iResearch
艾 瑞 咨 询

2023年中国AIGC产业全景报告

日就月将，学有缉熙于光明

部门：企业服务三组　　署名：王祺 李冬露 张云 李鑫

PREFACE

前言

研究背景：

2023年4月，艾瑞发布《AIGC系列报告——ChatGPT专题》，对OpenAI的研发路径及商业模式、ChatGPT下游应用场景、大语言模型（LLM）对AI产业影响、中国LLM产业价值链等话题进行了分析，初步探讨了AIGC时代下中国大语言模型产业的价值空间与发展方向。

古人有云：日就月将，学有缉熙于光明。人类对人工智能学的潜心钻研终于再度获得重大突破，大模型的涌现能力与AIGC的应用普及为那不一定是AGI但一定更AI的未来提供了确定性的加速度。AI2.0时代的加速到来，不仅是把AI能力融入到现有应用中，更是未来产业范式的再塑造。AI正跳跃式地加速渗透进各行各业，推动一场新的生产力与创造力革命。AI产业链各环节参与者的角色功能、产品服务和应用生态可能将发生变化。

2023年8月，艾瑞发布《AIGC系列报告——中国AIGC产业全景报告》，作为AIGC系列首发，报告将展开对AIGC产业的全景洞察、探究生成式AI技术对数字产业的影响变化、绘制“中国AIGC产业全景图谱”、分析主流参与厂商类型与格局策略、各类型厂商发展路径和能力要求变化等，为市场辨析产业发展价值与空间。

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

发展总览

AIGC技术作为新型内容生产方式，将以内容生产模式变革催动生产力革新，引领数实融合浪潮下的产业变革，对人们生产生活方式带来深远的影响，开辟人类生产交互新纪元。艾瑞咨询预测，2023年中国AIGC产业规模约为143亿元，随后进入大模型培育期，持续打造与完善底层算力基建、大模型商店平台等新型基础设施，以此孕育成熟技术与产品形态并将其对外输出。中国AIGC产业生态将日益稳固，完成重点领域、关键场景的技术价值兑现，逐步建立完善模型即服务（MaaS，Model As a Service）产业生态，2030年中国AIGC产业规模有望突破万亿元，达到11441亿元。

大模型层

大模型是AIGC技术变革的原生驱力。大模型的落地将提速AI工业化生产，并充分释放AI产业潜在市场空间，带来新一轮AI产业化扩散。从商业化路径来看：1）MaaS是大模型能力落地输出的新业态。2）闭源与开源市场将并存互补，呈现“轻量级模型陆续开源，助力开源生态建设，千亿级模型暂以闭源路径开展”的发展特征。3）基模落地因需求差异展开产业路径分化，以行业级、企业级大模型方式支撑上层应用。4）数据准备、ROI衡量、 Prompt工程是连接模型层与应用层的落地三要素，工具层成为AIGC产业新热点。

应用层

应用层是AIGC技术价值传递的实际落位，将通过对内容生产方式和人机交互方式的改变，深刻影响个人的生产与消费生活。对比国外，我国在开源生态、付费能力和创新力等方面的差距是AIGC应用发展必须面临的挑战。AIGC应用可分为个人消费和企业服务两个赛道。在个人消费领域，AIGC将以消费级内容和内容创作工具为载体，率先通过UGC进行产业渗透，垄断内容分发的各大流量、社交、视频平台将作为本轮变革的核心，借助AIGC内容与工具进行商业模式创新。在企业服务领域，AIGC技术在SaaS、决策AI、生成AI等多个领域的赋能路径已初步明朗，而在商业价值上，引入AIGC技术能为AI厂商带来显著降本效果，同时厂商借助AIGC技术能满足客户更多场景化需求，带来营收的第二曲线增长。

算力层

算力层是AIGC发展不可忽视的资源引擎。在OpenAI的GPT模型涌现能力后，AI产业迅速进入以大模型为技术支撑的AIGC时代，巨量训推算力需求让本就供需不平的算力产业结构进一步承压。算力产业模式将在AIGC时代有所演变，智能算力资源或将更多承载于云服务产品，以MaaS模式服务千行百业。大模型时代下，数据中心将进一步优化网络带宽、能源消耗与散热运维等，AI芯片需进一步升级内存、带宽、互联等能力。整体来看，中国正大力推进“东数西算” 工程，引导新型数据中心实现集约化、高密化、智能化建设，并坚持自主创新道路，静待国产替代曙光，实现国产“算力+应用”的正循环。

趋势挑战

从**技术突破**来看，当前Transformer仍具明显优势，但学界和业界都在积极突围，未来Transformer不会是唯一解；从**应用前景**来看，软硬结合、物联网应用升级是趋势，大模型低参化处理后带来全新的手机拍照、语音交互、具身智能机器人应用体验；从**社会影响**来看，AI将成为基础设施，将替代部分专业性岗位，进而带来社会人力结构和分配方式的重塑；从**监管展望**来看，政策鼓励AIGC相关研究，放宽了内容容错率，积极推动公开数据建设，但也强调了AI生成标识、境外服务严格监管等方向，宽松鼓励与整顿规范并存。

CONTENTS

目录

01/中国AIGC产业之“变”与“新”

Overview

报告研究范围 - AIGC

AIGC与大模型将引领“AI产业”与“产业AI”发展

AIGC（AI-Generated Content）指利用人工智能技术（生成式AI路径）来生成内容的新型内容生产方式。2022年11月上线的AIGC应用ChatGPT，凭借其在语义理解、文本创作、代码编写、逻辑推理、知识问答等领域的卓越表现，以及自然语言对话的低门槛交互方式，迅速获得大量用户，于23年1月突破1亿月活，打破前消费级应用的增速记录。ChatGPT等AIGC应用在多个领域的问题解决能力已超出一般人类水平，微软称其在GPT-4（ChatGPT Plus背后运行的大模型）中看到了AGI（通用人工智能）的雏形。大众的生活工作日常出现了Midjourney等新形态的各类AIGC应用，各行业的智能化升级也看到了新的可能性，“AI产业”与“产业AI”的想象空间进一步拓展。**AIGC应用创新的技术支撑为“生成对抗网络（GAN）/扩散模型（Diffusion）”与“Transformer预训练大模型”的两类大模型分支，在国外AIGC应用展示出大模型的能量的同时，我国企业也加强了相关产品技术布局**，云厂商、AI大厂、创企、各行业公司及技术服务商等产业各领域玩家纷纷发布大模型或基于大模型的应用产品及各类技术服务。相较于一般AI应用，大模型应用的训练及推理需要更强的算力支持。综上，本报告将围绕模型、应用、算力三个角度对AIGC产业的发展进行探讨，试图在讨论开源闭源、垂直通用、知识幻觉等大模型未来发展的各种不确定性的同时，为AIGC应用的迭代升级、产业的智能化应用，提供尽可能多的研究辅助，为那个不一定是AGI但一定更AI的未来提供确定性的加速度。

生成式AI显现通用人工智能雏形

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

本报告主要研究范畴

来源：艾瑞咨询研究院自主研究绘制。

中国AIGC产业发展环境-政策 (Politics)

以包容审慎的态度，支持引导AIGC“可靠、可控”发展

为促进AIGC产业健康发展、规范应用，央地各级政府围绕算力、数据、模型、应用等不同方面逐渐完善支持政策体系，且国家层面快速出台聚焦AIGC的合规监管政策。支持政策方面，**以完善算力与数据等要素供给为基础，以模型算法创新为关键，以场景应用为牵引，构建活跃的AIGC创新与应用生态。**分区域来看，以北京为代表的AIGC创新及产业要素聚集地在政策层面支持力度更大。合规监管政策方面，《生成式人工智能服务管理暂行办法》奠定了我国对于AIGC**包容审慎、分级分类监管**的主基调，明确生成式人工智能服务提供者应承担网络信息安全、个人信息保护等义务，提出需进行安全评估与备案、对生成内容进行标识等服务规范。

中国AIGC产业政策分析

中国部分AIGC产业相关政策

政策名称	发文单位	发文时间	类别
《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》	科技部等六部门	2022-08-12	支持类
《互联网信息服务深度合成管理规定》	国家网信办等三部门	2022-11-25	监管类
《中共中央 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》	国务院	2022-12-19	支持类
《生成式人工智能服务管理暂行办法》	国家网信办等七部门	2023-07-13	监管类
《北京市通用人工智能产业创新伙伴计划》	北京市经信局	2023-05-19	支持类
《北京市促进通用人工智能创新发展的若干措施》	北京市政府办公厅	2023-05-30	支持类
《成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施》	成都市经信局市新经济委	2023-08-04	支持类

支持引导类政策-强化基础资源，营造应用生态

应用

- 支持方式：**开放政策性场景资源；建设场景应用试点、场景实验室；发布场景机会清单实施揭榜挂帅；评选场景应用示范项目**等。
- 重点领域：政务（城市治理）、交通、医疗、金融、科研、商贸、教育、文旅、养老等社会重点领域应用。

模型

- 支持通用大模型与行业模型的开发，并给予专项奖励；
- 支持企业、高校院所等建设**开源社区（平台）**...

数据

- 支持**训练数据集、标准测试数据集**等数据资源的建设；
- **加快数据要素市场建设**，推进数据分级分类共享、交流、交易；
- 建设**数据安全管控体系**...

算力

- 建立统一的**多云算力调度平台**，增强算力统筹能力；
- 支持大型云厂商等市场化企业**建设商业算力基础设施；**
- **推动大型公共算力中心建设**...

合规监管类政策-包容审慎、分级分类监管

明确生成式人工智能服务提供者应承担的责任与义务：

- 依法承担网络信息内容生产者责任，履行网络信息安全义务
- 依法承担个人信息处理者责任，履行个人信息保护义务...

基于提供者的责任与义务，提出服务规范：

- 依法进行安全评估申报与备案
- 采取措施提高训练数据质量并保障训练数据安全
- 依法对相关生成内容进行标识
- 不得收集非必要个人信息...

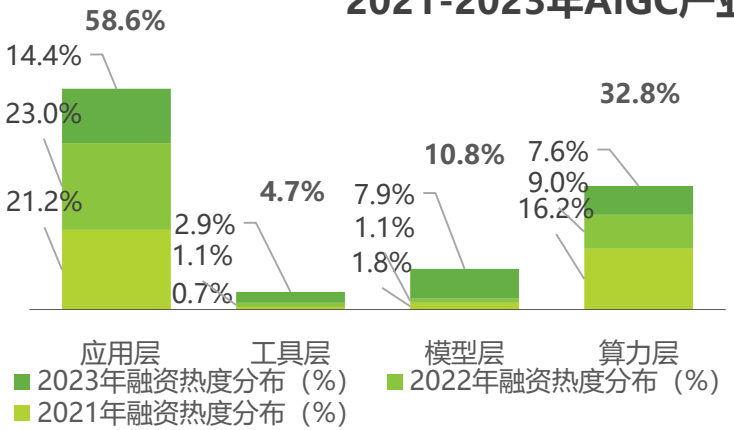
来源：艾瑞咨询研究院根据公开政策资料研究及绘制。

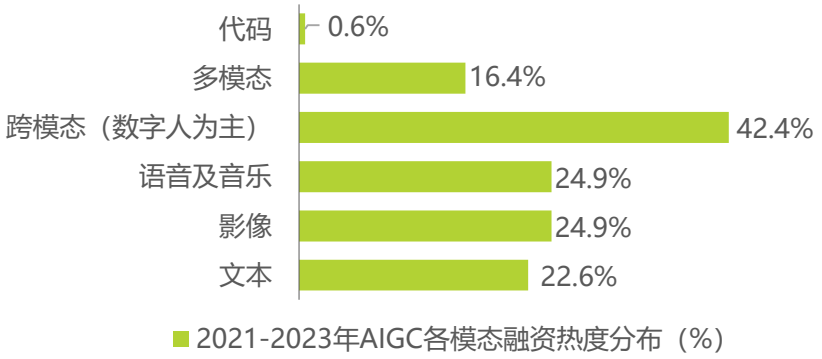
中国AIGC产业发展环境-经济 (Economy)

模型创业初抬头，多模态、跨模态备受青睐，资本扎堆优质项目

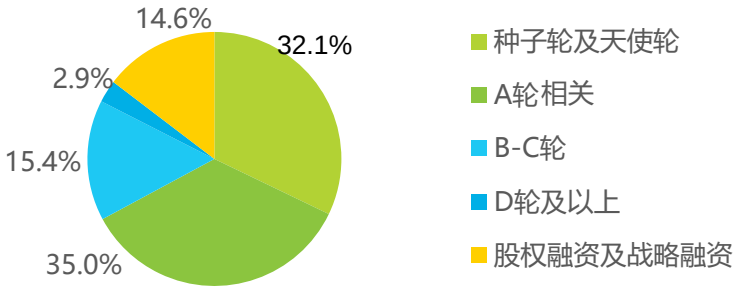
2021年至2023年7月期间AIGC赛道共发生280笔投融资，展现了其高热度与高成长性。从产业细分维度，应用层创业机会最多，模型层创业受到ChatGPT影响，在2023年集中涌现；在获投的应用与模型层创业项目中，文本、影像、语音平分秋色，但相比单一模态，**多模态和跨模态的应用前景更加为资本所看好。**从投融资轮次维度，70%左右的项目仍处于A轮及以前阶段，同时有高达14.6%的比例属于股权、战略融资，说明**赛道虽然处于起步期，但其战略价值已被公认。**在全部获投的170家公司中，获投3次及以上公司约占17%。**同一标的的高频融资，从企业需求侧反应AIGC前期创业所需大量资金支持，从资方供给侧表明优质创业项目仍非常稀缺。**

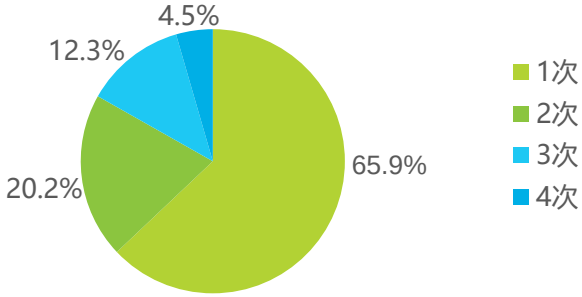
2021-2023年AIGC产业链各环节及各模态融资热度情况





2021-2023年AIGC产业总体及各公司融资轮次分布情况





注：数据截止2023年7月31日

来源：IT桔子；艾瑞咨询研究院自主研究绘制。

中国AIGC产业发展环境-社会 (Society)

引领数实融合新浪潮，以内容生产模式变革为根本引爆生产力革命

受惠于各行业不断丰富的数据资源、算力硬件资源的持续发展以及大模型技术的突破性发展，AIGC得以更好的抽象来自于真实世界的多模态数据源并进行有效表达，展现出其作为内容生产的通用工具在各行各业大规模应用的巨大潜力。**放眼未来，AIGC将以内容生产模式变革催动生产力革新，引领数实融合浪潮下的产业变革，对人们生产生活方式带来深远的影响。**一方面，AIGC将革新数字内容产业的发展范式，增加内容生产的价值和影响力。另一方面，AIGC将加速产业数字化进程，改善实体经济对于数据资源的应用模式与利用效率，赋能实体经济实现数智化转型。更进一步，AIGC将极大地激活数据要素潜能，更广泛地拓展数实融合空间，促进数字经济与实体经济的深度融合，数字产业化和产业数字化的范围将持续扩大交融，实体经济整体上出现创新驱动和结构升级的路径特征。放眼未来，随着实体经济中更多领域加速数字化进程，实体经济体系将进一步完成数字化效率变革。作为现阶段AI产业的排头兵，AIGC对生产力的革新，将一定程度引领产业涌进从IT化、互联网化到智能化的第三阶段数实融合浪潮。

AIGC将引领数实融合浪潮下的产业变革

AIGC大范围应用的奇点已经来临

底层大模型技术的快速迭代发展，支撑AIGC以内容生产模式的变革，引爆生产力革命，激发AIGC向社会经济生产活动广泛渗透、引领数实融合浪潮下产业变革的潜力。

革新内容产业

AIGC可降低数字内容生产的成本和门槛、拓展数字内容生产的空间和维度，从而创新数字内容生产的流程和范式，快速升级甚至颠覆以内容生产为核心的游戏、影视等行业的发展范式。

赋能实体经济

AIGC可以强大的信息获取能力、数据处理能力、逻辑推理能力、内容创作能力，辅助或承担部分繁复的基础性内容生产工作，**一定程度上解决b端边际成本和碎片化问题，加速产业数智化转型。**

孕育新赛道

AIGC可根据不同行业的特点和需求，定制化生成适应性强、价值高的数字内容，**促进数字内容与其他产业的深度融合，使数字经济与实体经济形成良性互动**，孕育新的爆发式增长赛道。

- **游戏**
  - 智能对话系统
  - 高效场景生成
  - …
- **影视**
  - 自动化剧本生成
  - 高效率后期制作
  - …
- **电商**
  - 沉浸式购物体验
  - 高效营销活动创作
  - …
- **交通**
  - 智能车载系统
  - 智慧交管系统
  - …
- **医疗**
  - 电子病例生成
  - 合成医护陪伴
  - …
- **教育**
  - 高效教案设计制作
  - 智能化教学测评
  - …

来源：艾瑞咨询研究院根据公开资料研究及绘制。

中国AIGC产业发展环境-技术 (Technology)

各模态生成质量均初步达到应用水平，可控性成为最大短板

AIGC技术可按照模态分为文本、图像、语音以及多模态等。音频生成技术成熟度最高，其余各模态技术发展稍缓，核心算法仍存在大面积黑箱，虽然在生成效果上整体能够达到人类平均水平，部分场景达到人类优秀水平，但在算力成本、生成稳定性、个性化精细化需求满足等方面存在明显瓶颈。如大部分AI生成图像目前无法支持画师对细节进行精细化的修改，文本生成内容仍会出现事实性错误，因而目前无法达到大规模成熟应用水平。从技术迭代速度看，各模态呈现出成熟度越低，迭代速度越快的特点，文本和图像生成领域几乎每1-2个月就能出现突破性技术进展，未来可期。

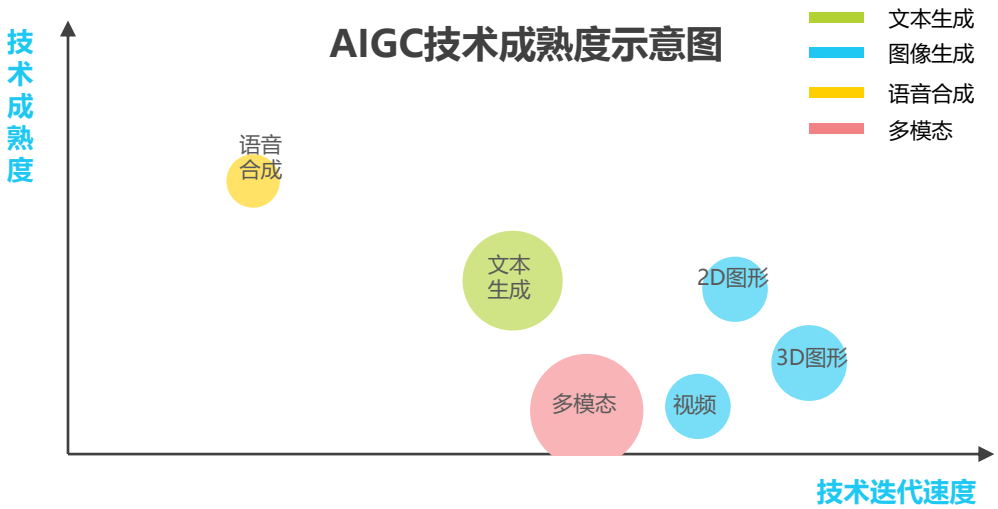
AIGC各模态技术成熟度分析

文本生成：
文本生成可细分为对话、提炼总结、创作等多种能力，以GPT-3为代表的大语言模型将AI文本生成能力拉升至接近人类平均水平。当前在语言理解能力、稳定性、可控性和安全性上还存在非常明显的应用瓶颈，且具体技术路径及对应的训练效果也尚无定论。

图像生成：
图像生成可分为2D、3D及视频生成三个分支。在GAN、dufffusion、NeRF等模型框架基础上，图像生成技术正在以超乎想象的速度迭代升级，部分散点应用已取得良好效果，在**生成精准可控性**及**算力成本**等方面仍需优化。

语音合成：
语音合成技术经历了拼接合成、参数合成和端到端合成三代技术演进，已实现规模化应用，当前技术升级方向在于通过提升语音的**韵律感**和**情绪表达**等达到完全**拟人**的效果，以及实现语音**实时合成**。

