成本的微调方案的出现，将有越来越多的机构和企业，会基于企业和行业的需求定制大模型。

垂直领域大模型成功的关键在于对场景的理解，只有深刻的行业know-how，才能指导行业数据的积累和语料的设计工作，这一点决定了垂直大模型的应用效果，只有在行业高度和广度上有超过通用大模型的表现，垂直大模型才能赢得生存空间，有机会形成业务和数据的飞轮效应，实现场景闭环。

百融云创是以“AI+金融”起跑，经过多年的数智服务，百融云创的产品和解决方案可以覆盖金融领域的全生命周期，涵盖信贷场景、增量用户获取、机构资产端、财富场景、存量用户运营、机构负债端等等，积累了大量的业务经验和用户反馈。在过去多年的积累基础上，推出了百融金融大模型BR-LLM，拥有更精准的知识引用、智能的工具使用能力和丰富的金融行业知识，可以赋能金融机构，在生产力场景中提升智能化水平。

©2024.4 iResearch Inc.

方磊

九章云极DataCanvas董事长，北京市卓越工程师，中关村高端领军人才，于清华大学和弗吉尼亚理工大学分别获得学士和博士学位

我们对于未来的愿景，即智算操作系统将定义的新的算力基础设施，让人类进入计算的新世界。

大模型的崛起带来的底层计算模式的变化，正引领新一轮底层计算变革。计算从来都是软硬件协同作用的。回顾计算演变历程，硬件为满足计算需求不断演进，软件则作为桥梁连接硬件与应用。随着硬件同质化趋势加剧，软件创新空间扩大，角色愈发重要。

算力经济时代下，软件将成为算力单元的定义者，AIFS人工智能基础软件同样迎来巨大发展机会。以AIFS人工智能基础软件为基石的DATACANVAS AIDC OS智算操作系统，以AI使用能力为驱动，不仅关注硬件资源的有效管理和利用，更着眼于如何更好地满足终端用户对算力的核心需求：将可用的算力、好用的算力提供给算力消费者。

AIDC OS不仅仅是软件与硬件之间的桥接，更是新计算世界中定义算力基础设施的关键力量。它定义的新的算力基础设施，将引领人类进入计算的新世界！

©2024.4 iResearch Inc.

姚志强

云从科技联合创始人

用技术创新赋予「躯干」灵魂，打造有竞争性的「神经」和「大脑」，让 AI 具备人机协同能力，真正成为各行业的专家。

套壳ChatGPT功能的初创企业大概率会技术迭代中被吞噬。因为数据不是护城河，行业经验才是。以后的基础大模型很可能会穷尽世界上所有公开数据（目前已经接近）和大部分半公开场景数据，所以，对行业工作流程、场景智能化的经验才是立身之本。单纯的文案撰写、文生图等单点功能不足以支撑企业的长远发展。

接下来期待不同场景的「AI精灵」（AI-agent）爆发。这和云从的长期目标是一致的。如果说人工智能远景是打造一个具备「四肢躯干」和「大脑」的机器，那云从要做的，就是用技术创新赋予「躯干」灵魂，打造有竞争性的「神经」和「大脑」，让 AI 具备人机协同能力，真正成为各行业的专家，全面提升效率和用户体验。

AI大模型在技术进步、数据驱动和多领域应用等方面取得了显著成果。随着深度学习、自然语言处理等技术的不断发展，在多模态领域的能力同样得到了重大提升。它的发展不应仅仅局限于技术层面的进步，更应注重其社会影响和伦理考量。随着大模型在各个领域的广泛应用，我们必须认真对待其带来的数据隐私、偏见和信息茧房等问题。在追求技术卓越的同时，不能忽视对人类价值观和伦理原则的坚守。

©2024.4 iResearch Inc.

员晓毅

力维智联 技术副总裁

AI智能体将成为下一代平台，从Copilot副驾走向主驾，让数字员工成为新常态。

以生成式AI为代表的人工智能技术迅猛发展，为AI领域突破提供了新的通用化解决方案，使得AI技术大规模普惠落地成为可能，不仅加速了与各行各业场景的深度融合，还掀起一场应用的AI革命，带来产品形态、开发模式、价值理念的一系列全新变化。

通用AI践行渐进，大模型走向多模态，解析世界本来面貌，加速AI从感知到认知转化。AI智能体将成为下一代平台，从Copilot副驾走向主驾，让数字员工成为新常态；具身智能加速进化，AI有望完成“感知-决策-行动”的闭环。

为让大模型技术为企业发挥真正作用，解决应用落地的最后一公里问题，力维智联推出Sentosa LMM零代码大模型平台，通过算力资源智能调度、大模型预训练与微调、智能体定义与研发、应用敏捷编排等能力，致力于把通用大模型打造成能随企业一起成长的企业大模型，赋能企业全业务提效。

©2024.4 iResearch Inc.