多模态：
多模态技术是文本、图像、语音等模态之间的转换与融合，技术难度最大，当前成熟度水平较低，仅在文-图，文-语音等领域有达到商用水平的表现。



技术成熟度评价维度

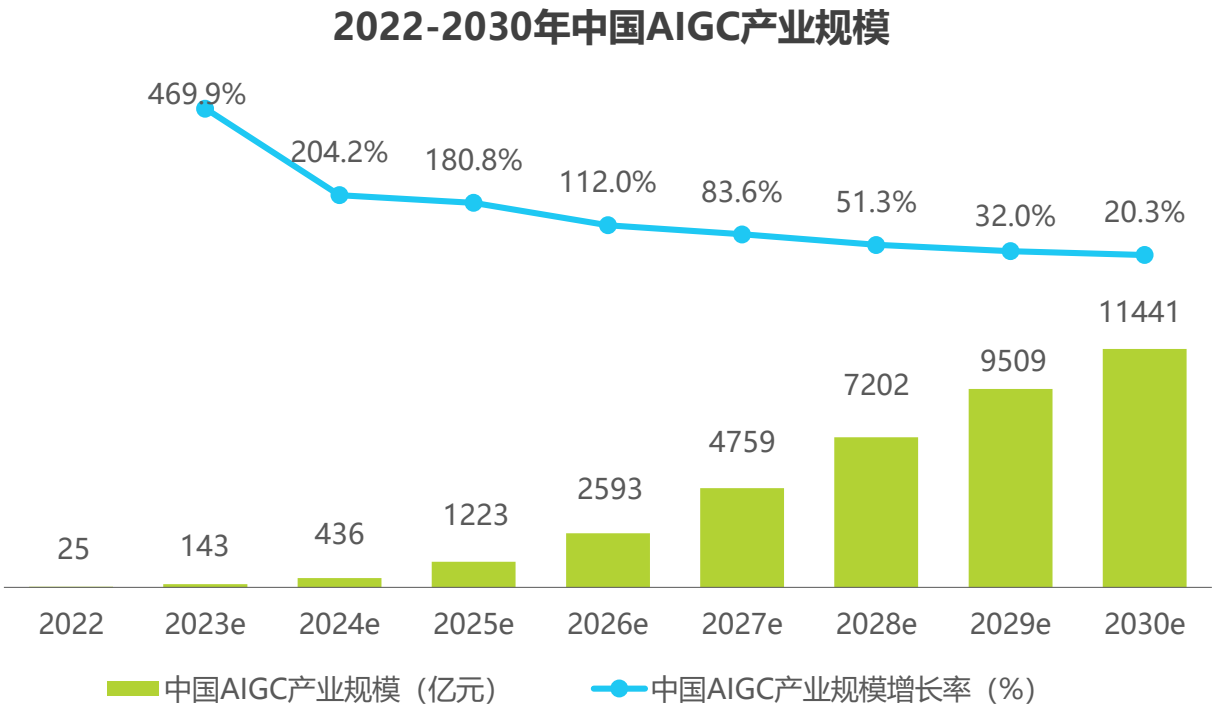
- 生成内容质量
- 个性化生成与定向微调效果
- 技术稳定性
- 技术可解释性
- 应用成本（生成时间周期、部署难度、算力成本等）

注：气泡大小代表该项技术的预期影响力。
艾瑞咨询自主研究绘制

中国AIGC产业市场规模

市场规模呈指数级增长，突破规模化临界点攫取万亿产业价值

根据第50次《中国互联网络发展状况统计报告》，截至2022年6月，中国互联网普及率已高达74.4%。**在网民规模持续提升、网络接入环境日益多元、企业数字化进程不断加速的宏观环境下，AIGC技术作为新型内容生产方式，有望渗透人类生产生活，为千行百业带来颠覆变革，开辟人类生产交互新纪元。**艾瑞咨询预测，2023年中国AIGC产业规模约为143亿元，随后进入大模型生态培育期，持续打造与完善底层算力基建、大模型商店平台等新型基础设施，以此孕育成熟技术与产品形态的对外输出。**2028年，中国AIGC产业规模预计将达到7202亿元，中国AIGC产业生态日益稳固，完成重点领域、关键场景的技术价值兑现，逐步建立完善模型即服务产业生态，2030年中国AIGC产业规模有望突破万亿元，达到11441亿元。**



来源：艾瑞咨询研究院根据公开资料、专家访谈自主研究绘制。

中国AIGC产业图谱全景图

iResearch 艾瑞咨询

2023年中国AIGC产业图谱

下游 / 中游 / 上游

AIGC应用层 Application for AIGC

内容消费赛道

创作工具赛道

企业服务

- 金融
- 零售
- 政务
- 电力
- 医疗
- 企业内部运行
- 数字化与大模型方案提供商
- AIGS (AIGC+软件生成)

AIGC大模型层 Models for AIGC

基模分化

行业垂直型基础大模型：金融 | 医疗 | 电商 | 建筑 | .....

业务垂直型基础大模型：智能问答 | 病例生成 | 设备运检 | 企业服务 | .....

通用基础大模型

- 闭源厂商：OpenAI | GPT | 阿里云 | 通义 | 百度智能云 | 文心 | 4Paradigm | 式说；腾讯云 | 混元 | 华为云 | 盘古 | 云知声 | 山海 | 科大讯飞 | 星火
- 开源厂商：Meta | Llama | 智源研究院 | 悟道 | MOSS；百川智能 | Baichuan | 阿里云 | 通义 | ChatGLM

AI开源社区：Hugging Face；阿里云 | 魔搭；百度智能云 | 飞桨

AIGC工具层 Tools for AIGC

AI Agents：AutoGPT | LangChain；澜码科技 | Vanus AI

模型平台/模型服务

- 阿里云 | 灵积模型服务
- 百度智能云 | 千帆大模型平台
- 火山引擎 | 火山方舟
- SoftStone | 软通动力天璇
- FIXIE | DUST

AIGC基础层 Infrastructure for AIGC

算力基础

- AI芯片（异构Fabless）
- 智能服务器：浪潮 inspur | 中科曙光 Sugon | 华为 HUAWEI
- 智算中心：企业自建智算中心 | 城市智算中心
- 智能云服务：阿里云 | 华为云 | 腾讯云 | 天翼云 | aws | 百度智能云 | 浪潮云 | 移动云 | 金山云 | 联通云 | 京东云 | Azure

数据基础

- AI基础数据服务：百度智能云 数据众包 | speechocean 海天瑞声 | appen | 标贝科技 | 数据堂 Datatang | Scardust | 云测数据 | MAGIC DATA
- 向量数据库：zilliz | Pinecone | Weaviate | TensorDB | Qdrant | 拓数派
- 数据集：公共开源数据集 | 高校数据集 | 企业私有数据集 | 政府数据集

算法基础

- AI算法框架：TensorFlow | PyTorch | Keras | Spark MLlib | [M]昇思 | 飞桨 | 旷视天元 MegEngine | mxnet | Caffe2
- AI开发平台：百度大脑 | 阿里云 | 腾讯云 | 华为云 | 火山引擎 | 讯飞开放平台 | HIKVISION | JDT 京东科技

中国AIGC产业机会前瞻

技术变革：模型层>工具层>算力层>应用层；资源要素：算力层>模型层>应用层>工具层；市场机会：应用层>工具层>模型层>算力层

2023年中国AIGC产业全景总览及机会前瞻

应用层

To C应用试水，合规性与付费意愿等要素限制有规模化难度；To B应用将在数字化基础做进一步渗透扩张，场景边界仍在探寻

- **C端洞察：AIGC进一步下放内容创作权，极大激发用户创作热情，加速内容裂变，并带来一系列AI-Native的新生机会。**从内容/社交平台角度出发，以社区形式通过用户自发创作交流形成粘性是未来发力方向。而国内用户在SaaS服务上仍是较低付费意愿和购买力，如何聚集流量、从尝鲜行为转为深入重复使用且满足强监管要求是C端运营难点。
- **B端洞察：在产业服务中，AIGC将从内容生产与交互方式改变企业数字化产品服务。**AIGC在B端应用推广与企业自身的数据基础、上云进程、数字化进度等息息相关。艾瑞从供需两侧访谈了解，目前B端AIGC应用正处于场景探索期，双方正努力搭建标杆业务场景与典型行业模型，共同推广AIGC技术的应用渗透。纯应用开发技术门槛的降低将数据要素与行业know-how的重要性置顶，拥有垂类数据积累与业务理解的B端厂商可利用AIGC赋能升级获得进一步增量空间。

模型层

MaaS是大模型能力落地输出的新业态，模型层将更贴近应用侧，工具链完完善度影响用户体验，进一步催生工具层发展

- **能力输出业态：大模型成为未来AI产业的操作系统，带来“以云计算为基础，将大模型作为一项服务提供给用户使用“MaaS模式的新业态**，重构AI产业链价值流通环节和技术传递路径。
- **模型路径演变：基础大模型落地会因需求差异展开产业路径分化，以行业级、企业级大模型方式支撑上层应用。**从开闭源角度来看，基模厂商普遍采用轻量级开源、千亿级闭源”的发展路径，而向上分化的垂直领域厂商将基于开源模型或基模平台开发部署细分领域模型产品，厂商优势在垂类数据与业务理解。

工具层

- **AI Agent与大模型服务/平台是AIGC时代下新衍生的工具层，已成为继大模型之后，更有想象空间却也更贴近应用的下一爆点。**对于AI Agent来说，将宝贵的垂类数据与业务理解集成到Agent框架之中，保证大模型应用在执行任务时可以访问到正确的信息并高效执行产出，是未来AI Agents能发挥出实际效用的关键。随着大模型工程化能力提升，模型服务定位的人才及资源投入需求将降低，市场机会不明朗。而模型商店/平台将呈现明显双边效应，技术资源聚集及应用生态搭建是关键。

算力层

带动算力基础设施建设，大模型运行对其提出更高要求

- 作为数智化时代的资源引擎，算力正逐渐成为影响国家综合实力和经济发展的关键性要素。随着AIGC时代大模型参数的量级提升，算力供需结构承压持续加大。训推算力需求先呈指数级上涨。顺应先训练后推理逻辑，未来仍有巨量边缘及端侧算力需求待释放。**预训练大模型的训练推理需要巨量数据资源与高性能计算机的全天候高速运转，对数据中心的网络带宽、能源消耗与散热运维能力，AI芯片的内存、带宽、互联能力、软硬协同均提出更高要求，极大影响算力利用率与芯片性能发挥。**MaaS将云计算、智能算力、模型能力等资源实现高度融合，**艾瑞判断，未来智能算力资源或将更多承载于云服务产品，以MaaS模式服务千行百业**，而随着大模型轻量级开源版本的发布，大模型有望进行进一步剪裁优化，将推理能力部署在端侧，并带动手机、机器人等端侧芯片发展。

来源：艾瑞咨询研究院根据公开资料与专家访谈自主研究绘制。

1）大模型将成为AI应用开发的操作系统

模型即服务（MaaS）构建新型AI基础设施，重构AI开发部署范式

AI产业的场景落地一直面临碎片化困境。随着企业上云进程中智能化转型需求的逐步增多和传统行业领域数据的不断积累，AI应用开发过程中逐渐面临大量细分领域的深耕、非典型客户需求，对算法的通用性和延展性提出了较高要求。传统“小模型”范式的AI应用开发流程一般针对单一场景，独立完成模型选择-数据处理-模型优化-模型迭代等一系列开发环节。因此，AI应用在定制化需求、长尾需求下的开发效率较低，且模型精度、性能、可扩展性等指标质量也会受到影响。随着AI产业深入及智能化需求增加，AI在研发门槛及开发效率的问题日益凸显。“预训练大模型”应运而生，其将数据中蕴含的知识通过无监督或者自监督学习方式提取出来，存储在具有大量参数的神经网络模型中。AI应用开发流程转变为，调用通用流程-结合行业经验-解决实际问题。未来，**大模型将成为AI产业的操作系统，其基础设施特性可为AI应用开发做好底座，将AI模型变得可维护、可扩展、可迭代，极大降低AI应用的开发门槛。从需求侧来看，客户能通过更低成本、高效率的MaaS（Model As a Service）路径获得AI能力，完成AIGC应用的个性化开发、优化及部署，持续兑现大模型的技术红利，将AI能力应用渗透到各行各业的场景业务中。**

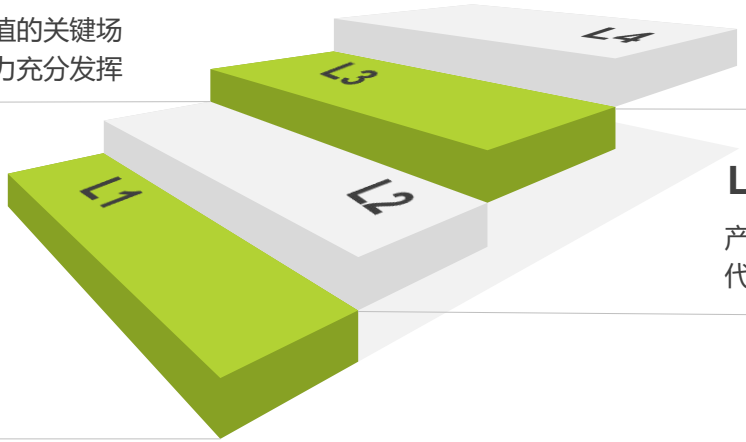
模型即服务（MaaS，Model as a service）范式演进历程

L3：核心场景稳定期

挖掘充分体现其核心价值的关键场景，从而让大模型能力充分发挥

L4：产业生态期

随着业务流程、产业基础设施的发展而完善和融入，模型即服务产业生态建立



L1：大模型成熟期

通用/行业/场景基础大模型的评测指标趋于稳定，是产品和技术持续输出的关键和基础；底层算力基建等基座打造和完善

L2：产品形态成熟期

产品优化，符合创作者使用习惯，可提供低代码或者零代码开发服务

来源：《华为预训练大模型白皮书》，艾瑞咨询研究院自主研究绘制。

2）厂商合作关系演化及周边工具服务发展

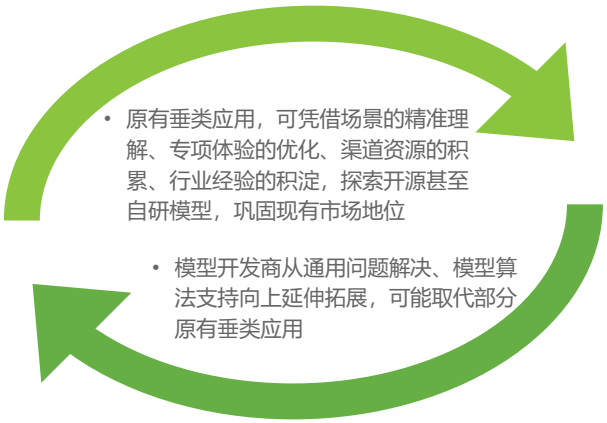
模型层与应用层边界渐弱，带动数据层、开发平台等工具服务高效发展

伴随大模型通用性的提升，模型开发厂商可能因其模型被广泛使用调用汇聚多维场景数据、积累行业场景认知并集成部分垂类功能特性，进而向上延伸拓展至完整功能应用；原有垂类应用，为巩固市场地位，可能探索开源甚至自研模型，凭借既有资源、经验积淀及领域聚焦，同样打造模型开发及应用服务的闭环迭代，因此**模型层、应用层有交错发展之势。此外，企业客户参差的数字化基础及个性化的软件、流程需求依然需要解决方案厂商定制优化并部署实施，而AI开发平台也将与大模型合力，通过“稀疏、蒸馏、剪裁”等手段助力大模型解决训练、推理部署困难问题，进一步实现“低门槛、低成本、高效率”的开发部署与应用。数据标注、安全合规等周边工具服务亦是促进AIGC产品高效开发、产业有序发展的可观商机。**

AIGC厂商合作关系演化及周边工具服务发展

针对企业服务市场，各行业客户的数字化基础及建设的发展规划、具体需求各有不同，在AIGC重构企业软件及业务流程的同时，依然需要方案集成商来贴合具体客户的特定需求优化并部署实施

在模型层、应用层的交错发展中，在部分领域两类公司的原有合作关系弱化，背后是模型开发与应用服务的一体化快速迭代，AIGC产品体验跃升



类比“淘金潮”中售卖铁锹、牛仔裤、水的生意，AIGC产业的模型、应用可视为“金矿”，芯片等算力支持可视为“铁锹”，AI应用开发平台、数据标注、 安全合规、开发平台等周边工具服务亦是促进AIGC产业高效有序发展的可观商机。

AI应用开发平台	提升模型的微调及管理、评估效率，以弹性、稳定、低成本的方式保障模型的训练、部署及运转
数据标注	模型训练基于海量标注数据，AIGC大模型的重要成因之一在于对RHLF（基于人类反馈的强化学习）的使用，需人工对模型生成的多个结果进行排序标注
安全合规	服务商可通过数据集筛查预处理、偏见评估系统、伪造监测算法等为模型及应用的健康发展保驾护航

AI Agent　向量数据库　……

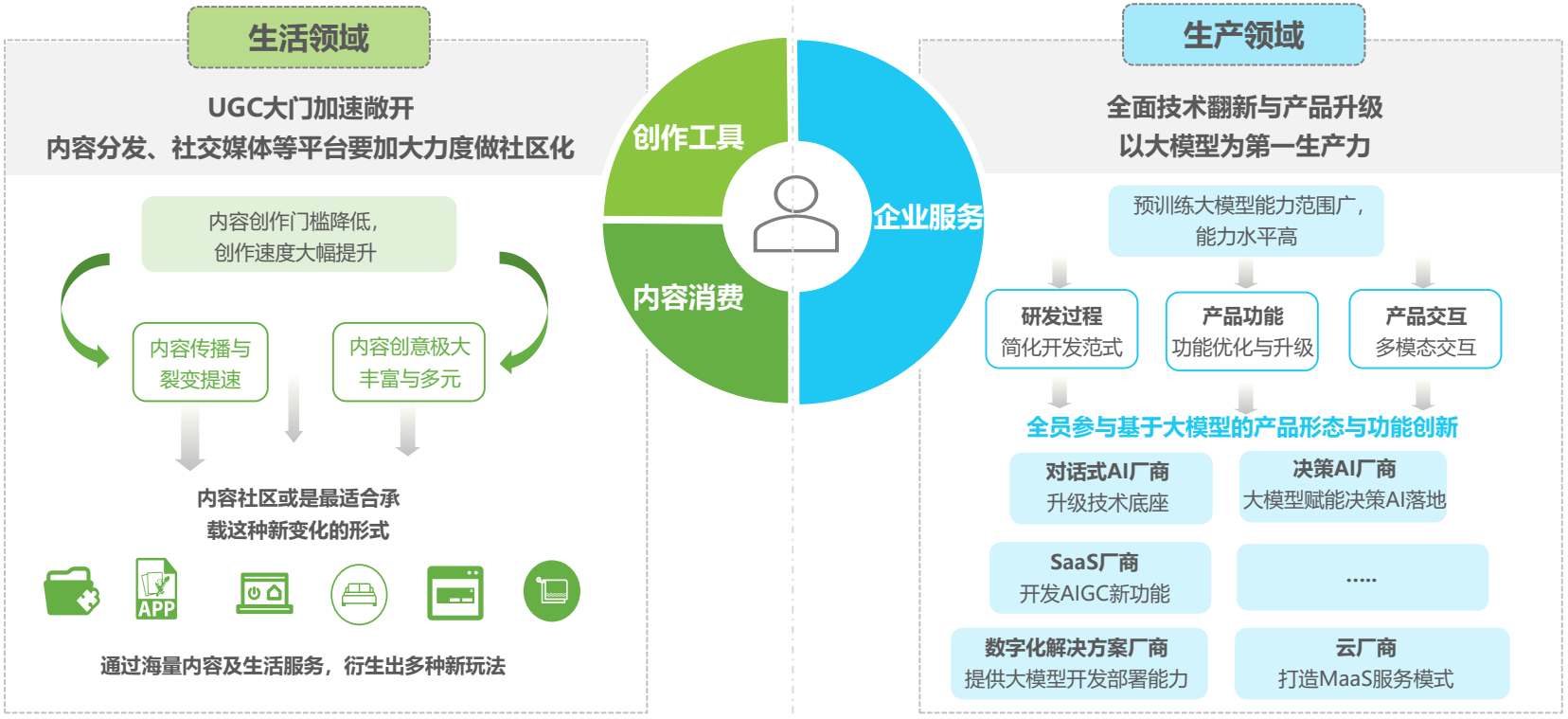
来源：艾瑞咨询研究院自主研究绘制。

3）生产力变革带动海量下游应用优化

生活领域充分释放用户创新能力，生产领域全面革新交互体验和效率

以大模型为标志，生成式AI是一次新的技术革命，同时还具有极强的普适性，能够对人类生产、生活的方方面面进行改造与升级。**在生活领域，AIGC将通过进一步下放内容创作权，激发UGC创作热情，加速内容裂变。**加之社区玩法在部分内容平台的良好盈利表现，内容消费领域从技术到商业模式的路径已全线打通，以社区形式，通过用户自发交流自主创作形成粘性，是各类平台的发力方向。**在生产领域，大模型能从研发流程、产品能力和交互上全方位为企服软件带来提升，也充分开拓了新的服务场景**，因此各类企业数字化厂商都将围绕大模型寻找自身优势空间与定位。

AIGC全面落地应用的影响分析



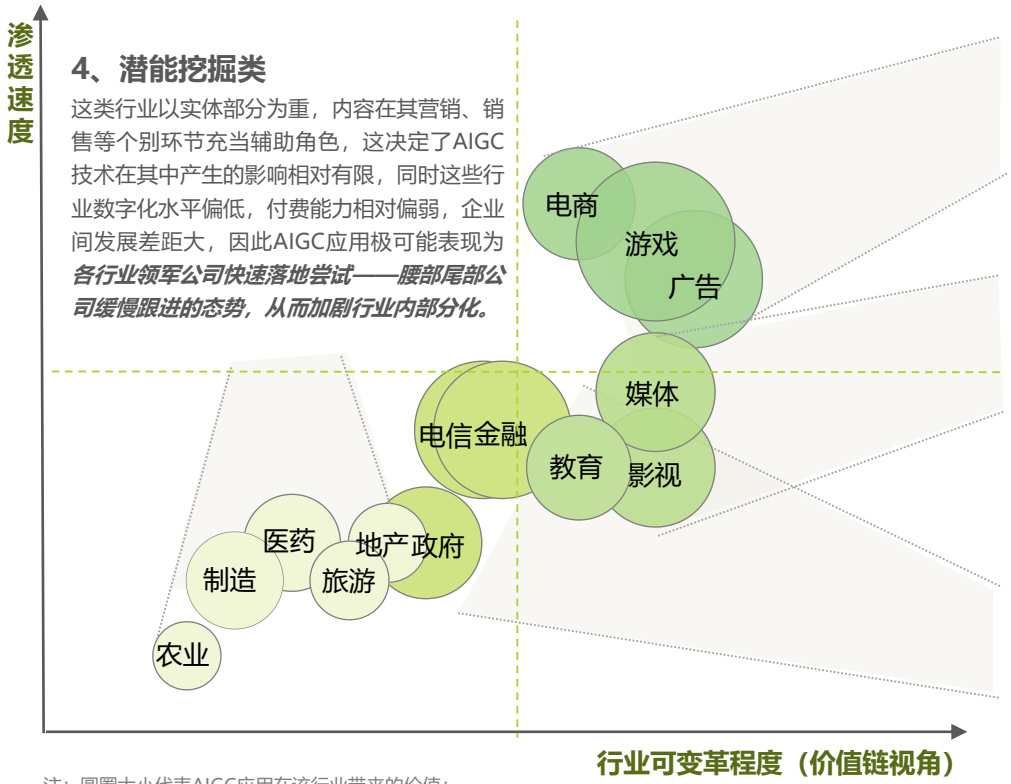
来源：艾瑞咨询研究院自主研究绘制。

4）AIGC将在全行业引发深度变革

线上化程度、数字化基础、行业内容占比等影响AIGC应用前景与渗透速度

总体而言，AIGC主要影响内容创作与人机交互，因此价值链线上化程度越高，内容在价值链中占比越高，AIGC对其颠覆效应越明显；另一方面，行业自身的数据、知识、监管要求等特点也会深刻影响到AIGC技术的渗透速度。比如电商、游戏、广告、影视传媒等以内容生产为价值核心的行业，以及电商、金融等研发设计、营销等环节在行业价值链中地位较高的行业，能够快速看到AIGC应用对原有生产工具的替代和业务流程的变革。

AIGC对各行业影响与变革分析



1、快速颠覆类

电商、游戏和广告行业线上化程度高，且内容质量直接决定其价值创造，这两大特征使得AIGC应用在其中能够产生最大化的价值，并能够迅速渗透至核心生产环节。据统计，AIGC相关应用已经帮助游戏行业在研发制作环节节约**50-70%**人力或时间成本。

行业中大小公司均有机会抢占先机，甚至受到个人开发者的冲击，行业可能面临洗牌。

2、匀速增长类

这类行业也以内容为价值核心，但与第1类区别在于其内容生产更多环节在实体环境完成，具有更强的专业性、灵活性，在这些行业中AIGC可变革上限略低，且会更加考验AIGC技术的成熟度，当前应用多属于单点尝新而非刚性替代，未来市场空间大，但实现大规模行业渗透需要更长时间。

行业格局预计产生局部波动，行业价值链各环节地位排序面临重构。

3、稳中求变类

金融、电信和政府行业价值链中涉及大量内容生产和人与人交互的环节，数字化转型付费意愿和付费能力强。但同时因为其具有业务流程与组织架构的可变性低，对数据安全极为敏感等特质，因此对AIGC应用的态度最为保守。***解决大模型私有化部署的ROI问题是打开这类市场的关键。***

注：圆圈大小代表AIGC应用在该行业带来的价值；
行业可变革程度：在行业价值链中涉及到内容生产和人机交互的环节占比，以及AIGC对相关环节的影响程度；
渗透速度：由各行业政策监管、数字化特别是数据基础建设水平、数据安全要求、行业创新能力等指标构成

来源：艾瑞咨询研究院自主研究绘制。

5）算力资源决定大模型发展高度

中国坚持自主创新道路，大模型为算力设施带来更高要求与发展机遇

顺应大模型趋势，算力需求急剧攀升，算力正在成为影响国家综合实力和经济发展的关键性要素。面对AI“大模型”算力挑战，数据中心会建设大量服务器节点，通过网络构建集群互联协作完成训推任务。若网络带宽不够大、时延不够低，不仅会让算力边际递减，还会进一步增加大模型训练的时间成本。**未来，数据中心需夯实优化算力基础设施建设，积极提升网络带宽、能源散热等方向以应对大模型带来的高运行要求。此外，实现AI芯片的自主性供给，是中国中长期发展算力产业的重中之重。为了适配大模型的训练及推理，AI芯片对其内存、软硬架构协同、片间及片内互联能力等提出更高要求，给国内厂商带来挑战与机遇，**可进一步关注存算一体、Chiplet等技术发展方向。当前国内寒武纪、华为、海光、昆仑芯、燧原等一二线厂商推出的AI推理芯片产品成熟度较高，处于规模化商用进程中；AI训练芯片普遍与国外旗舰产品在性能上存在1-2代际显著差距，会率先在国家智算中心推广应用，并积极与国内互联网大厂适配调整，优化软硬件适配及生态成熟度。整体来看，中国算力层正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”为两条主线发展路径，静待国产替代曙光，实现国产“算力+应用”的正循环。

AIGC浪潮下的中国算力产业分析

数据中心将进一步优化网络带宽、能源消耗与散热运维等能力

- Transformer密度模型加大基础设施的带宽要求，模型训练要求网络架构走更多东西向流量，未来推理逐步加大南北流量比重。
- 电力占据大模型训练成本的近六成，AIGC下的电力消耗问题值得关注，进一步发展如核聚变等能源方向。
- 大算力需求需要更高散热效率。液冷有望替代风冷达到更高效的散热预期。散热预计将从房间级、机柜级、服务器级向芯片级演进，更贴近核心发热源。

AI芯片需进一步升级内存、带宽、互联、软硬协同等能力

- 为了适配大模型的训练及推理，AI芯片要求有更大的内存访问带宽并减少内存访问延迟，因此带动由GDDR到HBM的技术升级，另一方面需要更高的片间互联甚至片内互联能力以满足AI加速器访存、交换数据的需求。
- 面对海量数据，AI芯片异构化需求加大，需要与软件保持更好地一体化协同，让软硬件的功能完成更紧密的加强，以满足目前高速爆发的数据增长。

来源：艾瑞咨询研究院自主研究绘制。

02/ 技术变革的原生驱力
大模型层

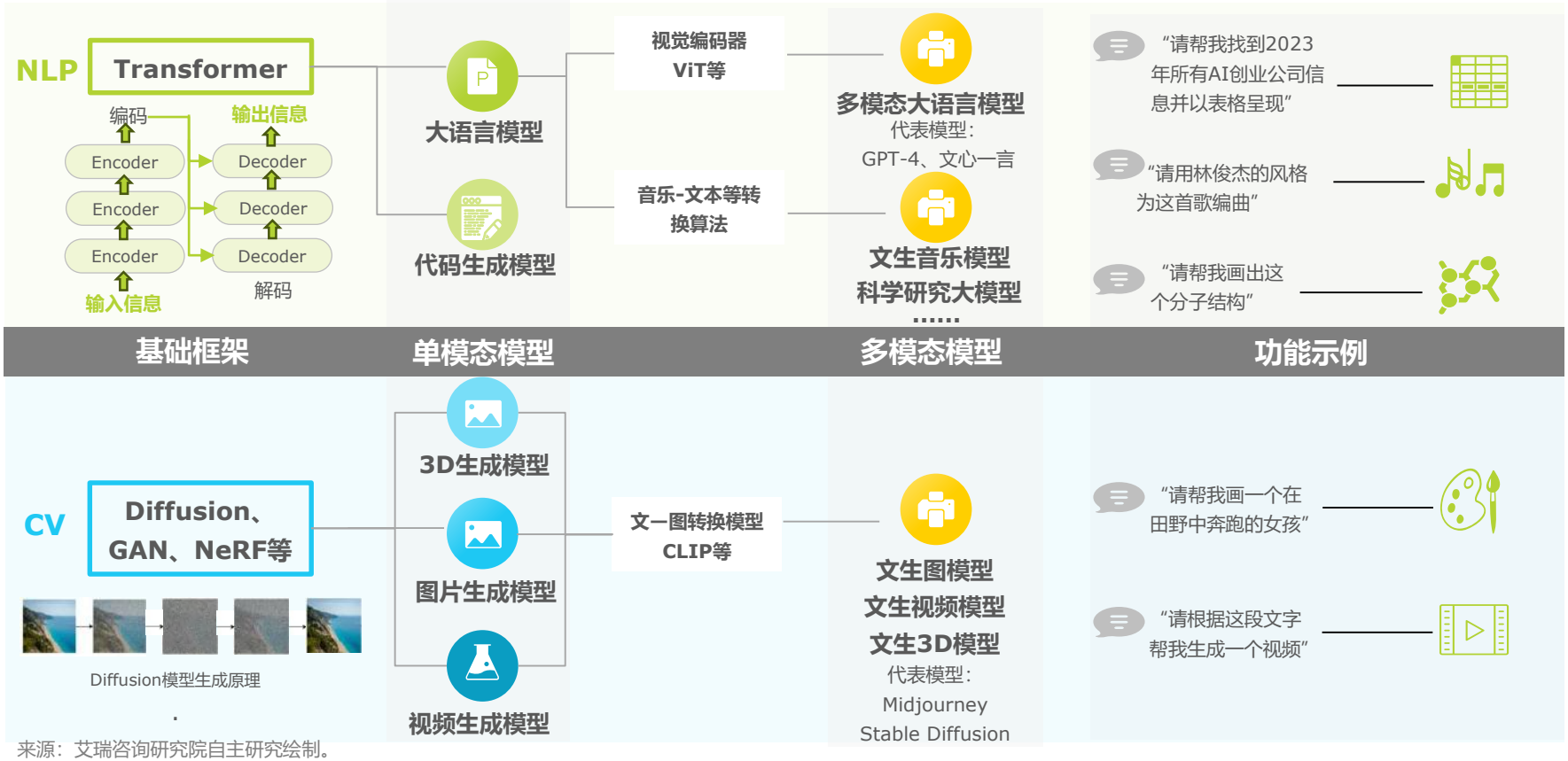
Large Model

预训练大模型分类与介绍

基于两大类基础架构衍生出各类大模型；多模态已成趋势

预训练大模型按照模态可以分为文本、图像、视频、代码、音乐生成等多种，但从底层架构上都分属两类。Transformer是一种编解码模型框架，适用于处理文本、代码这类强连续性生成任务；Diffusion、GAN、NeRF等框架善于处理图像生成类任务。叠加文-图转换技术可以形成文生图模型。由单模态模型在实际训练时融合其他模态技术，可形成多模态、跨模态大模型，如GPT-4、文心一言、 Mid journey等，由于多模态模型可接受文本、图像等不同输入输出形式，对应用场景能够更广泛适配，着力发展多模态模型成为产研两界共同趋势。

预训练大模型各模态技术分支与功能定向



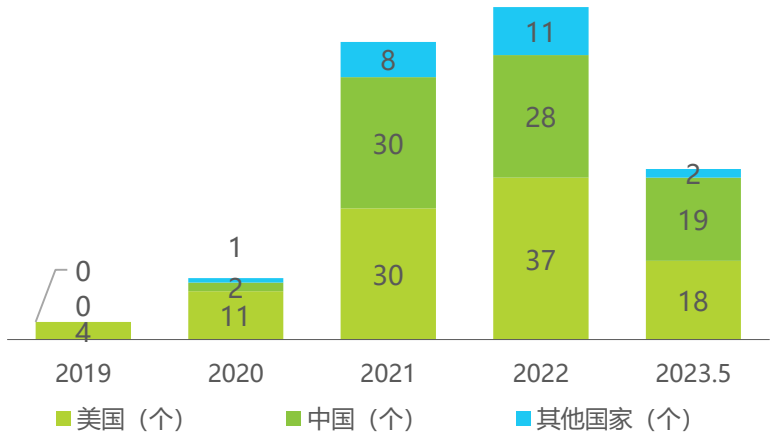
来源：艾瑞咨询研究院自主研究绘制。

预训练大模型的发展业态

中美引领大模型产业发展，NLP仍是大模型的重点活跃领域

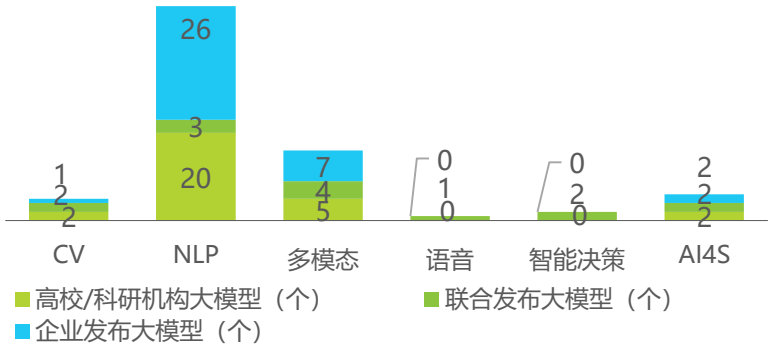
从全球范围来看，中美两国在大模型领域都取得了令人瞩目的成果。2019年，美国率先在大模型领域取得突破性进展，推出了BERT、GPT等具有里程碑意义的预训练模型。2020年，ERNIE系列模型和TinyBERT等轻量化模型的推出拉开了中国大模型产业快速发展的序幕。2021年以来，中美在大模型领域逐渐呈现出分庭抗礼的趋势，共同引领全球大模型产业的发展；聚焦国内，从技术领域来看，国产大模型广泛的覆盖了自然语言处理、多模态、机器视觉等多个技术分支，形成了紧跟世界前沿的大模型技术群。其中，自然语言处理是目前国内大模型最为活跃的技术领域，超六成的国产大模型主要基于自然语言处理技术进行预训练和微调；多模态领域活跃度仅次于自然语言处理技术，超两成的国产大模型可处理图像、视频、音频等多模态数据；而聚焦在计算机视觉和智能语音等领域的国产大模型数量相对较少。从研发主体来看，国内企业、高校、科研机构等不同创新主体均积极参与大模型研发。其中，企业仍是国内大模型研发的主力军，约46%的大模型由企业独立研发；高校及科研机构也对大模型的研发做出较大贡献，约37%的大模型由高校/科研机构独立研发。同时，我们也观察到目前由企业与高校/科研机构联合研发的大模型尚不足20%，展现出大模型开发在产学研合作方面仍有较大潜力。

2019年至2023年全球大模型数量统计



来源：《中国人工智能大模型地图研究报告》，中国科学技术信息研究所、科技部新一代人工智能发展研究中心，艾瑞咨询研究院自主研究绘制。

2023年中国各技术领域大模型数量统计



来源：《中国人工智能大模型地图研究报告》，中国科学技术信息研究所、科技部新一代人工智能发展研究中心，艾瑞咨询研究院自主研究绘制。

预训练大模型的路径探讨

了解人工智能时代的“ios”与“安卓”，闭源与开源市场将并存互补

在以OpenAI为代表的闭源模型厂商开放对外技术服务后，开源模型厂商也在加紧发力，以Meta的Llama模型为代表陆续开源迭代，意图进一步实现生态层面的跑马圈地，2023年上半年LLM与数据集迎来开源季。大模型的开源可根据开源程度分为“可研究”与“可商用”级别。2023年2月，Meta发布了开源大模型LLM的第一个版本Llama，授予“可研究”用途。2023年7月进一步发布“可商用”的Llama2版本，虽然有日活超过7亿产品需额外申请、不能服务于其他模型调优等的商用限制，但海外很多中小企业已可用Llama2的模型来做私有化部署，基于Llama2开源模型训练出定制化的可控模型。由于Llama2基本不支持中文，对中国的大模型商用生态暂时不会产生实质性变化，中国仍需开发培育适配于中文数据土壤的开源生态。闭源LLM可为B端用户和C端消费者持续提供优质的模型开发及应用服务；开源LLM可从研究角度促进广大开发者和研究者的探索创新，从商用角度加速大模型的商业化进程与落地效果。未来，开源和闭源的LLM会并存和互补，为大模型发展共同创造出多元协作的繁荣生态。

中国AIGC产业大模型层开闭源分析

PART1: 盘点中外闭源模型

- **闭源模型**：通过付费的API或者有限的试用接口来访问。目前，**OpenAI的GPT模型、谷歌的PaLM-E模型，及国内阿里、腾讯等互联网大厂的大模型目前均处于闭源状态。**

- **供给侧：AIGC模型层发展中，谁会采取闭源策略?**
大模型技术前沿厂商出于打造自身先进模型壁垒、构建技术护城河的商业考虑，会选择闭源或逐步从开源走向闭源，以**保证模型的先进性、稳定性、安全性**等。

PART2: 开源生态的搭建与意义

- **开源模型**：公开模型的源码与数据集，任何人都可以查看或修改源代码。 如**Stability AI 开源 Stable Diffusion，Meta开源Llama。中国智源开源Aquila，**国内外开源生态愈加丰富。

- **供给侧：AIGC模型层发展中，谁会采取开源策略?**
训练出开源LLM模型背后仍需要大量资金、精力、人才投入；此外，相较于大模型技术前沿厂商，选择开源厂商在技术上仍处于追赶地位，将**通过开源路径培植生态，并追赶优化模型。**

1）可研究类别

2）可商用类别(或进行商用限制/授权收费)

PART3: 需求侧：开源与闭源的选择之虑

闭源模型		V.S.	开源模型	
前期投入成本低	运行稳定		数据隐私安全	迭代更新快	私有化部署
……	完整工具链&工具平台		深度优化&Fine tune	依赖专业团队	……

来源：艾瑞咨询研究院自主研究绘制。

着力打造中国AIGC开源社区生态

轻量级模型陆续开源，助力开源生态建设，千亿级模型暂以闭源路径开展

2023中关村论坛上，科技部副部长吴朝晖表示，中国将坚持开源协作，加强大模型技术持续创新，协同解决透明性、稳定性等共性问题，进一步推动算力资源和数字资源开放共享，加快形成大模型的产业生态。而AIGC开源社区的建设可以吸纳更多的开发者及拥有定义用户的主导权，以AI开源创新平台为杠杆，带动支撑底层AI芯片、智算中心及云服务等基础设施发展。从供给侧逻辑来看，大模型开源早期由高校和机构推动，如清华大学的ChatGLM-6B、复旦大学的MOSS，陆续有头部云厂商加入，如百度的文心系列与阿里的通义系列，共同为中国AIGC开源社区的建设“增砖添瓦”，以阿里云魔塔社区、百度云飞桨社区为代表的开源社区建设成果初现，而千亿级模型暂以闭源路径开展，凭借稳定、优质效、完整工具链等产品特点定位应用市场；从商业化路径来看，参考海外明星开源社区Hugging Face的商业模式，中国AI开源社区同样会先免费提供基础算力，为客户提供免费的社区体验、demo部署及测试，并进一步通过付费服务推送轻量级迁移的微调推理服务或深度开发的训练调优平台，提升模型产品性能，通过开源社区吸引开发者、企业客户完成更多部署应用资源的引流变现。

中外AI开源社区发展洞察

海外开源社区普遍采用“免费+增值”的商业模式

01 GitHub ——代码托管云服务网站

GitHub是一个面向开源及私有软件项目的托管平台，因为只支持git作为唯一版本库格式进行托管，故名GitHub。GitHub是全球最大的开源社区，**允许用户免费创建无限的公共和私有存储库，付费可获得更多功能。**

高级订阅服务 | GitHub应用市场 | GitHub周边商店

02 Hugging Face —— AI/机器学习/NLP界的”GitHub”

Hugging Face是一家以自然语言处理(NLP)技术为核心的AI初创公司,凭借开源项目Transformers（提供了数以千计的预训练模型）积累巨大影响力，并**通过渐进式商业化路径，逐步向SaaS产品和企业服务拓展。**

平衡开源社区与商业化路径 – 商业模式

付费制会员 | 数据托管 | 定制化解决方案

中国AI开源社区建设成果初现，旨在为云服务引流变现

阿里云 – 魔塔社区 | 百度云 – 飞桨社区 | ……

优秀模型聚集交流区 → 模型试用体验 → 微调部署/深度开发

开源模型（以6B 7B 13B轻量级为主）

以开源社区为生态建设，引流资源服务

AIGC开源社区

- 试用体验 Demo部署
- 轻量级迁移 微调+推理服务
- 深度开发 训练+调优+评测 等服务

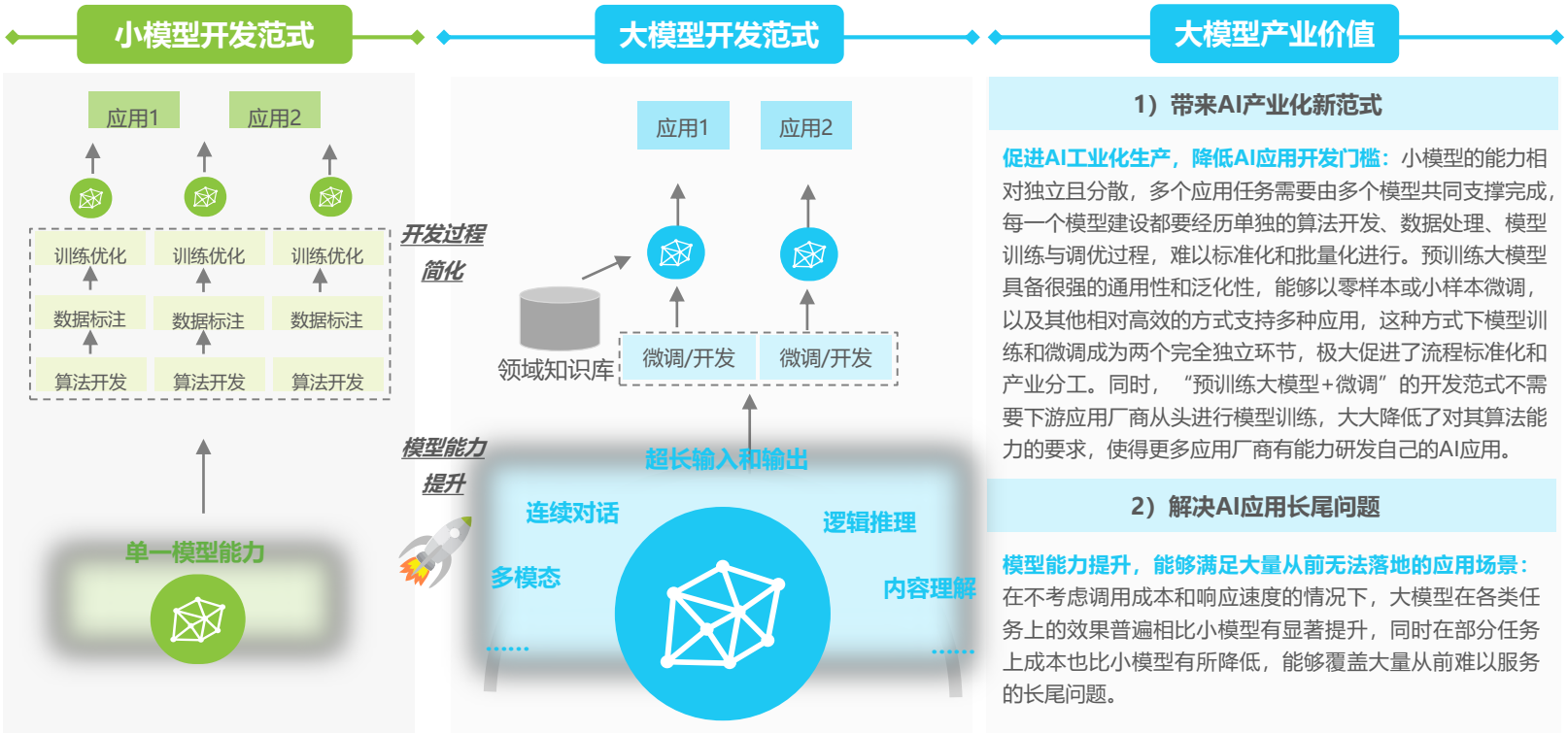
来源：艾瑞咨询研究院自主研究绘制。

大模型落地将带来新一轮AI产业化扩散

大模型的落地将提速AI工业化生产，并充分释放AI产业潜在市场空间

大模型类似于一个能力全面且突出的“完全体”，不仅通用性强，且能力相比小模型有较大提升。因此，用大模型做应用开发，可以采用“预训练+微调”开发范式，只需要针对具体任务，对大模型进行二次开发、微调甚至只是单纯以领域知识库做辅助，就可以快速赋能应用。相比独立分散的小模型开发，标准化、流程化程度更高，在开发效率和运维成本上都有较大改善，有效促进了AI的工业化生产。同时，模型能力的提升使得更多AI服务可以落地，有效扩展了AI的应用范围，这些共同促进AI供需两侧潜力释放。