林松涛

拓尔思 副总裁

面向企业级的大模型应用正在开花，基于智能体的服务模式将是主流。

随着不断推陈出新的开源大模型的快速迭代，国内自训的闭源大模型将逐渐收拢，但是围绕大模型在行业和企业级的应用需求在持续探索中不断的发芽开花。作为面向B端应用的企业级大模型，更多的强调内容生成的高质量和真实性，减少幻觉将是行业大模型落地的长期课题。

拓天行业大模型从高质量数据、可控生成和信创安全等多角度出发，融合自研的海贝向量数据库、利用RAG和知识图谱等技术将大-小模型与行业权威数据源结合，实现了内容生成的合规可控。在业务融合角度，拓天基于自主演化的任务链，实现知识和数据混合驱动的AI Agent应用框架。围绕B端用户的应用场景，根据不同的输入需求，使用大模型Agent自主构建对应结构的工具链。

未来，人工智能在千行百业有效落地，需要不断进化的模型基座与业务知识的集成，我相信以可控规模参数的大语言模型所迭代的应用将会指数成长，AI智能体将成为主流交互模式，而工程化能力则是落地的重要保障。

©2024.4 iResearch Inc.

张杰

中关村科金技术副总裁

随着技术的不断成熟，大模型的应用场景将变得更加多样化，尤其是数字员工或智能助手类的应用场景商业机会将会非常大。

今年，政府工作报告中提出了“人工智能+”行动，明确了国家层面对于推动人工智能与各行各业深度融合的战略意图。这一行动不仅释放了国家对人工智能技术重视的信号，也为人工智能和大模型领域带来了前所未有的发展机遇。

当前，大模型呈现出快速迭代和广泛应用的特点，技术的进步为行业带来了无限可能。从长远来看，通用大模型确实能给行业带来巨大的价值，但目前来看，通用大模型不能满足企业对专业性、合规性、规模化的需求。相反，领域大模型不仅不需要依赖特别大的算力和参数，还能切实解决领域内细分场景的问题。

作为领先的对话式AI技术解决方案提供商，中关村科金率先布局大模型技术和应用，发布了国内首个企业知识大模型、AgentGraph应用开发平台以及“超级员工”系列AIGC应用，全面升级云呼叫中心、智能客服、智能外呼、质检陪练、智能音视频等产品。未来，随着技术的不断成熟，大模型的应用场景将变得更加多样化，尤其是数字员工或智能助手类的应用场景商业机会将会非常大，从提升工作效率到助力产业升级，大模型将成为推动新质生产力发展和社会进步的重要力量。

©2024.4 iResearch Inc.

张伟贤

活树科技 创始人

在当今数据驱动的时代，高质量/多语种的数据不仅是大型AI模型开发的基础，也是推动整个AI行业向前发展的关键力量。

在当今快速发展的人工智能领域，数据的质量、多样性和有效管理被认为是关键因素，这些因素共同决定了大型模型的性能和可靠性。准确和相关的数据集是训练高效AI模型的基础，任何数据偏差或错误都可能削弱模型的可靠性和有效性。此外，持续的数据验证和清洗是确保数据真实性的重要步骤，它们帮助在数据驱动的决策过程中维护信任与透明度。

数据的多样性对于增强模型的泛化能力至关重要，尤其是在不同的文化和地理背景下。一个多元化的数据集有助于模型在多样化的实际应用环境中表现出更好的稳定性和公平性，从而减少潜在的算法偏见。此外，随着数据量的激增，高效的数据存储、处理和保护措施变得尤为重要。现代工具如数据湖和云服务的运用不仅提高了数据处理效率，也加强了数据合规性和隐私保护，这对于提升公众对AI应用的信任尤为关键。

活树科技公司，凭借20余年的深厚经验，已在全球16国构建了覆盖50+语种的数据采集网络。公司的全球数据平台涵盖多种语言和文化，确保数据多样性与质量，从而加强AI模型的适应性与智能。活树的先进工业化流程和高标准数据产品，不仅提升了全球AI技术的公平性和普及率，也推动了技术创新与社会共赢的发展目标。

©2024.4 iResearch Inc.

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。

iResearch
艾瑞咨询

THANKS

艾瑞咨询为商业决策赋能