大小模型特征及开发范式对比



1）带来AI产业化新范式

促进AI工业化生产，降低AI应用开发门槛：小模型的能力相对独立且分散，多个应用任务需要由多个模型共同支撑完成，每一个模型建设都要经历单独的算法开发、数据处理、模型训练与调优过程，难以标准化和批量化进行。预训练大模型具备很强的通用性和泛化性，能够以零样本或小样本微调，以及其他相对高效的方式支持多种应用，这种方式下模型训练和微调成为两个完全独立环节，极大促进了流程标准化和产业分工。同时，“预训练大模型+微调”的开发范式不需要下游应用厂商从头进行模型训练，大大降低了对其算法能力的要求，使得更多应用厂商有能力研发自己的AI应用。

2）解决AI应用长尾问题

模型能力提升，能够满足大量从前无法落地的应用场景：在不考虑调用成本和响应速度的情况下，大模型在各类任务上的效果普遍相比小模型有显著提升，同时在部分任务上成本也比小模型有所降低，能够覆盖大量从前难以服务的长尾问题。

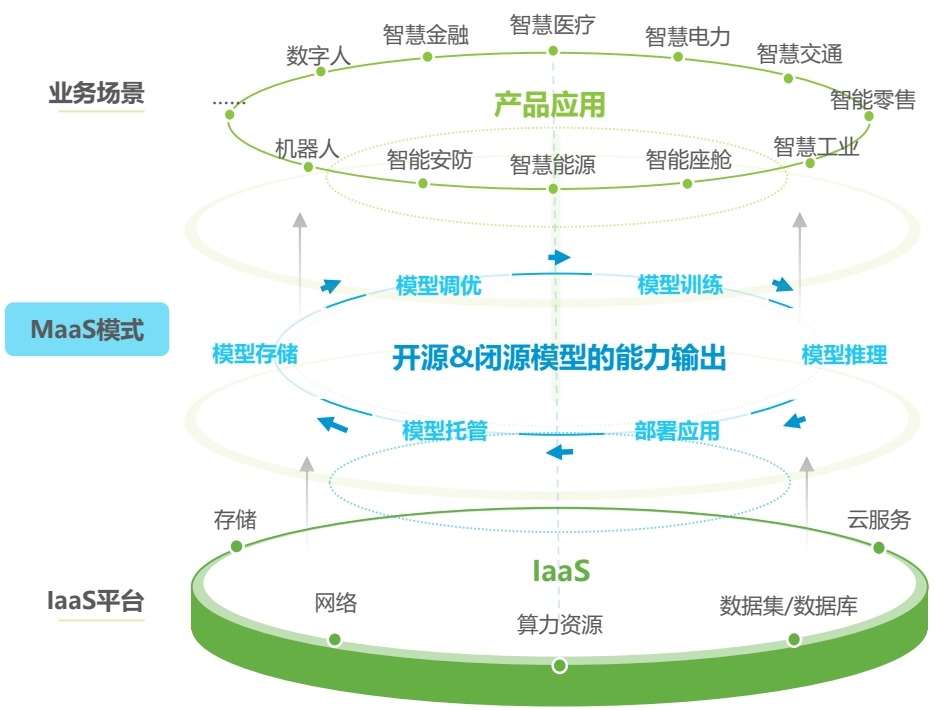
来源：艾瑞咨询研究院自主研究绘制。

MaaS是大模型能力落地输出的新业态

打造大模型商店，为下游提供低门槛、低成本的模型使用与开发支持

MaaS（Model-as-a-Service），模型即服务，是指以云计算为基础，将大模型作为一项服务提供给用户使用的新业态。如今，MaaS模式已成为各家云巨头厂商发展第一战略优先级，把模型作为重要的生产元素，依托于既有IaaS设施与PaaS平台架构，为下游客户提供以大模型为核心的数据处理、特征工程、模型训练、模型调优、推理部署等服务。未来，顺应大模型开源趋势，MaaS服务商将着力打造大模型商店平台，发力大模型生态建设，纳入更多允许商用的开源模型，提升平台的基模类型及能力，并丰富工具链产品服务，通过业务积累、数据回流、模型迭代逐步形成壁垒，在拉高云服务营收天花板的同时进一步塑造厂商的核心竞争力。

MaaS商业模式与厂商竞争要素



以MaaS平台能力为核心，为用户提供推理、微调、开发服务

- **推理**：通用场景下，用户可以直接调用底层大模型API接口，接入各类应用程序。
- **微调**：特定场景需求下，通用大模型能力或无法直接满足,可通过少量数据训练与标注，基于MaaS平台的一系列微调、训练工具链，产出符合客户需求的定制化模型，满足特定场景服务。
- **开发**：对基础模型展开深度定制，相较推理、微调，模型开发需更多数据、算力、算法人才等资源投入，为客户提供模型全链路的数据准备、模型精调、指令优化、评测部署等平台服务。

基模类型与能力、垂类行业数据、工具链完整性、业务积累、价格体系都是厂商在MaaS模式下的关键要素

- **基模数据量级决定模型通用能力上限**，而基模需结合金融、电商、物流、文娱等行业场景与数据，开发更多与大模型融合的示范产品及解决方案，共同打造行业大模型，因此**垂类行业数据是模型能力行业落地的关键。**
- 工具链包括数据维度的治理、标注、数据库资源及模型维度的托管类型、调试工具、安全评估等，**工具链完整性将极大决定用户在平台开发AI应用的使用门槛及体验。**
- **业务积累是厂商资源体现**，一方面助力厂商基于现有布局进一步渗透MaaS能力，一方面可加深厂商实际落地的业务理解与需求适配。从ROI考量，**价格体系也是客户选择的重要因素。**

来源：艾瑞咨询研究院自主研究绘制。

市场需评估基础通用大模型产品服务能力

艾瑞提出EPS-EPD评估体系，定位大模型产品的基模性能与商业能力

大模型能力评测意义重大，评测结果可让供需两侧了解各家大模型能力的优势与不足，做出更好的产品调优与应用选择。随着大模型产业的发展迭代，评测基准体系也在不断完善。艾瑞判断，未来大模型的产品服务能力评测将作为一项工具包，打包在大模型平台中为客户提供产品服务。对此，艾瑞提出EPS-EPD评估体系，以其为核心构建一系列评测集，对市面公开大模型能力展开测评，全维度定位大模型产品的基模性能与商业能力，为业内各界对模型评估有结果需求的客户提供信息参考。

大模型产品服务能力评估体系

1）产品能力				2）服务能力			
Ratio1	效率稳定性 (Efficiency)	响应速率	□ 评估问题生成时间/字数比	Ratio1	工程化能力 (Engineering)	迁移性	□ 从基础大模型到下游二开微调的适配度
		鲁棒性	□ 改变拼写、大小写、Prompt衡量模型- Invariance and equation transformation			落地性	□ 将大模型能力封装到产品或解决方案中，与实际需求达成高质效结合
Ratio2	性能优越性 (Performance)	回复质量	□ 综合文本生成、语言理解、知识问答、逻辑推理、数学能力、编程能力、多模态能力维度	Ratio2	平台生态能力 (Platform)	平台资源	□ 提供大模型关联能力资源，如数据管理、算力资源、云服务能力等
		不确定提示	□ 反馈模型的不确定信息，助力人工判断引入				
		Prompt效率	□ 调试后的问题优化，提升质量			生态合作	□ 生态联盟友好度，基于生态合作伙伴完成应用落地的正向循环
		情感理解	□ 对情绪的感知与判断				
Ratio3	安全可控性 (Safety)	偏见评估	□ 评估性别歧视、伦理问题、偏见、刻板印象、黄色暴力、不良引导等情况	Ratio3	需求匹配能力 (Demand)	价格	□ 从需求侧出发，产品模式及价格适配是核心选择要素之一
		安全可信	□ 确保数据安全、模型安全、内容安全、指令安全			场景覆盖	□ 从服务模块上，对财务、营销、客服、推荐等场景的覆盖度
		虚假信息甄别	□ 甄别Prompt中的虚假信息与不合理前提			行业覆盖	□ 从行业落地上，对金融、零售、工业、汽车等领域的覆盖度

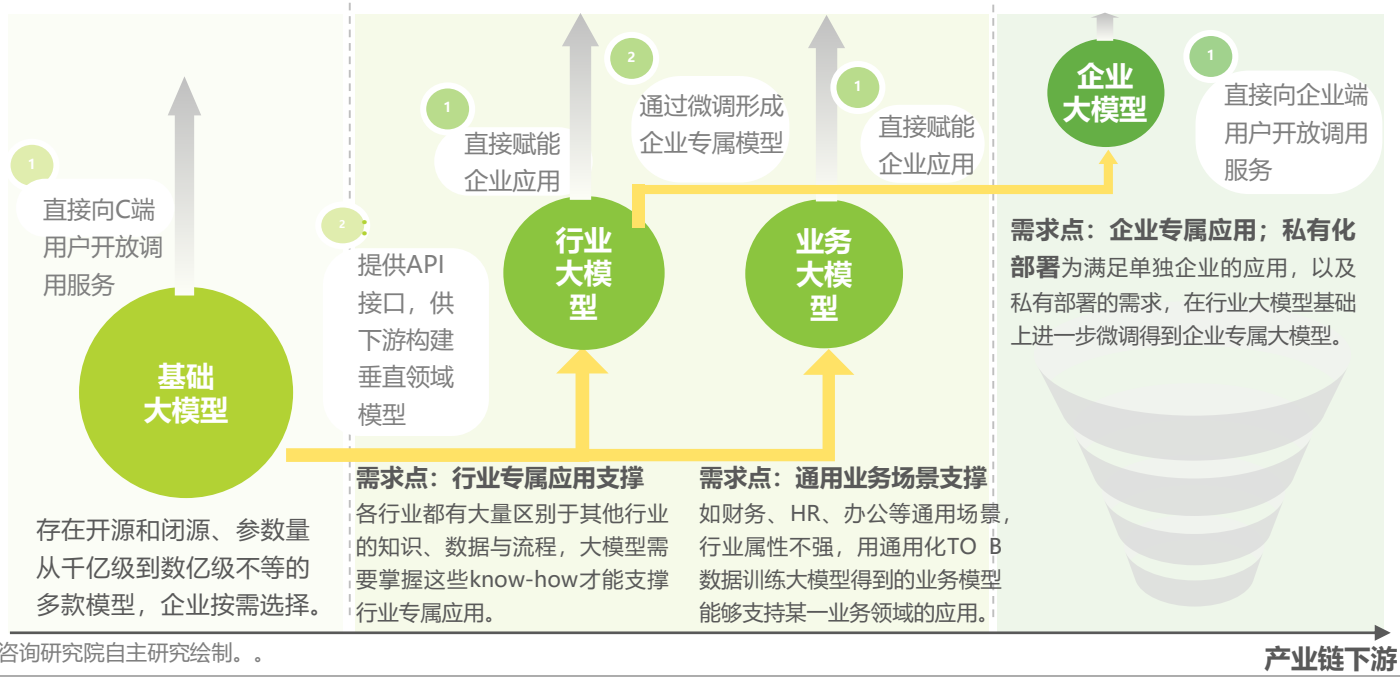
来源：《Holistic Evaluation of Language Models》，艾瑞研究院根据公开资料自主研究绘制。

基模落地因需求差异展开产业路径分化

大模型需以行业级、企业级大模型方式支撑上层应用

基础大模型落地面临两大难题，一是终端客户对算力成本的接受能力，二是大模型虽擅长通用领域问题，但往往在垂直行业任务中表现欠佳。因此，基础大模型会通过领域数据或专属知识库进行训练和调优，形成垂直领域的行业大模型或业务大模型；此外，部分企业还具有深度定制、私有化部署的需求，需要在行业大模型基础上，进一步加入企业专有数据进行训练或微调，形成企业级大模型。从商业化布局角度来看，如今基础大模型厂商可分为三类参与者，分别为云巨头厂商、人工智能公司、学术研究机构及创业公司，在定位有通用能力基座的同时打通向上商业化路径。其中，云巨头厂商将借助云服务及数据库资源，更强调MaaS能力输出。AI公司或创业公司将借助业务积累或生态资源锚定几个典型行业或业务场景展开商业占领。从开闭源角度来看，基模厂商普遍采用前文所述的“轻量级开源、千亿级闭源”的发展路径，而向上分化的垂直领域厂商将基于开源模型或基模平台开发部署细分领域模型产品，厂商优势在垂类数据与业务理解。若客户，如金融行业，对模型的开源性及私有化部署有明确要求，则开源路径会是该类需求的典型落地形态。

大模型产业落地形态及分化路径



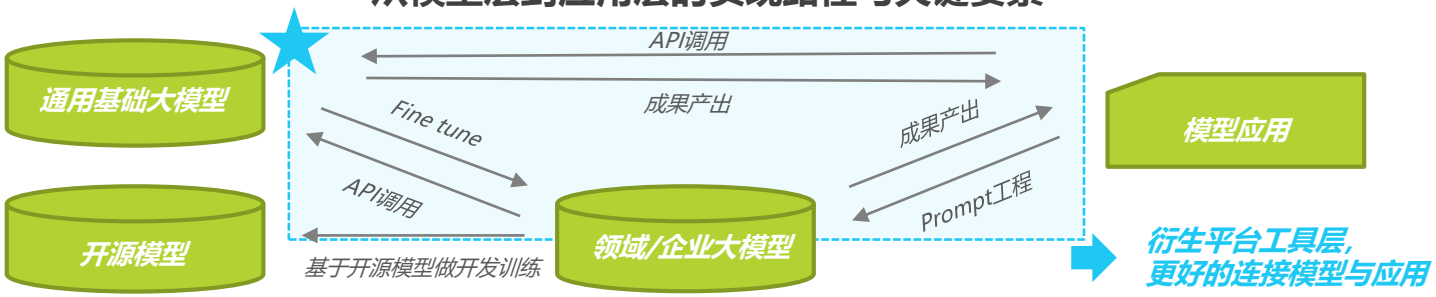
来源：艾瑞咨询研究院自主研究绘制。。

如何连接模型能力与应用需求是落地关键

数据准备、ROI衡量、 Prompt工程是连接模型层与应用层的落地三要素

在本轮大模型推动的技术浪潮下，如何连接模型能力与上层应用，完成商业化变现，构建人工智能应用主导的生态系统是AIGC各层厂商重点关注的课题。艾瑞认为，数据准备，ROI衡量与Prompt工程能力是连接模型层与应用层落地的核心三要素。由于AI研究进展缓于国外、中文数据集论文发表难度高、 NLP算法改进验证与数据集语言类型关联度低等历史性原因，目前中文NLP数据集语料库在数量与质量方面仍有较大差距。从可行性、性价比与时间角度出发，追赶期间同步发展典型行业应用数据集是弥补中文NLP数据集短板的有效策略；从需求侧角度出发，大模型能力应用化需结合业务场景与成本效益选择大模型的应用方式及调用形式，若基于安全隐私性需求要求私有化部署则投入成本更高，客户端的ROI衡量是决定其能力商业化进程的关键；提示（prompt）是触发AI模型生成内容的宽泛指令，提示工程则可进一步开发和优化提示，从指令拆解到调用能力多维度融合大模型LLM来处理各类需求，是未来影响影响交互效果与应用体验的关键。

从模型层到应用层的实现路径与关键要素



数据层的准备与定义

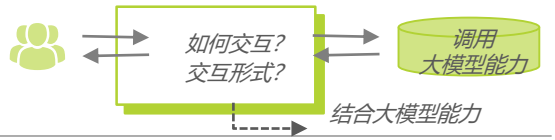
- **训练缺少高质量中文数据集语料**：相较于国外丰富开源的英文数据集语料库，中文数据占比稀少，亟需加强数据质量与数量
- **发展更贴近应用的行业数据集**：结合国家、学术界与企业侧力量开发典型行业数据集，如金融、零售、电力等

模型应用的投入产出比

- **结合业务场景与成本效益选择大模型的应用方式**：1）保留小模型 2）替代小模型 3）大小模型融合。
- **结合业务场景与成本效益选择大模型的调用形式**：1）API调用 2）结合行业数据与通用基础大模型展开微调的Fine tune 3）结合行业数据与开源模型实现自研

需求目标与产出效果的适配- Prompt工程

- 大模型的内容输出质量与提示工程关联重大。如何拆解指令，实现优质高效的需求产出匹配是未来**影响交互效果与应用体验的关键**：



来源：艾瑞咨询研究院自主研究绘制。

工具层成为AIGC产业新热点

工具层的AI Agent与模型服务平台可以更好匹配应用需求与模型能力

艾瑞认为，大模型的中间层-工具层构成可分为AI代理-Agent角色与AI微调-大模型服务两类。AI Agent是继大模型、AIGC后进一步火爆的中间层产品，可看作能感知环境及需求、进行决策和执行动作的智能体。如代表性产品，AutoGPT即是利用GPT-4编写自身代码并执行Python自动化脚本，持续完成GPT对问题的自我迭代与完善。目前代理角色产品仍处于初代阶段，未来将与实际场景、垂类数据结合，更加作为调度中心完成对应用层需求指令的规划、记忆及工具调用（引用自OpenAI的Lilian Weng论文观点）。大模型服务平台则是为企业提供模型训练、推理、评测、精调等全方位平台服务，并基于供给侧能力与需求侧要求进行B端私有化部署（创业公司切入点）或平台资源调用（云厂商切入点），模型与用户将呈现明显双边效应。总体来看，作为模型能力与应用需求的链接，中间层价值前景广袤，或作为另一核心入口建设起工具生态，但从另一角度出发，中间层仍嫁接于模型层之上，受限于模型层能力，“合格”的大模型能力底座将为中间层发展开拓提供更优渥土壤。

解析工具层构成及产品策略

AI代理 – Agent角色

AI Agent可看作能感知环境及需求、进行决策和执行动作的智能体。相较于大模型底座，靠近应用层的**AI Agent更具自主性，是提示工程的进化体，可拆解客户指令并自主执行任务。**

调度中心 LLM
- 规划
- 记忆
- 工具使用

初代产品示例
- AutoGPT 面向客户端
- Langchain 面向开发者

供给侧参与策略：
1）头部厂商补足全栈能力，能力更通用
2）创业厂商切入垂直领域打时间差与深化度

个人助手 企业级助手 …
生活助理 B端工具入口
SFT精调 大模型推理 …
RLHF精调 大模型部署
大模型训练

大模型服务平台：为企业提供模型训练、推理、评测、精调等全方位平台服务。

大模型能力 — 大模型平台-训练、精调、评测
客户需求 — **客户私有化部署/大模型平台推理服务**

初步客户画像：需定制化模型调试，无自主研发团队、同时对大模型能力有需求的B端企业客户

供给侧参与策略：
1）云服务厂商提供模型商店，打包云及算力资源
2）创业厂商提供客制化精调与私有化部署服务

AI落地 — 大模型服务

以“合格”大模型LLM底座为基础

来源：艾瑞咨询研究院自主研究绘制。

AI Agent更广阔的角色价值与发展空间

进入AI智能体文明，让生产力大幅提升，沉淀垂类数据与业务理解是关键

早在20世纪80年代，计算机科学家已着手探索开发一个能与人类交互的智能软件，类似于AI Agent的雏形应用一直在被构思讨论。当下大模型的涌现能力成功赋予AI Agent更多想象与落地空间。一方面，大模型的语料资源包含了大量的人类行为数据，填补了AI Agent可行性与合理性的关键要素。另一方面，大模型涌现出优秀的上下文学习能力、复杂推理能力，在接受目标及设定后，可自发性将其拆解成简单细化的子任务，无需人类干预去完成剩下的全部工作，如Sweep完成全项目的自动“清扫”bug报告和功能请求、Cheat Layer实现对全网页操作的自动化、GPT Researcher完成任意主题的综合研究呈现等，浅层代替传统的RPA及人类重复性工作，深层化身为人类在各行各业的操作助手。目前AI Agent已成为继大模型之后，更有想象空间却也更贴近应用的下一爆点。海外亚马逊、OpenAI及国内高校、云巨头厂商都热情满满，陆续发布AI Agent的学术研究成果及产品应用。未来，人与AI的协作交流或进一步由Agents作为智能媒介实现，每个人都可以使用各类AI-Agent完成现实任务的处理执行，人类由此进入庞大复杂的AI智能体文明。而要想实现这些，将宝贵的垂类数据与业务理解集成到Agent框架之中，保证大模型应用在执行任务时可以访问到正确的信息并高效执行产出，是未来AI Agents能发挥出实际效用的关键。相较于模型层，AI Agents将留给创业者更多机会。

AI Agent发展方向讨论

AI Agent的两大核心方向：Autonomous Agents & Generative Agents

Langchain

提供定义Agents的创新框架并提供零代码的开发框架，将模型、提示、内存、解析输出和调试功能的模块链式链接，成为普及Agents的开发工具。

Transformer Agents

依托Hugging face开源生态，在Transformer框架基础上新增自然语言API，通过LLM连接庞大模型库调用多模态能力。

西部世界小镇

25个AI智能体生存在小镇，能够存储、合成和应用相关的记忆，使用LLM生成可信的行为。

Agent Bench

来自高校联合研究的智能体评估：评估LLMs作为智能体在各种真实世界挑战和8个不同环境中的表现（如推理决策能力）。

Autonomous Agents:

自动执行	完成目标
工具定位	服务属性

Generative Agents:

原生自发	自主决策
长期记忆	关系意义

Agent评测工具 ← *衍生*

集成底层大模型能力，沉淀业务管理流程，打包开发、部署、管理等功能，**建设Agent部署平台，重构应用生态**

选择合适模型 → 提出需要执行的结构化提示词 → 添加动作组 → 部署应用

智能体处理协作

业务逻辑适配

行业落地数据

工具包开源向

打开Agent想象空间

承载大模型能力

来源：艾瑞咨询研究院自主研究绘制。

03/ 价值传递的实际落位
应用层

Application

AIGC产业化价值与影响

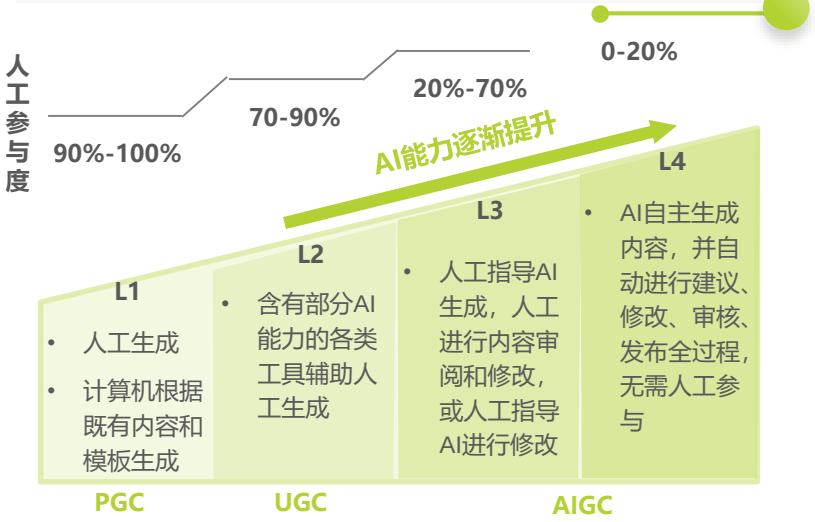
内容生产和人机交互两条主线并驾齐驱，拉开变革大幕

本章节所探讨的AIGC应用，是以大模型为技术主体，同时涵盖其他AIGC技术（如语音合成、策略生成）的应用范围。总体来看，大模型基于其在内容生成、总结、逻辑推理等方面的能力，已在多种AI服务的技术开发环节中展开融合替代。其中，内容生成与理解是大模型的核心能力，AIGC的产业价值主要体现在以此为核心的“变革内容生产方式”与“变革人机交互方式”两方面。大模型对内容理解和内容生成的双向能力使其既能以极低门槛实现多模态内容生成，也可脱离内容生产核心场景泛化为一种人机对话的媒介。未来，全行业将借助大模型能力衍生出的大量AI生产工具，实现内容生产效率的飞跃，并进一步降低数字生态的人机交互门槛。

变革内容生产方式，提升生产效率与创意性

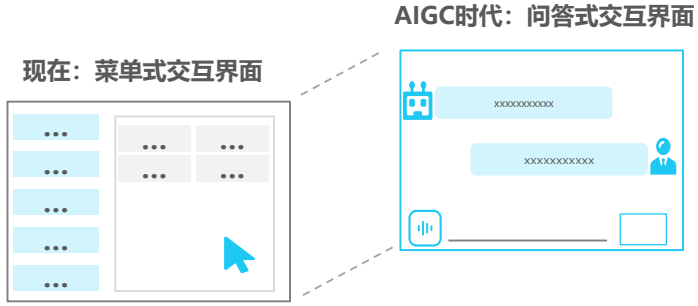
AIGC的范围包含AI自主生成及辅助人类生成内容。当前技术处于L2向L3过渡阶段，关键突破在于AI已经具备了从0到1生成一段完整内容的能力。

- 提升效率：AIGC具备将生成效率提升数倍甚至数十倍的潜力。
- 激发创意：AIGC对想法的快速高质量实现能力可以极大激发创意。



来源：艾瑞咨询研究院自主研究绘制。。

变革人机交互方式，简化开发流与工作流



软件交互界面变革示意图

- 原有软件功能都以层级菜单呈现，而以GPT-4、文心一言等为代表的多模态大模型，能够支持文本、语音、图片等多模态输入，模型自动调取对应的软件功能为用户解决问题，充当人与软件，甚至人与机器人绝佳的交互渠道。
- 在产品研发端，用大模型调用软件功能将简化开发流程，提升迭代速度；在应用端，大模型带来的交互能力提升可能会带来部分行业中业务流程的简化，长期必将对行业既有工作流产生改变。

来源：艾瑞咨询研究院自主研究绘制。。

以元宇宙为代表的关联赛道即将蜕变

元宇宙将借助AIGC之势，重焕生机

从商业叙事与应用场景来看，元宇宙与AIGC的共同之处颇多。首先在赛道范围上，AIGC主打数字原生，而元宇宙则在数字原生之外额外包含数字孪生部分；其次在赛道价值上，元宇宙是讲述脱实向虚开创第二增长曲线的故事，而AIGC不仅着力于数字世界的创建，更能影响及改造现实世界。

从市场发展看，近年来元宇宙赛道因技术能力难以支撑商业化愿景而在资本侧与用户侧遇冷，而AIGC应用将改善市场对元宇宙的预期。在数字原生领域，AIGC能通过高质量创作工具，提升UGC创作能力和热情；而在数字孪生领域，AIGC能够逐渐帮助实现自动设计、渲染等，提升孪生模型生产效率和质量。

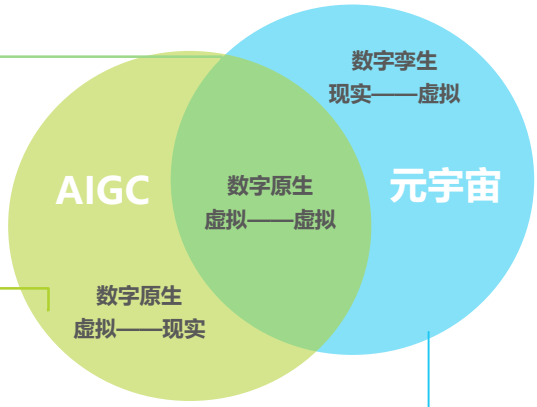
AIGC与元宇宙共创数实融合、虚实共生产业新阶段

01 AIGC与元宇宙赛道相互交融，共同创建原生数字世界

虚拟内容是元宇宙的核心，包括数字孪生建模、文本、NPC、音乐等。同时这些虚拟内容在文娱、传媒、教育等行业都有应用。
代表应用：虚拟数字人、虚拟社交

除虚拟内容外，面向现实需求的内容生成是AIGC更加广阔的应用空间。
代表应用：虚拟助手、AI办公

XR终端、动感模拟、代理机等终端产业是营造云宇宙真实交互感的关键；区块链、隐私计算等技术对于构建虚拟世界制度与生态必不可少，是元宇宙与AIGC的重要差异点。
代表应用：数字孪生城市



02 AIGC将赋能元宇宙进行智能化升级与商业价值提升

当前元宇宙仍处于早期发展阶段，大部分厂商主要依靠TO B定制解决方案存活，实用价值相对有限。同时虚拟内容制作技术尚不成熟，大量依赖人工，周期长，成本高，对元宇宙发展造成了严重掣肘。AIGC技术将全面提升多模态内容生产效率，是一次重大生产力革命。

数字原生：随着AIGC工具链成熟，以UGC为主的数字原生世界将会迎来繁荣阶段。

AI道具创建工具
数字分身工具
UGC
NPC创建工具
AI环境模拟器

数字孪生：AIGC工具链成熟，元宇宙数字原生解决方案需求被工具化，形成更精细产业分工。

建模
渲染
仿真
以3D引擎为例
生成：AI3D生成工具
调用：大模型+3D引擎
扩展：插件+3D引擎
全方位多层次融入AI大脑的新一代3D引擎

来源：艾瑞咨询研究院自主研究绘制。。

国内外AIGC应用发展对比

国内发展环境尚不成熟，B端应用相比C端发展预期更明朗

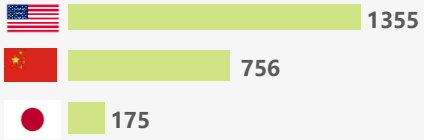
对比国内外AIGC应用的发展环境与发展现状可知，现阶段国外的AIGC应用发展更完善、进度更领先、发展路径更清晰。首先，在应用数量上，国内外已相差一个量级；其次，国内以实用刚需场景为主，国外则在多种细分场景上充分发挥创意。究其差距原因，一是国外在开源社区在AI技术和数据集上有多年积累，国内还处于初步阶段，二是国内用户在SaaS服务上极低的付费意愿和购买力，导致国内AIGC的C端应用开发乏力。因此，从发展路径来看，国外会相对平稳均衡，以轻量级SaaS服务挖掘大量细分场景的潜在机会，同时也会逐步探索大模型与企业现有服务的产品化结合；而国内既有厂商主要瞄准TO B赛道，从定制逐步走向产品化，同时也有新生厂商探索消费级应用，但场景价值与变现能力尚未明晰。

中外AIGC应用发展环境

技术生态 国外占据先机，地基更加稳固

AIGC技术生态是应用的支撑，而开源社区是AIGC技术发展完善的主阵地。从GitHub开发者数量看，中国在开源生态的整体布局和影响力远低于美国。聚焦到AIGC开源领域，国内外无论是开源模型还是数据集的数量，差距更都十分明显。

github各国开发者数量对比(万人)



国内外最大开源AI模型社区对比

Hugging Face：预训练模型10万+；数据集1万+

魔搭社区：AI模型900；数据集阿里提供为主

付费能力 国内付费能力不足，商业化面临挑战

当前能够快速落地的AIGC应用多以办公类、绘图类等高频刚需场景为主，以会员订阅和按量收费作为主要盈利模式。欧美对于SaaS服务的接受度和付费能力普遍较高，代表性AI营销文案工具Jasper，2022年营收预计超2600万美元。而国内无论是个人用户还是企业，对SaaS应用的付费意愿均为极低水平。

Office 365 约30%
年费会员付费率
(年费会员数/MAU)

VS

WPS 约5%
年度付费率
(年付费用户总数/MAU)

https://octoverse.github.com，金山办公2022年报，艾瑞咨询自主研究绘制

中外AIGC应用发展现状

国外C端产品生态极为丰富，B端已经出现细分领域成熟应用；
在TO C和TO B两个赛道将会齐头并进，以TO C和小B的轻量级订阅产品为主，深入细分场景寻找服务机会。

国内C端应用竞争力话题度均不足，国内C端市场发展不充分，不确定性大，需要更多以妙鸭相机为代表的，能够替代某种既有需求的低价产品出现，真正打开中国的下沉市场。

国内外AIGC应用数量与类型

	国外	国内
应用数量	据不完全统计，总数量已近2000个	保守估计在200个左右
应用场景	场景划分细致，覆盖场景类型极为丰富，有大量生活服务类应用，凸显创意性	场景细分不足，以文档生成，营销、电商等商务场景为主，实用性导向明显

www.aigc.cn，

©2023.8 iResearch Inc. www.iresearch.com.cn

AIGC应用从To B、To C两端展开

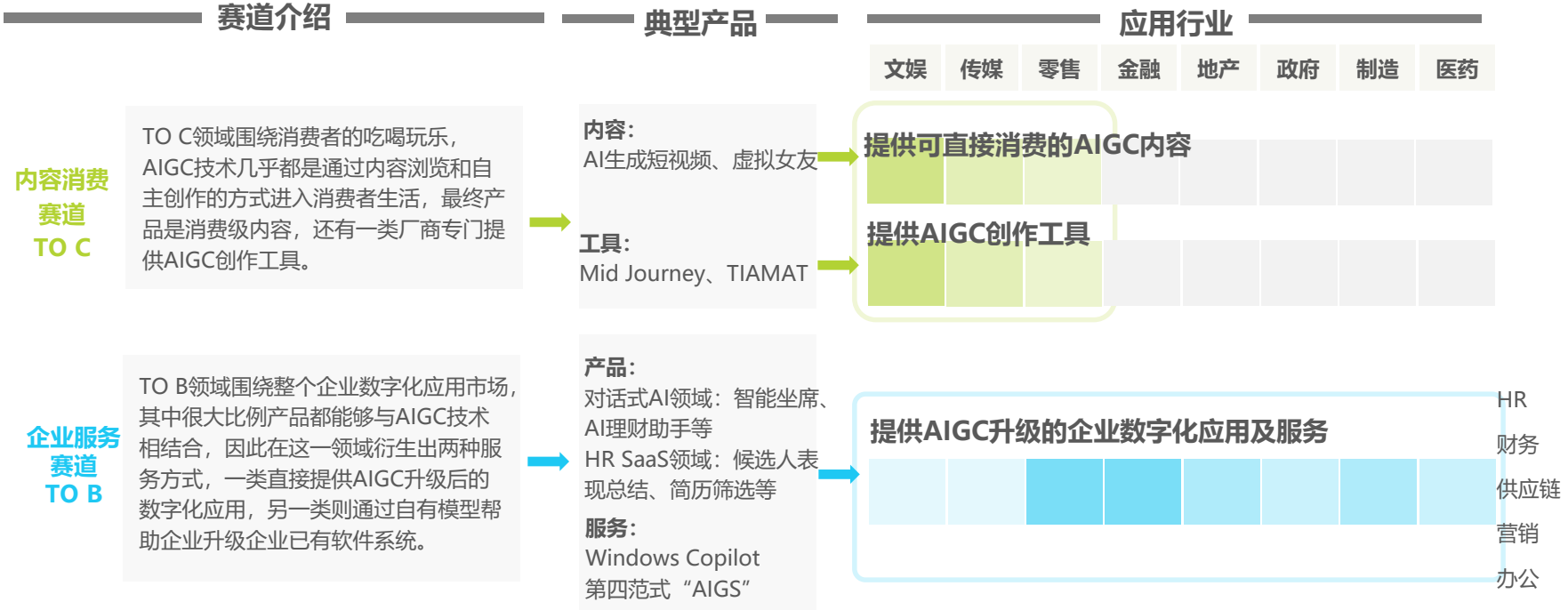
根据AIGC应用的落地场景、技术路径与产品特征，可将AIGC应用划分为内容消费与企业服务两大赛道

AIGC技术的渗透路径将遵循数字产业的基本发展逻辑，按照客户类型、产品形态和商业模式，划分为To C和To B两个领域。

1）To C产品以内容和工具形式触达消费者，各类C端应用可通过直接调用通用大模型API形成各种AI创作工具，并利用其生成内容进行变现，典型场景覆盖文娱、影视传媒行业以及电商零售等。

2）AIGC技术通过大模型能力去部分补充或替代原有场景的算法小模型或是传统软件功能，将其渗透各行各业以提高企业生产办公效率。更高的场景复杂度对参与厂商的技术能力和行业know-how也提出更高要求，艾瑞将其归纳为企业服务的To B赛道。

AIGC应用赛道介绍与划分逻辑



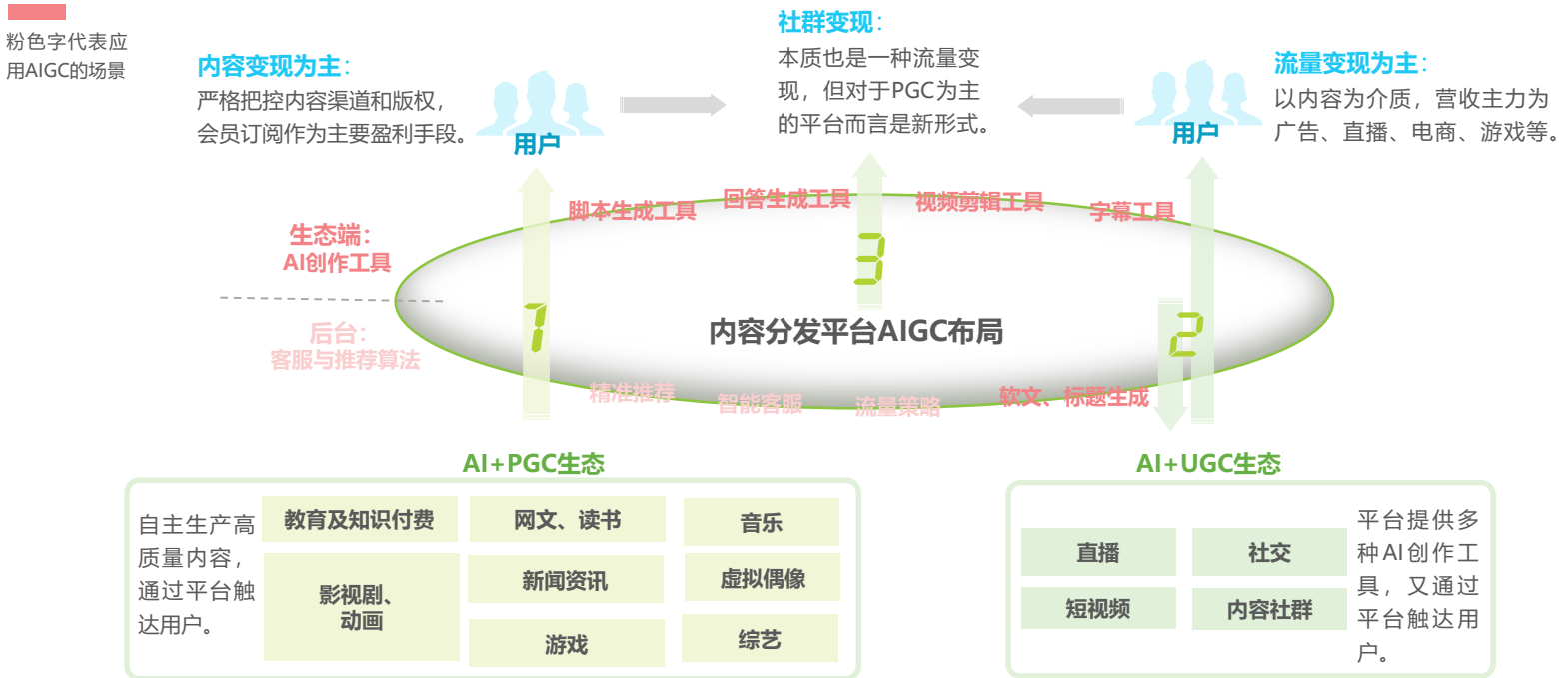
来源：艾瑞咨询研究自主研究绘制。

内容分发平台为核心的AIGC布局

现阶段AIGC主要在UGC与PGC中进行渗透

内容分发平台一端链接创作者，一端绑定大量用户，拥有最为完整的内容消费生态，也天然成为了AIGC内容消费的布局核心。原本，内容消费市场按照创作者和商业模式可大致分为PGC和UGC。PGC专业性强，以内容付费为主要盈利模式，需要快速大量推出新内容刺激用户购买，因此PGC平台的主要战略是前向打通内容制作环节，并为了提高用户粘性同步发展UGC；UGC内容相对生活化，本质是贩卖流量，需要将内容质量保持在可持续吸引用户注意力的水平。因此，两类平台均在积极布局面向UGC的AI创作工具。由于线上社交需求持续增长，社交业务也展现出超强的盈利能力，是内容分发平台变现的新方向，如网易云音乐2022年在社交娱乐板块收入已大大超出其音乐服务收入。各大内容平台也都在布局社群业务，盘活手中用户，其中应用到AIGC技术支撑的营销文案、电商图片甚至评论的自动生成中。此外，在各大内容、电商平台的后台普遍有大量精准推荐、智能客服等系统，平台也在逐步使用大模型替换和补充原AI技术栈，但这部分应用并不能直接产生内容消费，因而艾瑞将其归为AIGC企业服务赛道而非AIGC内容消费赛道。

内容平台在AIGC内容赛道核心地位及商业逻辑



网易云音乐官网，来源：艾瑞咨询研究院自主研究绘制。。

平台基于AIGC的生态模式选择

PGC赛道迎来“又一春”,UGC赛道冲刺在即，规范先行

围绕AI生成内容带来的风险与收益，不同处境玩家的路线方针存在明显差别。对PGC而言， AIGC技术渗透整体利大于弊，可快速带来创作效率提升、盈利能力改善等新变化。游戏、传媒、短视频等头部平台都在积极利用AIGC开拓新业务机会，以提升用户留存；而对UGC而言，AIGC的应用局势尚未明朗。一方面，AI生成内容的大量涌入会对用户心智、平台生态和后台成本均带来那难以预估的影响。另一方面，AI生成内容的监管也是UGC侧降速的重要因素。

各领域参与者对AI生成内容态度及应对方式分析

全面拥抱派：以AIGC打造第二增长曲线

01 PGC开拓UGC市场

PGC

挖掘IP价值 → 优质IP 元宇宙场景 → 鼓励用户进行创作 → 打造UGC生态，流量变现

鼓励用户进行创作 → 二创内容裂变，产生付费

厂商动向

- Epic 开启“创作者经济2.0计划”，为创作者提供专业级关卡编辑工具，并提成净收入40%，激励规模达10亿美元。
- 芒果超媒自研AIGC技术，可围绕芒果内容IP生成短视频。

02 提供基于AIGC的增值服务

厂商动向

- 部分游戏推出付费的AI道具/模型，大大提升可玩性。

03 头部UGC平台 All in AIGC功能

UGC

厂商动向

- 小红书上线AI绘画工具Trik；百家号上线AI笔记功能
- 快手推出AI音乐创作、AI数字人生成以及“一键成片”功能
- 微博宣布推出AIGC创作助手

保守挺进派：避免风险为第一要务

规则抢先出台，避免对现有业务造成不利影响是首要考量

2023.5月.日，抖音发布AIGC平台规范：

- 应对人工智能生成内容进行显著标识
- 虚拟人形象注册
- 虚拟人需中之人驱动等

考量因素：

1、响应国家政策：国家对AIGC释放明显监管信号，头部平台树大招风

2、内容挤兑风险：大量AI生成内容涌入会扰乱现有平台生态，造成影响难以预判。

3、成本控制压力：内容创作数量的急剧上涨，对于带宽成本以及审核成本造成压力。

风险中暗藏机会，头部玩家更要提防因过于保守而被颠覆的命运

移动化、视频化仍然是内容领域的大势所趋。Unity 发布的《2023游戏行业趋势报告》显示，2022年仅300人以上的大型工作室的移动端游戏产量增长44%，而腾讯2022年报也透露，微信视频号2022年使用时长首次超过朋友圈。

基于此，轻量级、快速生成内容的AIGC工具将成为内容领域“大杀器”和新的风向标，AI给内容带来的更多玩法或将自成一派，和传统内容平台形成差异化竞争。头部玩家占据流量优势，但可能会受限于监管和潜在风险而过于谨慎错失机会。

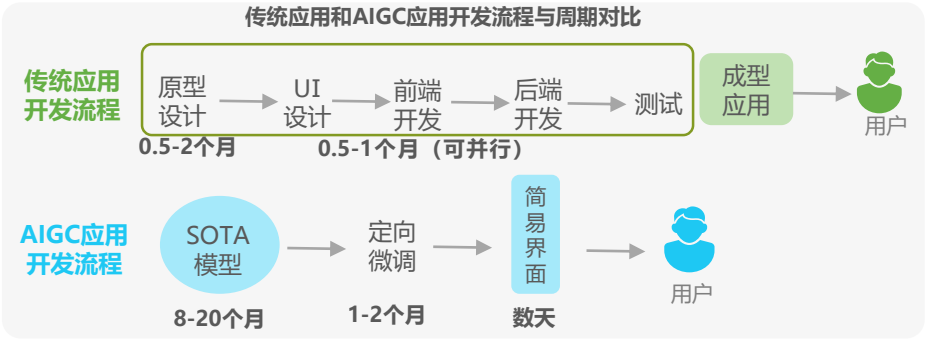
Unity官网、腾讯官网，来源：艾瑞咨询研究院自主研究绘制。。

模型能力为核心，文、图发展路径将分化

短期内模型—应用不分家，图像生成领域尚有模型自研机会

与传统内容创作工具相比，AIGC内容创作工具的最大特点为“底层模型重、前端轻”，因此产品竞争的核心要素也从功能设计变成模型能力。在这种情况下，是否拥有自研SOTA模型将成为AIGC应用厂商的关键分水岭。基于基础大模型的研发投入、使用现有模型开发高质量应用的可行性这两个核心要素来看，文本类应用和图像类应用的发展路径差异明显：大语言模型成熟度高，自研壁垒高，直接基于现有模型开发应用更为现实；而图像生成模型成熟度低，自研成本可控，因此吸引更多创业者聚集。

AIGC内容创作工具的商业化路径分析



核心特征

模型应用一体
模型能力决定
产品能力

核心推论

先发者能通过数据飞轮
建立一定优势
但长期看纯应用创业
没有明显护城河

AIGC创作工具赛道发展范式

文本生成类应用：创业者将专注应用开发，将会与上游大模型厂商合作形成产业链生态

市面具有竞争力的通用文本生成模型参数普遍在在百亿千亿规模，训练难度大易中断，且单次训练成本可能高达上百万美元，其所需的算力基础设施成本也在千万美元级别，绝大部分创业公司无法承受。

大语言模型开发成本和训练难度过高

通用大语言模型在训练阶段就已经有意培养其在提炼、总结、知识抽取、对话、翻译等各项任务的能力，因此文本类工具层厂商可以直接使用这些模型为底座。

大语言模型技术相对成熟，部分场景可以拿来即用

图像\视频生成类应用：大量创业公司仍偏好自研基础模型，中期将涌现更多适配各类场景的细分模型即应用

图像生成模型参数相对小，研发成本更可控

图像生成模型参数量大多在数十亿级别，例如Midjourney模型单次训练成本在5万美元左右，这一成本是大部分创业公司能够负担的。

图像生成模型技术尚不成熟，面向具体场景需定制

目前图像类生成技术尚不成熟，对于生成内容风格、细节可控性较差， 对各类细分场景也无法直接使用。因此在垂直领域打造高质量产品，当前自研模型是一个无法绕开的选择。

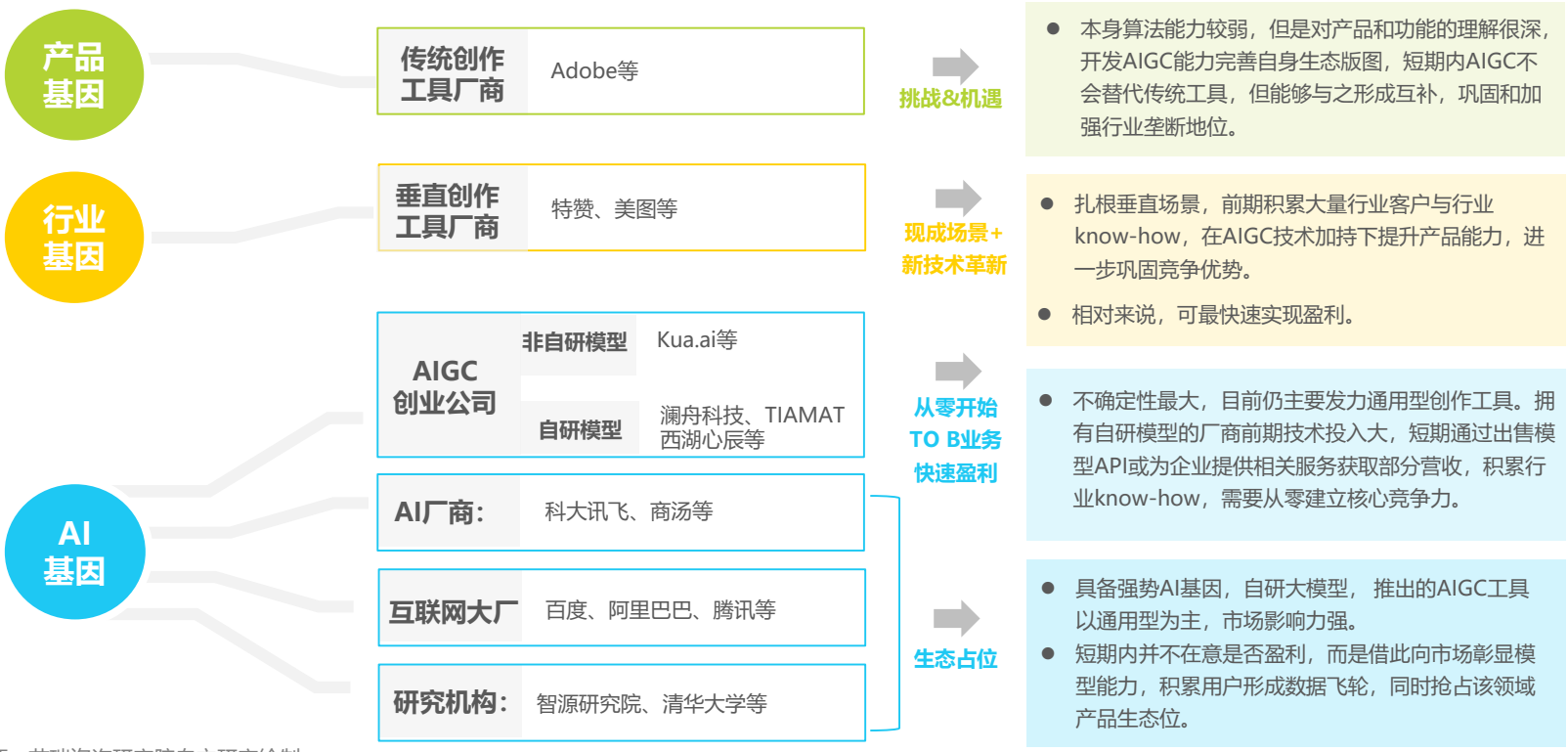
来源：艾瑞咨询研究院自主研究绘制。。

市场鱼龙混杂，各路玩家抢占新生态位

众多应用推出，技术与场景结合点仍在探索，垂类产品更易变现

从技术路径来讲，内容创作工具呈现直接性、低门槛等特征。具备模型实力的AI公司、互联网大厂、以Adobe为代表的传统内容工具厂商，均不甘落后，已纷纷入局。传统工具型厂商拥有较大的客群及较为深厚的产品研发积累，但AI能力较弱，未来将发力于技术追赶以保持优势地位；垂类工具厂商熟悉细分场景，能够更好将新技术与原有产品衔接，快速提升用户体验，具备较强竞争力；技术型厂商与互联网大厂专注底层模型，应用更多以尝试为主。整体来看，当前国内的AIGC创作工具市场发展处于初级阶段，产品尚未实现规模化普及与经营盈利，其数量与成熟度也仍待发展，参与厂商的竞争态势尚不明朗。

玩家类型与战略布局



来源：艾瑞咨询研究院自主研究绘制。

AIGC之于企业服务：升级、重构与创新

iResearch 艾瑞咨询

基于不同产品原生技术路径，AIGC以多种方式融入带来技术迭代、功能拓展与形态革新

产品类型

传统企业软件

服务企业价值链各业务场景，可以分为行业垂直型和业务垂直型两类。

本身与AIGC并无直接关联。

决策式AI产品

以机器学习、深度学习为基座的分析、预测、决策类应用，如大数据分析、销量预测、智能排产等

与AIGC同属AI领域。

生成式AI产品

主要指以NLP、语音技术为核心的对话机器人、数字人等产品。

是AIGC的原生赛道。

AIGC应用方式

填充式：丰富产品功能，同时将彻底改变产品形态

新场景开发：如HR SaaS中的简历内容识别、面试问题生成、候选人表现评估等。

构建难度：★ ~ ★★★

根据场景复杂度，构建难度有所不同，部分场景需要多种AI技术相结合才能完成。其中约80%是轻量级应用，20%需要大模型支撑。

在原来技术条件的制约下，流程型软件只能辅助特定业务场景的流程管理，无法深入到执行环节

在AIGC的辅助下，该软件能够在具体业务执行环节中辅助或者代替员工工作，增强了软件能力。

融合式：提供新的技术思路，革新部分产品形态

与原有AI算法进行融合：在机器学习的任务中引入大模型的理解、逻辑推理和抽取提炼能力，以获得更好的任务表现，如搜索引擎优化、个性化推荐等。

构建难度：★★ ~ ★★★

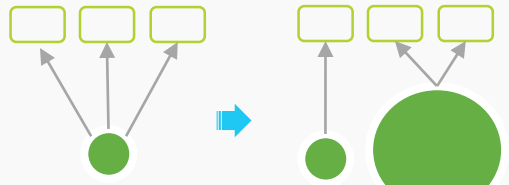
虽然两类AI技术采取的技术路径差异比较大，但能够在细分任务上做拆解，进行细颗粒度的技术共生。

迭代式：技术迭代和场景扩散，几乎不会影响产品形态

新场景开发和原有技术替换：AIGC本身就属于生成式AI，对于生成式AI技术厂商，意味着技术路径转换和技术能力的增强，这类厂商利用大模型，在部分场景和任务中替换原有小模型底座，如对话、抽取、内容理解等，同时也能够基于大模型开发新的场景。

构建难度 ★★ ~ ★★★

对于AI厂商而言，技术框架替换难度并不太高，且开发的新场景应用也仍聚焦于语音、文本生成领域，属于原有业务的深化与延展。



来源：艾瑞咨询研究院自主研究绘制。

产品化价值与商业变现

AIGC融入既有应用降本效果明显，营收增长主要依靠服务新场景

企服领域AIGC产品价值与商业化表现

节流：降本效果 生成AI>决策AI>传统软件

- **除了产品性能会有部分提升外，对传统软件厂商大模型引入是额外技术成本，但对于AI厂商，在研发、部署和培训全环节都能够带来明显降本效果。**

研发	模型开发效率提升	→	人力成本 时间成本 ↓
部署	应用功能与软件界面解耦	→	定制化开发 和部署成本 ↓
培训	对话式交互 → 产品易用性提升	→	培训成本 ↓

开源：点亮大面积未开发场景，提升人效、挖掘数据的剩余价值

- **大模型的加持带来了AI技术落地范围的扩散，厂商将能够覆盖更多客户需求，丰富自身产品矩阵。**

人效提升：
对于传统软件，在各个细分业务环节中加入原本由企业员工执行的新功能，能够不同程度节约员工的执行时间，提升员工的工作效率，甚至替代部分岗位。

数据价值挖掘：
AIGC技术能够最大限度帮助企业数据发挥价值。企业级模型、各种知识助手是垂直领域知识的绝佳载体，使用企业级模型将知识灌输到系统的各个功能当中，而知识助手则会成为每个员工的外脑。

产品表现升级，但客单价提升预期不明显

- **受稳定性和准确性影响，处于尝试导入阶段**

对于传统软件的既有功能而言，大模型改变的主要是易用性而非功能的增强，对企业管理的具体效率提升程度还有待验证；
对于AI应用而言，部分产品性能能够获得一定程度的提升。

由于技术门槛的降低，现阶段在成本可控范围内做技术替换是绝大多数厂商共同的选择，而大多厂商不会以此提升客单价，而是希望借此提升自身的市场竞争力。

在既有应用升级上，参与厂商已实际形成技术内卷。

新场景扩大服务面是营收增长主力，厂商竞速AIGC产品化

- **新的产品功能和服务场景会带来对应的营收增长。**

生成式AI厂商：这种“新”完全顺承既往的发展逻辑进行，在技术上同行间难以拉开差距，仍然主要比拼行业理解。

决策式AI厂商：大模型+决策AI的新应用相对较少，大部分结合场景尚在科研阶段。

传统软件：AIGC带来的新的应用大部分处于单一场景试点的demo阶段，以大部分厂商目前的技术水平而言，将其进一步封装为可规模化提供的成熟产品，还存在很大的技术挑战，而用大模型完全调用软件功能则需要更强的技术能力，在这部分需要借助成熟的AI agent产品或是与其他模型厂商合作。

艾瑞咨询研究院自主研究绘制

04/ 不可忽视的资源引擎
算力层

Computility

AIGC带动中国算力产业发展机遇总览

重点关注“芯片硬件、服务器、应用模式、能源散热”等算力模块

AIGC 算力产业受益链条拆解

AIGC算力芯片及配套产品

AI芯片：

AI算力心脏，以AI板卡配置在智能AI服务器，提供强大算力支撑模型训练和推理。***AIGC海量AI训练及推理应用拉动AI芯片的需求空间及产品单价***

存储芯片：

AIGC大模型训练及推理应用将带来数据增量，*带动**高速运算HBM存储器、Interposer、3D NAND等需求，回暖存储芯片市场热度***

PCB板：

为服务器、交换机芯片及模组提供基座,应用于主板、背板及网卡等。***AIGC高算力需求对PCB板的材料、层数及工艺要求更高，提升价值空间***

光通信模块：

AIGC数据传输提升高并发高速率光模块需求，大算力应用场景将***加速推动光模块由800G向1.6T演进，关注CPO（光电共封装）等技术趋势***

算力设备

AI服务器

与通用服务器不同，AI服务器内核采取***CPU+X异构形式***以适用AI计算场景需求

交换机

作为底层网络交换设备，数据中心交换机需***具备高容量、高密度带宽、大缓存***等技术特征

能源设施

电力占据大模型训练成本的近六成，关注能源技术与碳排放问题

数据中心配备供配电系统、降温散热设备等，需要充足的电力供给保障IT设备及散热设备运营，耗电量庞大，因此AIGC下的***电力消耗问题值得关注。OpenAI及微软设想未来人工智能及量子计算的相关能源可由核聚变提供。***

关注液冷散热技术、散热设备贴近发热源趋势

AIGC的大算力需求需要更高散热效率。面临散热不足、能耗过大的问题。基于冷却介质的选择上，***液冷有望替代风冷达到更高效的散热预期。此外，散热预计将从房间级、机柜级、服务器级向芯片级演进，更加贴近核心发热源。***

数据中心

AIGC对底层算力的训练及推理需求将***推动AI基建扩容***，通过加载AI服务器建设智算中心化解AI计算需求爆发与传统算力不足的供需冲突。由此爆发的数据洪流，***对基础设施的运算、网络、储存及安全管理的处理效能同样提出更高要求。***

商业模式带动

云端算力适配AIGC浪潮提出MaaS新业态

1）算力租赁服务　2）数据中心自建，配置更多AI服务器，并加速建设智算中心

3）提出MaaS新业态，***以算力资源承载云服务产品，赋能千行百业***

端侧仍在初探，探究大模型规模与端侧芯片的适配

随着大模型轻量级开源版本的发布，未来大模型有望进行进一步剪裁优化，***将推理能力部署在端侧，并带动手机、机器人、汽车等端侧智能芯片发展。***

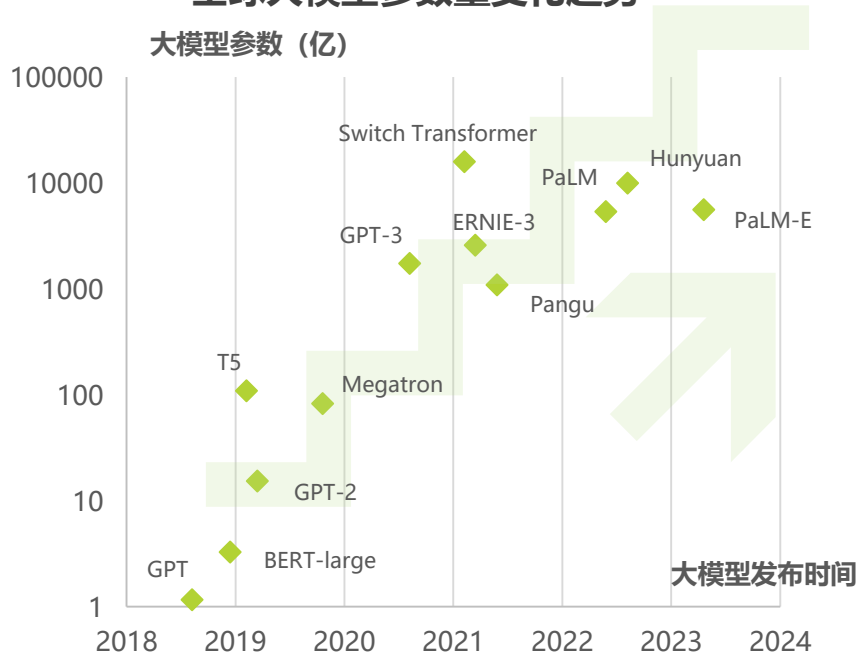
来源：艾瑞咨询研究院根据公开资料自主研究绘制。

全球将大力发展算力基础设施建设

算力支撑与模型需求存在gap，AIGC的大算力需求让供需结构进一步承压

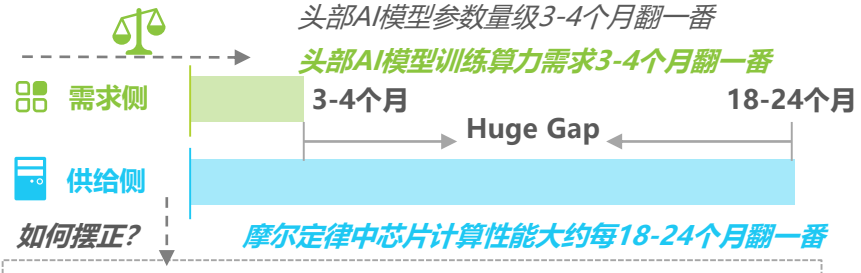
自2017年谷歌提出Transformer架构后，人工智能发展逐步迈入预训练大模型时代。2018年6月，OpenAI的GPT模型参数量已经达到1.17亿，模型参数量开始实现亿级基底的飞越发展，平均每3-4个月即呈现翻倍态势，由此带来训练算力需求也“水涨船高”。算力正在成为影响国家综合实力和经济发展的关键性要素。浪潮信息发布的相关报告表明，计算力指数平均每提高1个点，数字经济和GDP将分别增长千分之3.3和千分之1.8。面对算力层的供需结构矛盾，各国积极发展算力层基础设施建设。在计算力指数国家排名中，美国坐拥全球最多超大规模数据中心，以75分位列国家计算力指数排名第一，中国获得66分位列第二，随后为日本、德国、英国等国，算力建设已然成为国家高质量发展的战略级方针。2022年末，在OpenAI的GPT模型涌现能力后，AI产业迅速进入以大模型为技术支撑的AIGC时代，巨量训推算力需求让本就供需不平的算力产业结构进一步承压。目前中国各地正加快新一批数据中心与智算建设，持续优化算力资源，满足未来高速发展的大算力需求。

全球大模型参数量变化趋势



来源：艾瑞咨询研究院根据公开资料自主研究绘制。

AIGC算力供给与需求的倾斜天平



1) 头部AI芯片产品性能以超越摩尔定律的速度在加速翻倍迭代中

NV AI算力	2017	2020	2022
产品型号及对应算力峰值	V100 – 125 TFlops (SXM2)	A100 – 312 TFlops (FP16)	H100 - 1,979 TFlops (FP16 SXM)

2) 各国对AI基础设施的大量布局，以数量增量来满足庞大算力需求

No.1 – 国家计算力指数75分

商业主导：坐拥**四大超大规模数据中心平台**，亚马逊、谷歌、Meta和微软。

No.2 – 国家计算力指数66分

商业+政府主导：顺应《**智能计算中心创新发展指南**》30~40+智算中心建设

来源：《2020全球计算力指数评估报告》，浪潮信息，艾瑞咨询研究院根据公开资料自主研究绘制。

©2023.8 iResearch Inc. www.iresearch.com.cn

算力释放顺应AI模型逻辑：先训练后推理

亿元美金训练算力需求打底，巨量边缘及端侧算力需求仍待释放

大模型需要历经训练、调优、二开与推理四个核心步骤。从算力应用角度出发，可拆解为训练算力与推理算力。

1）**训练算力需求=模型参数量*数据集token数*系数k。**由此可知，模型训练的算力需求与其参数量、数据集token数成正比关系。承载千亿、万亿级参数的AIGC预训练大模型需要巨额算力支撑，对相关厂商高筑算力门槛。以OpenAI训练GPT模型为参考，1750亿参数模型训练约需要3.14*10^23次浮点运算，对应10天训练时间消耗约一万张GPU。国内厂商算力资源紧张，且此前购买的大部分GPU被常规业务占用。为了训练自身国产大模型，部分厂商采取算力资源调配策略，以大模型训练为优先级实现资源汇聚。

2）**推理算力需求 = 模型参数量 * （“输入+输出”token数）* 系数k。**对于云端大模型来说，推理算力需要支撑成百上千万用户频繁应用。2023年4月6日，ChatGPT就曾因需求量太大而暂停升级服务，并停止Plus付费项目的销售。国内大模型厂商也因推理端算力资源容量而限制AIGC大模型的公开测试名额。AIGC初期，市场把更多目光放到模型预训练大算力层，关注GPU等高算力芯片的资源供给及消耗。未来，顺应AI模型先训练后推理逻辑，AIGC算力层将逐渐迎来推理算力更广泛开阔的主场，带来更多分散性机会。

AIGC算力需求逐层释放逻辑图

注：系数k与模型种类相关，Encoder-only与Decoder-only系数为6，Encoder-Decoder的系数为3

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

算力产业模式将在AIGC时代有所演变

智能算力资源或将更多承载于云服务产品，以MaaS模式服务千行百业

过去数据中心以租赁与自建为主，算力需求方基于自身业务量级、财务预算情况、数据隐私要求等情况进行选择租赁或自建。在AIGC时代背景下，数据中心将配置更多AI服务器满足日益剧增的智能算力需求，云厂商更是提出MaaS（Model as a Service）模型即服务的商业模式，将云计算、智能算力、模型能力等资源做高度融合，客户可以直接在云端调用、开发与部署模型，更好适配于客户的个性化需求。未来，更多长尾企业的需求体量将拥抱MaaS商业模式。相较于云端算力发展，端侧大模型虽然发展较缓，仍是各家终端厂商发展的技术焦点，如从苹果招聘信息中可观测到其对“在端侧实现推理和加速大语言模型功能“的人才需求及产品规划。未来，随着大模型轻量级开源版本的发布，大模型有望进行进一步剪裁优化，将推理能力部署在端侧，并带动手机、机器人等端侧芯片发展。

中国AIGC算力产业模式洞察

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

数据中心需对高速巨量运转需求做出应对

大模型时代下，数据中心将进一步优化网络带宽、能源消耗与散热运维等

预训练大模型的训练推理需要巨量数据资源与高性能计算机的全天候高速运转，对数据中心的网络带宽、能源消耗与散热运维等能力提出更高要求。首先，网络是数据中心最为重要的组成部分，随着数据量与计算量的飞涨，数据中心需优化网络带宽，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；其次，能源消耗与碳排放问题是数据中心亟需关注的重点问题。普通服务器的标准功耗一般在750~1200W，而AI模型运行时会产生更多的能耗，以CPU+AI芯片（搭载4卡/8卡）异构服务器为例，系统功耗一般会达到1600W~6500W。根据斯坦福大学发布的《2023年AI指数报告》数据显示，GPT-3模型训练耗费的电力可供一个美国家庭使用数百年，CO_2排放量也相当于一个家庭排放近百年。由此，OpenAI创始人Sam Altma下注核聚变公司Helion Energy，向其投资了3.75亿美元，Helion Energy也已与微软签署购电协议，承诺将2028年之前把世界上第一台商业核聚变发电机接入电网，交付给微软；另一方面，基于大模型算力需求的高能耗运行，其热量释放呈现倍增态势。为了确保服务器能够长期处于适合的工作温度，数据中心将更注重系统设计和散热技术的发展应用。大模型散热需求加速由风冷到液冷的技术升级，进一步提升经济性、节能效果和散热效率等。散热也将更贴近发热源，由机柜级散热、服务器级到芯片级发展。目前，中国大力推进“东数西算” 工程，并发布《新型数据中心发展三年行动计划（2021-2023年）》等政策性文件，引导新型数据中心实现集约化、高密化、智能化建设，在AIGC时代下完成中国算力产业在规模、网络带宽、算力利用率、绿色能源使用率等方面的全方位提升。

大模型时代下的数据中心优化方向

网络带宽

高吞吐、低延迟的数据互联

性能效率

优化集群设计提升运行与调度效率

能源消耗

新型能源供给数据中心高能运行

碳排放

节能环保技术与设计，优化PUE能效

散热运维

催化散热液冷技术,优化中心运维能力

来源：《2023年AI指数报告》，斯坦福大学，艾瑞咨询研究院根据公开资料自主研究绘制。

AI芯片是算力皇冠，关注其性能与利用率

为服务于大模型的训推，AI芯片需进一步升级内存、带宽、互联等能力

算力是评价AI芯片的核心要素，而除了运算次数外，芯片的性能衡量还需考虑运算精度。基于运算数据精度不同，算力可分为双精度算力（FP64）、单精度算力（FP32）、半精度算力（FP16）及整型算力（INT8、INT4）。数字位数越高，代表运算精度越高，可支持的运算复杂程度越高，以此适配更广泛的AI应用场景。此外，AI芯片的性能峰值算力是指芯片能够输出的最大算力，而由于硬件架构的限制，算法模型特性，以及工具链，软件框架等各方面因素，AI芯片算力不会被百分之百充分利用。为了适配大模型的训练及推理，AI芯片要求有更大的内存访问带宽并减少内存访问延迟，由此带动由GDDR到HBM的技术升级，另一方面需要更高的片间互联甚至片内互联能力以满足AI加速器访存、交换数据的需求。最后，大集群不等于大算力，在大规模集群部署下，集群训练会引入额外通信成本，节点数越多算力利用率越低，且单点故障影响全局运行。因此，同比增加GPU卡数或计算节点，不能线性提升算力收益，中国面临的单卡芯片性能差距将更难通过堆料等方式解决。

服务AIGC大模型的AI芯片关键要素

1. **算力&精度**

   算力是AI芯片性能的核心表现标准，算力表现与运算精度相关，支持大模型训练的AI芯片**在满足太算力需求外，一般还要支持FP32以上的精度计算。**

2. **内存/显存&带宽**

   AI芯片需具备足够内存放置大模型数据，高带宽可提高芯片对数据吞吐量，因此**内存带宽影响数据处理量上限，进而影响训练结果准确性。**

3. **片间互联&片内互联能力**

   基于CPU+X的异构连接及多个GPU/AI芯片的互联需求，大模型训练时提供**多芯互联/片间高速互联、交换数据**的解决方案（如NV Link、Intel CXL）。

4. **AI框架支持度**

   大模型基于Transformer模型搭建，**对PyTorch、Tensorflow等主流AI框架的支持度会影响到AI模型训练的稳定性与开发效率。**

5. **软硬件产品适配及稳定性**

   AI芯片产品应用依赖于软件生态与工程能力，实现软硬件的适配协同与响应调整，**在发挥性能的同时尽量减少出现运行时出现断点、宕机的情况。**

6. **制程、能耗、价格**

   取决于芯片部署场景及规模量级。如部署在数据中心，规模越大越需考虑算力密度、运维成本、能耗散热等，极大影响数据中心的利用率与运行状况。

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

AI芯片算力性能拆解公式

算力 = **（单芯片）性能** × **规模/芯片数量** × **算力利用率**

- **定点运算**：INT采用定点数进行数值运算，常用单位为**TOPS**，代表芯片每秒能进行多少次定点运算。
- **浮点运算**：FP采用浮点数进行数值运算，常用单位为**FLOPS**，代表芯片每秒能进行多少次浮点运算，同样长度下浮点运算比定点运算数字表达范围更大，结果更精确。AI场景常用到FP16(半精度)/FP32(单精度)/FP64(双精度)。
- **不同场景对运算类型及精度要求不同**：FP64多满足HPC高精度场景；FP16/FP32可用于常见在图形处理、深度学习、人工智能领域的训练与推理场景；INT8精度相对较低，更多用于深度学习中的智慧识物、图库分类、人脸监测等分类推理问题。

- 性能峰值算力只反映AI芯片理论上的最大计算能力，而非在实际AI应用场景中的处理能力。
- 算力利用率区间跨度大，在10%-60%不等 **1）挖掘芯片算力潜力的同时，必须考虑算力的资源调度、芯片与算法模型匹配、算力与内存带宽匹配等问题。2）堆料难以解决芯片性能差距问题，数量堆积导致节点增加，进一步降低算力利用率**

影响算力利用率

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

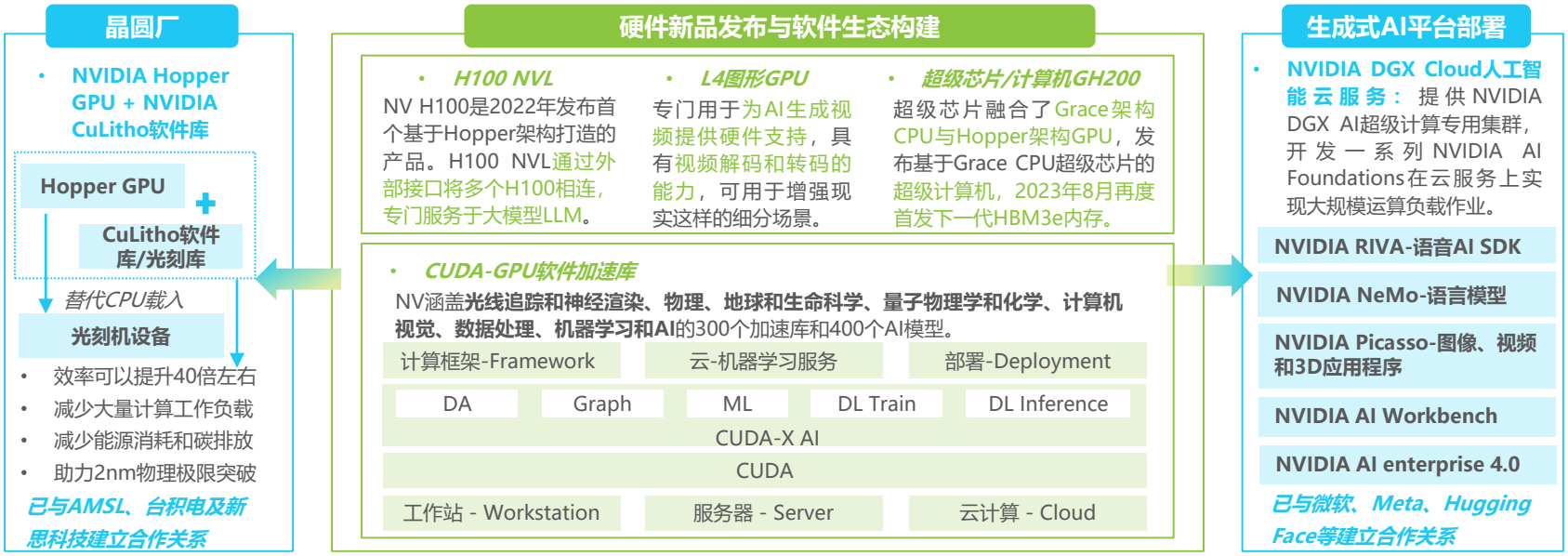
©2023.8 iResearch Inc. www.iresearch.com.cn

AI芯片巨头英伟达正由硬件拓展到平台

英伟达标榜AI的“iphone”时刻，All in “生成式AI”

受益于AIGC爆发，英伟达再度股价飞升，如今英伟达估值已达到1万亿美元，成功进入苹果、微软、谷歌、亚马逊所在的“万亿俱乐部”，成为美股有史以来首个市值触达1万亿美元的芯片公司。2023年3月，英伟达举办GTC（GPU Technology Conference）大会，介绍推出基于Hopper架构打造的产品H100、为AI生成视频提供硬件支持的L4图形GPU、融合了Grace架构CPU与Hopper架构GPU的GH200等重磅产品。此外，英伟达更是将产品布局延伸到上游，推出历时四年的cuLitho光刻库，与台积电、ASML和Synopsys等上游厂商合作，将计算光刻加速了40倍以上；并积极拓展下游模型应用场景，推出一系列围绕生成式AI发布的系列加速模型训练和推理软硬件产品及服务。8月，英伟达在计算机图形年会SIGGRAPH上宣布全球首发HBM3e内存——推出下一代GH200 Grace Hopper超级芯片，并宣布与Hugging face建立合作伙伴关系，助攻生成式AI模型的高效开发与部署。总结来看，英伟达早期以软硬产品结合策略构筑起AI芯片龙头地位，当下顺着生成式AI浪潮，英伟达已进一步开拓上下游布局，意图构建一套围绕产业上下游运转的应用开发生态，进一步加深公司技术与生态的护城河。

英伟达的AI生态版图



来源：艾瑞咨询研究院根据公开资料自主研究绘制。

中国算力产业将坚持自主创新道路

英伟达能否延续强者恒强？中国何时迎来自主创新芯片曙光？

作为AIGC产业的基建层，算力是AIGC生产力卡脖子的关键环节。对此，算力生产商纷纷发力，如AMD、英特尔等追赶型企业针对AIGC的产品新品动作频频。对标英伟达的Grace Hopper，AMD推出“CPU+GPU”双架构的Instinct MI 300进军AI训练端。英特尔即将在2025年发布Falcon Shores GPU，将其混合架构改为纯GPU解决方案。全球掀起一阵GPU采购热潮，马斯克抢购一万张卡加入AIGC大战，而国内厂商除过往存货外，受中美禁令限制仅能采购英伟达H版GPU，在算力及带宽方面受到极大限制。目前，国内大模型训练芯片仍以英伟达GPU为主，且英伟达作为首批训推部署框架成品及平台生态将进一步巩固其在生成式AI的优势地位，但国内客户正积极与海内外追赶型企业如AMD接触，意图打破英伟达的溢价与垄断体系。自2018年以来，美国陆续对中国企业实行贸易管制，进入到美方黑名单上的中国企业已达到了千余家，尤其在半导体、人工智能等先进科技领域，国产芯片实现自主创新迫在眉睫，中国科技部也陆续出台政策推动人工智能公共算力平台建设。目前国产芯片虽在成片进度有所突破，但整体还尚未进入成熟期。以适配AIGC大模型训练角度出发，国产产品会出现宕机、兼容性差、AI框架支持度低及核心IP受限等过渡性问题。在AIGC浪潮下，AI芯片发展路径更加聚焦于AISC品类，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”为两条主线发展路径，静待国产替代曙光，实现国产“算力+应用”的正循环。

国内外AIGC芯片发展概况

海外：领先者、追赶者、诸多尝试者

“领先者”

英伟达：以GPU通用性及性能等优势获得头部地位，尤其训练场景，并试图进一步拉大在产品性能及产业链上下游优势。

谷歌：研发出专为机器学习定制的专用芯片（ASIC）TPU，支撑了谷歌的搜索、语音/图像识别、自然语言处理等业务。

“追赶者”

AMD MI300/388：产品测试推进中，对标英伟达GH200，可用于HPC、AI计算，支持CUDA生态兼容仍存在一定差距。

Intel Gaudi2：收购Habana旗下产品，Gaudi2对标英伟达A100，开源oneAPI生态对标英伟达CUDA生态。

“尝试者”

Meta：计划开发一款内部芯片MITA，预计于2025年推出，芯片类型为ASIC，用来加速AI训练和推理。

中国：国内产品各有优劣，整体尚未进入成熟期

训练 *与国外头部厂商在硬件产品存在1-2代际差距，生态弱*

推理 *产品成熟度较高，但规模化应用仍有门槛*

01 产品稳定性

E.g. 基于一些早期架构问题需解决大规模并行环境下的产品可靠性，减少出现宕机情况。

02 兼容性

E.g. 若顺应谷歌路线，则芯片只能基于自身开发框架运行大模型业务，会在通用性上受到限制。

03 AI框架支持度及生态

E.g. 如果AI芯片对Tensorflow、PyTorch等AI框架支持度较低，指令兼容性差，影响开发及部署效率。

04 受禁令限制影响

E.g. 部分核心IP（如HBM、PCIE接口等）及代工工艺被美国所把控，对华禁令极大影响国产公司的设计及投片。

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

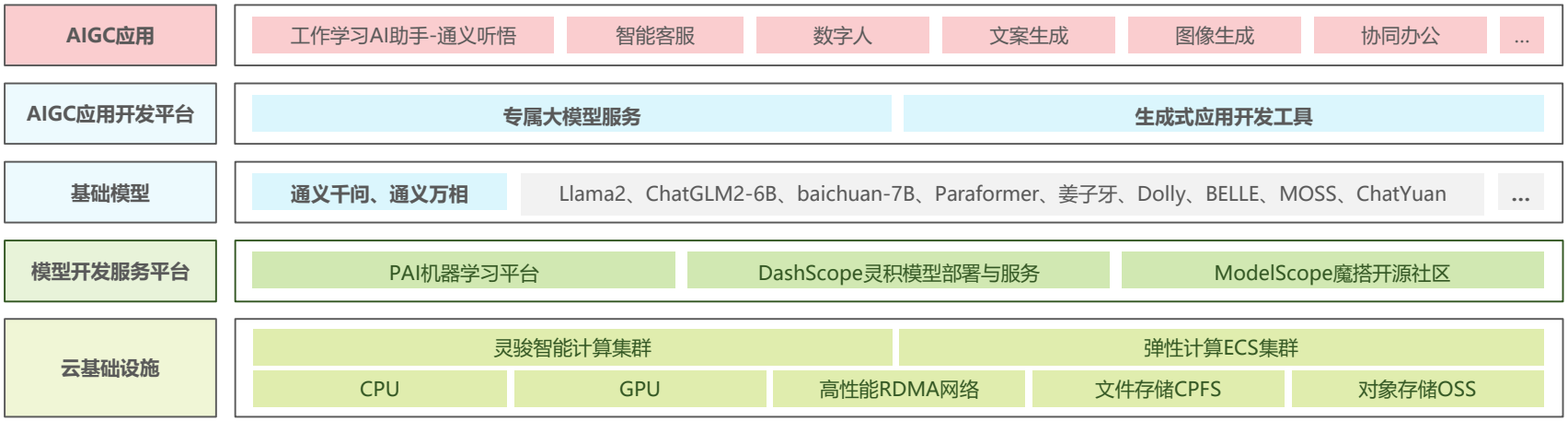
05/ 中国AIGC产业之标杆企业

Case

以模型为中心的AI开发新范式，打造MaaS平台服务

2022年，阿里云在业界率先提出大模型时代的“MaaS模型即服务”——以模型为中心的AI开发新范式，重新定义了云计算技术和服务架构，集成先进的大模型、机器学习、云基础设施等服务，为企业和开发者提供模型开发、应用开发、算力基础设施等的全方位服务。AI的发展进入大模型时代，参数规模的快速扩大带来对超大规模训练集群的需求，模型开发者难以高效使用计算资源并确保稳定性，造成大模型训练成本的高企。阿里云机器学习平台PAI可提供一站式的机器学习工程平台，覆盖模型开发、训练、调优、推理和部署全流程，降低用户开展大模型研发的门槛。结合灵骏智能计算集群，单训练任务可扩展至万卡级别，训练性能提高近10倍，千卡规模的线性扩展效率达92%，帮助客户高效实现大参数规模的大模型训练和推理，应对高额算力成本的挑战。在大模型和服务层面，阿里云推出“通义千问”和“通义万相”大模型，完整覆盖文生文、文生图等核心应用场景。在此基础上，阿里云推出“通义专属大模型”，帮助用户结合行业数据和企业私有数据调优和训练专属大模型，生成个性化API。不仅限于模型本身，阿里云灵积模型服务平台DashScope可为用户提供灵活、弹性的大模型API和定制服务，覆盖阿里云通义千问以及诸多业内领先的开源大模型，帮助用户快速基于大模型构建生成式应用。在开源生态层面，阿里云秉持开源开放的理念，在2022年联合CCF开源发展委员会推出魔搭社区ModelScope，仅半年时间便成为中国最大、最活跃的AI开源社区，吸引了超过180万开发者和超过1000个优质模型入驻。魔搭社区打通了与灵积平台的部署链路，支持社区的模型通过灵积快速实现服务化，加速中国AI开源生态的繁荣。

阿里云MaaS技术体系示意图



来源：阿里云官网，艾瑞咨询研究院根据公开资料自主研究绘制。

全面拥抱智能化时代，基于大模型重塑AIGC应用

生成式大模型和AIGC的爆火使传统应用迎来了革新的新机遇。阿里云在2023年4月发布“通义千问”的同时宣布将基于大模型“重做”一系列应用，已推出“通义听悟”工作学习AI助手，提供高精度的语音转写、文件转写、翻译、智能生成纪要和待办，大幅提升用户在会议、学习、访谈场景下的生产效率。在更广泛的协同办公场景中，钉钉接入“通义千问”大模型。在个人办公场景，用户可通过钉钉斜杠“/”随时唤起AI，获得邮件内容、策划方案生成、文生图、图生图等多种AIGC服务，全面辅助办公。在应用开发场景，钉钉可基于一张功能草图，无需代码输入，快速生成轻应用。生态合作方面，阿里云发布“通义千问伙伴计划”，以MaaS服务体系为基础，携手金融、交通、能源、通信、电力、酒店多个行业及通用场景的合作伙伴，加速基于大模型的AIGC应用落地。

阿里云AIGC架构体系

阿里云AIGC应用

通义听悟——工作学习AI助手

实时语音识别	音视频语音识别	发言人角色区分	音字回听
全文概要生成	章节速览	发言总结	关键词识别
问答回顾	PPT提取	待办事项生成	一键笔记生成
内容规整	内容问答	全文总结	文字解释

钉钉——智能协同办公

内容创作	会议交流	协同办公	应用创新
• 拟写标题 • 润色文案 • 扩写方案	• 总结会议纪要 • 生成待办事项	• 基于海量聊天记录，自动生成摘要	• 基于草图，快速生成轻应用 • 自动学习资料，化身专属客服

应用领域

通用场景 | 金融 | 法律 | 教育 | 文娱 | 媒体 | 交通 | 能源 | 通信 | 电力 | 酒店

通义千问伙伴计划

用友网络 | 中金财富 | 千方科技 | 昆仑数智 | 亚信科技 | 朗新科技 | 石基信息

来源：阿里云官网，艾瑞咨询研究院根据公开资料自主研究绘制。

第四范式

以生成式AI重构企业软件（AI-Generated Software）

第四范式是一家以机器学习为主的AI产品及技术提供商，也是中国决策智能市场的领军企业，至今已推动金融、零售、制造、能源电力等行业的多家头部企业的数字化转型进程。针对B端企业软件交互体验差、开发效率低的业务瓶颈，第四范式自主研发了生成式预训练大模型“式说”，并开发出以“式说”大模型为底座，通过多模态交互理解用户意图，并统一调用软件功能的AIGS服务。AIGS以知识库、Copilot、思维链CoT等能力为核心，具备良好的学习和扩展性，在不断充当用户与软件交互的代理人后，能够从基础的查找和应答能力，进化至根据既有规则自动执行任务，甚至在执行中不断学习和内化，形成全自动的任务执行能力，极大提升用户工作效率。

AIGS工作原理及阶段性能力演进

1.0 AI协助执行任务

核心能力：企业级Copilot

具备Copilot能力的大模型，与企业内部应用库、私有数据联网，调用企业AI模型、数据库及软件内置的数据、功能，“对话框式“完成任务。

应用场景

只需对话，AI就能替你查找到空闲会议室、撰写邮件、查询销售记录、执行生产计划...

例：通过对话让AI把照片亮度调亮20%。

2.0 AI参照规则执行工作

核心能力：企业级Copilot + 知识库

在阶段1的基础上，大模型能够参照企业内部知识库形成的规则，自动执行复杂工作（由多个任务组成）。

应用场景

AI替你完成海外市场调研报告、搭建网站...

例：AI查询了“人像美化”知识库后，能按照步骤完成修图工作。

3.0 AI自动执行任务

核心能力：企业级Copilot + 思维链CoT

具备Copilot和思维链CoT能力后，大模型在执行软件使用过程中，通过学习大量数据和规则，形成有步骤的推理，可以拆分并执行复杂工作。

应用场景

AI自动修改，无需人工调整干涉。

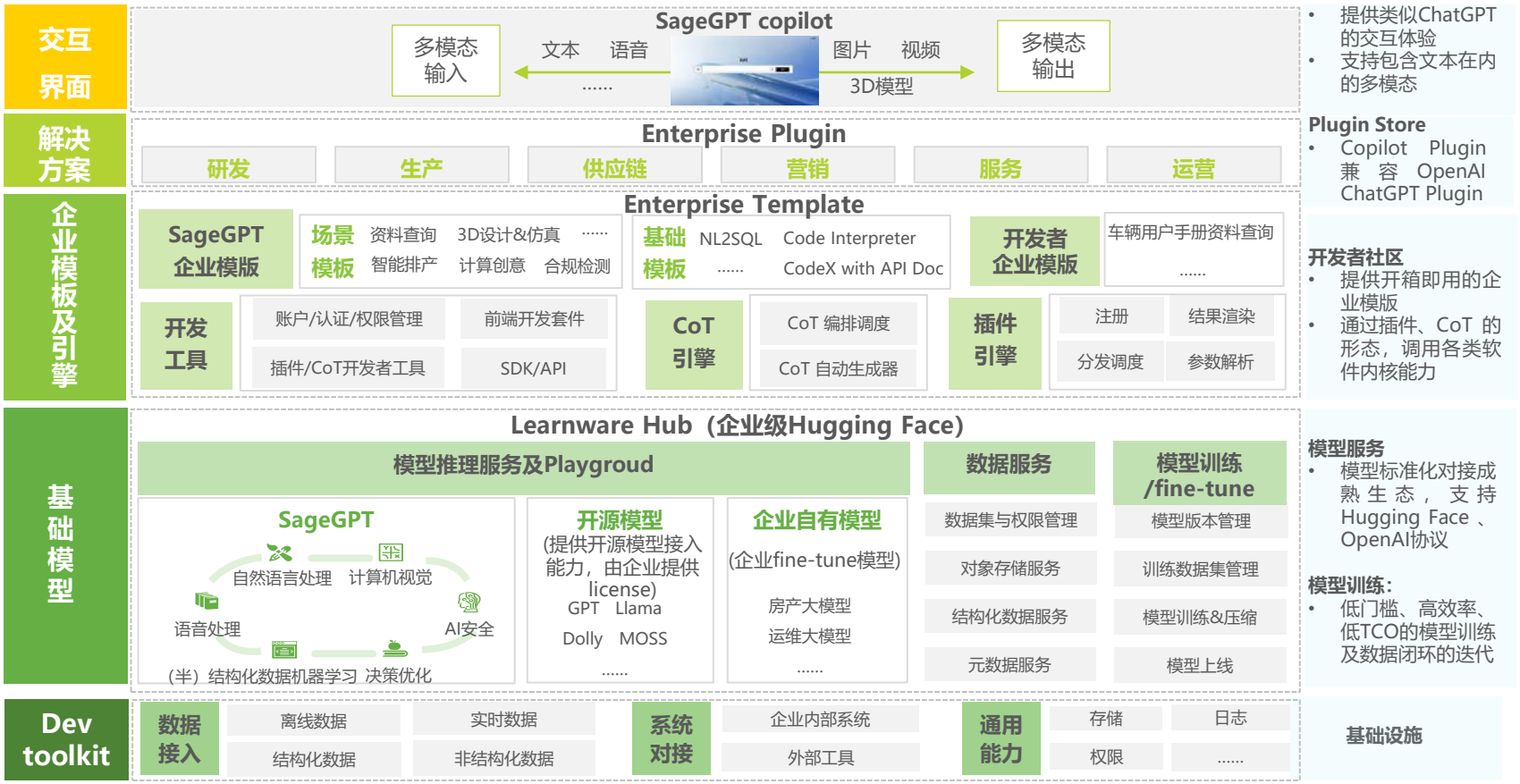
例：导入照片，AI直接进行优化修改。

来源：艾瑞咨询研究院自主研究绘制。

提供大模型落地全矩阵能力，全面保障业务效果

现阶段，大模型能力存在明显天花板，且从基础/行业大模型到应用落地之间存在诸多环节，厂商需要具备完善的全环节服务能力才能够帮助企业达到预期的应用效果。第四范式的大模型能力矩阵为企业提供大模型训练、微调、数据服务、应用开发、应用插件和多模态交互等全方位能力，帮助大模型达到最佳的落地效果，让大模型真正成为善于理解、准确执行的工作助手。

第四范式大模型落地能力矩阵



来源：艾瑞咨询研究院自主研究绘制。

云知声

具备全栈式AI技术能力的人工智能企业，自研并发布山海大模型

云知声成立于2012年，是一家拥有全栈式AI技术的人工智能企业，面向智慧物联与智慧医疗两大关键场景提供以AGI技术为基础的产品服务与综合解决方案。2023年5月，云知声发布了自研的“山海”大模型，其能力体系涵盖语言生成、语言理解、知识问答、逻辑推理、代码能力、数学能力、安全合规、领域增强等十大能力。云知声致力于打造MaaS 模式的AI 2.0解决方案，在通用能力基础上，增强物联、医疗等行业能力，为客户提供更智能、更灵活的解决方案，打开更大的 AI 技术产业化商业空间。迄今为止，云知声已经为华为、美的、格力、长虹、京东、北京协和医院、福建省立医院、平安集团、吉利汽车、民生银行等行业龙头企业提供了智能化产品和技术服务。

“山海”之力，十项全能，顶天立地

全栈AI技术体系

企业荣誉资质

国家级专精特新“小巨人”

北京市科技进步一等奖

连续5年入围CB Insights“全球人工智能独角兽榜单”

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

云知声

基于山海大模型技术底座，打造行业数字专家和智慧物联空间

在保持大模型高速演进的同时，云知声也在探索与具体场景深度融合的更多可能。云知声沿袭了一以贯之的U+X战略，即以“U”(AI技术和产品能力)，深度结合“X”(行业应用场景)，解决行业深层问题。山海大模型已开始全面接入并重塑云知声现有的各类人工智能应用场景，从效率、成本、体验等多角度，为千行百业的智慧升级按下加速键。目前，云知声山海大模型已深入到智慧医疗、智慧物联、智慧交通、智慧教育、智慧车载等行业中，致力于打造医疗、销售、知识管理、口语对话等懂行业的数字专家，以及智慧座舱等懂人的智慧物联空间，基于山海大模型打造的应用场景正不断丰富和拓展。同时，云知声已经和中建电子、京东科技、360等知名企业达成战略合作协议，共同构建山海生态合作体系。

“山海”之用，聚焦场景，因“地”制宜

行业数字专家

医疗专家

全面升级医疗业务线各产品智能化水平，实现从“助手”到“专家”的关键跃迁。

- 手术记录撰写
- 门诊病历撰写
- 医保智能审核

销售专家

助力销售管理者提升团队销售技能，把每个人都培养成金牌销售，不错过任何一个商机。

- “把脉”销售链路
- 精准洞察客户需求
- 打造差异化营销策略

知识管理专家

企业级New Bing！不限行业，秒级训练，私有化部署，提供精简式回答以及回答依据溯源。

- 专业知识辅助理解
- 问题回答精准溯源
- 多行业秒级训练

口语专家

智能情景对话，让学生声临其境；由浅入深，进行总体评价、语音推导、语法建议三维智评。

- 教你说的更标准
- 让你说得更准确
- 陪你说得更地道

智慧物联

智慧座舱

深度理解用户需求，提供一站式语音交互方案，满足高情商对话、行程规划、用车指南等多元需求。

- 打造理想人车交互
- 类人交互、复杂对话
- 出行专家、助理、伙伴

U+X 山海大模型赋能千行百业

行业：智慧医疗、智慧物联、智慧交通、智慧教育

企业：中建电子、京东科技、360 www.360.cn

战略合作协议

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

中关村科金

中关村科金

构建“人机协同”智能平台，助推企业智慧升级

中关村科金是领先的对话式AI技术解决方案提供商，致力于通过对话式AI技术构建人机协同的新型生产关系，帮助企业打造“超级员工”，实现具有分析决策能力的强人工智能应用。中关村科金依托自主研发的领域大模型、大数据分析、多模态交互三大核心技术，打造数字化洞察与营销、数字化服务与运营、数“智”底座三大矩阵，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，实现高效率、低成本、规模化的AI创新应用。目前已服务900余家头部企业的200多个应用场景，将AI创新广泛应用于金融、政务、零售、医疗、制造等行业，助力企业高效赢得客户，实现数智化转型升级。

中关村科金能力全景图

行业

银行 | 保险 | 信托 | 证券基金 | 互金 | 零售电商 | 汽车 | 大健康 | 公用事业 | 政务 | 制造业

应用

数字化洞察与营销

- 视频矩阵
- SCRM
- 智能助手
- 营销自动化
- 活动/权益管理
- 积分商城系统
- 智能质检
- 用户行为分析
- ABTest平台
- 客户洞察平台

数字化服务与运营

- 外呼机器人
- 文本机器人
- 智能IVR
- 云呼叫中心
- 音视频平台
- 可回溯系统
- RPA+
- 智能财务
- 智能培训
- 智能招聘

数“智”底座

对话式AI中台

- 数据采集
- 数据治理
- 数据存储
- 数据智能

客户策略平台

- 企业预置数据库
- 预置客户策略
- 对话式AI组件库
- 对话式AI工厂

AI人工智能能力平台

- OCR识别
- 多模态防伪
- 自训练平台
- 数字人

能力

多模态交互

- 交互策略
- 用户洞察
- 策略洞察

领域大模型

- 提示工程
- 领域知识库
- 领域适配
- 基础大模型

大数据分析

- 交互行为数据
- 一方业务数据
- 三方数据

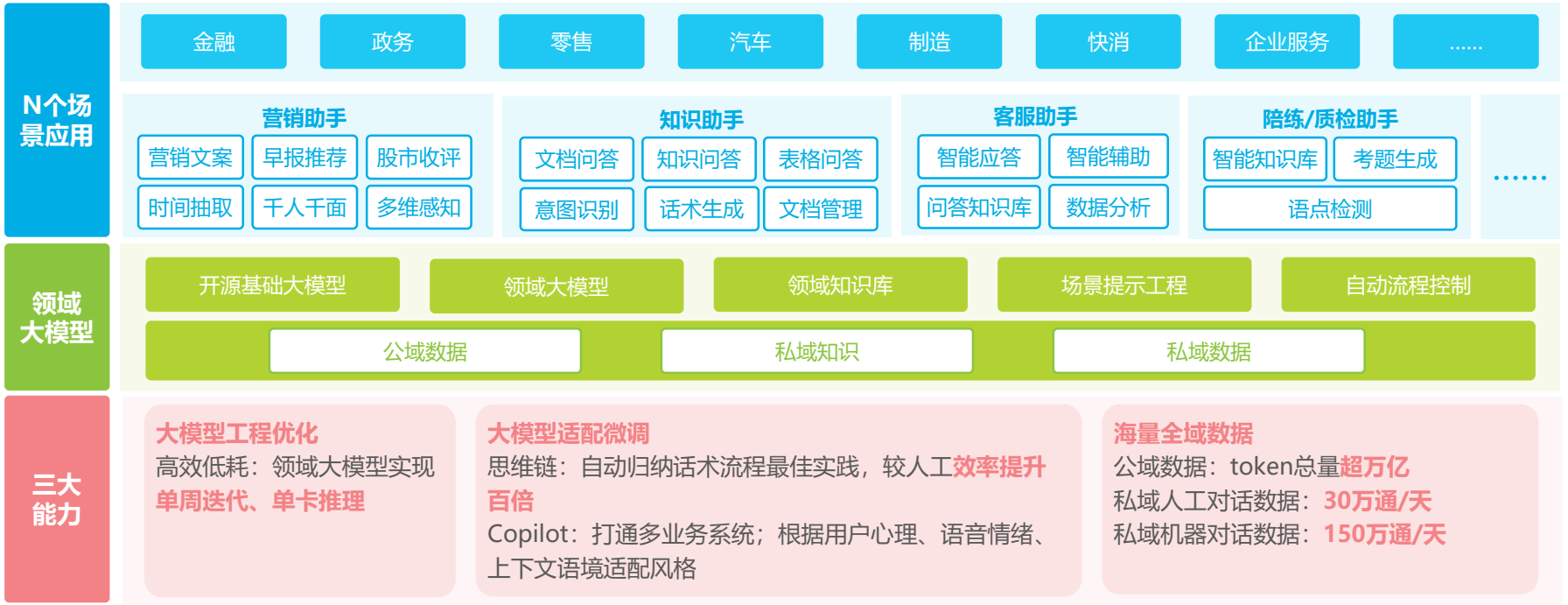
来源：艾瑞咨询研究院自主研究绘制。

中关村科金

为企业提供开箱即用、系统无缝衔接、成本可负担的专属领域大模型

中关村科金依托生成式大模型推出全新的“超级员工”系列AIGC应用，包括营销助手、知识助手、客服助手、陪练助手、质检助手等，为企业提供开箱即用、系统无缝衔接、成本可负担的专属领域大模型，以虚拟助手的形态打通企业对话场景数智化转型的“最后一公里”。中关村科金领域大模型应用已在诸多行业落地，金融领域，为诺亚财富打造智能知识库，以知识助手的形式赋能智能客服产品，大幅提升客服系统问答意图识别和回复的准确率，预期后期可减少70%以上的系统运营工作；营销助手为某财富公司理财师赋能，将营销文案撰写时间从10分钟缩短至10秒。政务领域落地医保全知大模型，帮助群众实现医保问题即问即答，一站式获取医保全量信息。零售行业落地话术师助手，原有30个话术师的工作量，现在2人即可完成，语义理解准确度从85%提升至94%。

中关村科金领域大模型架构



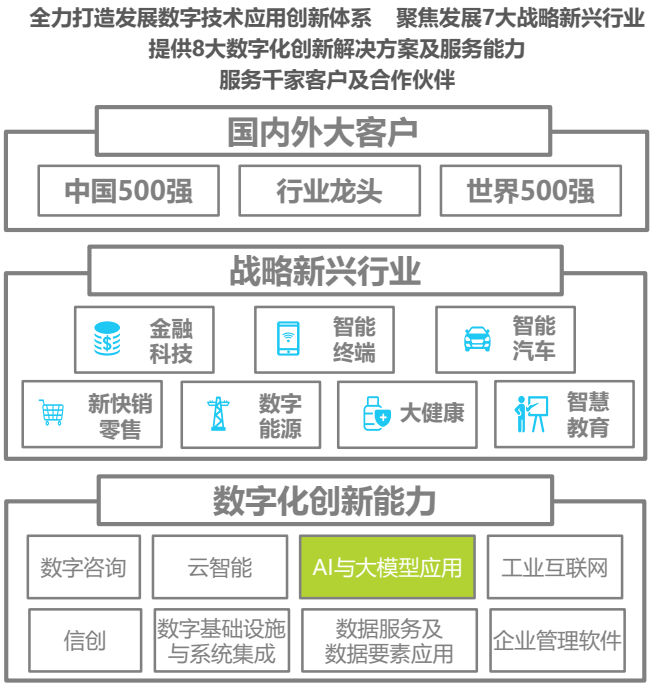
来源：艾瑞咨询研究院自主研究绘制。

软通动力

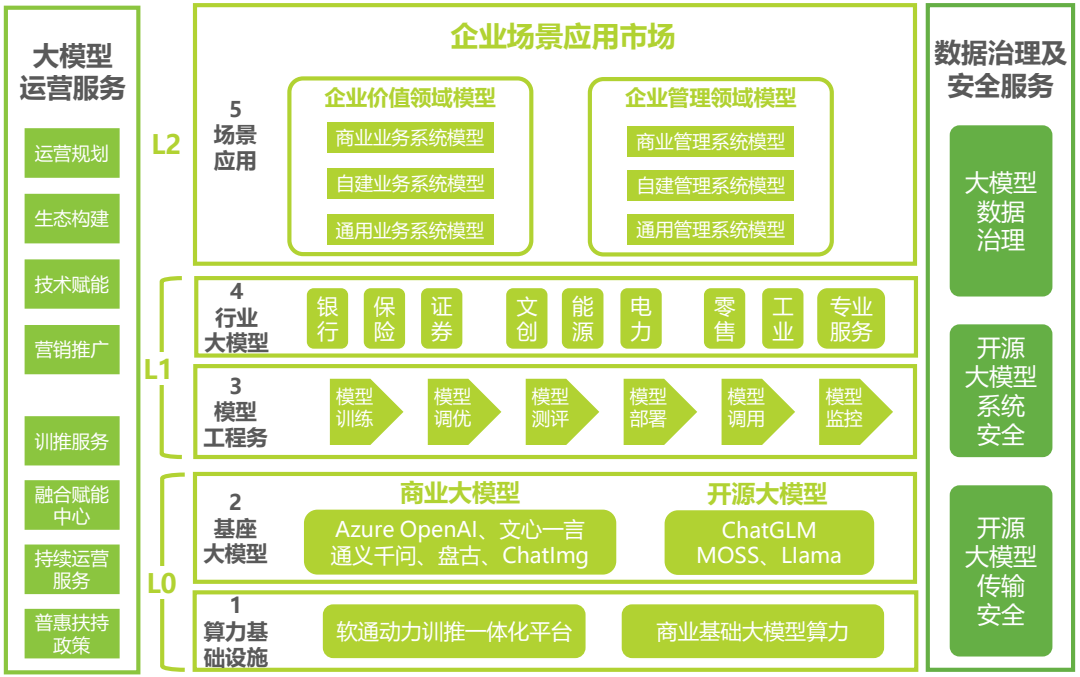
打造企业智能化核心解决方案和服务能力，释放企业价值

软通动力信息技术（集团）股份有限公司（简称“软通动力”）是中国领先的软件与信息技术服务商，致力于成为具有全球影响力的数字技术服务领导企业，企业数字化转型可信赖合作伙伴。2005年，公司成立于北京，目前在全球40余个城市设有近百个分支机构和超过20个全球交付中心，员工近90000人。软通动力长期以来在信息技术领域坚持自主可控，未来将积极推动行业信创实践，全力打造发展创新应用技术体系。软通动力已在10余个重要行业服务超过1100家国内外客户，其中服务过的世界500强和中国500强企业数量超过230家，500强企业中10年以上合作客户占比超过60%。作为大模型技术应用的赋能者，软通动力重点打造了“软通天璇2.0平台”， 基于L0级（大模型技术底座）、L1级（行业大模型及管理）、L2级（场景大模型应用）、大模型数据治理与安全、大模型一站式运营服务五要素，加速企业大模型工程化落地实施，助力企业智能化升级。

软通动力数字化业务发展布局



软通天璇2.0 MaaS 平台——大模型一站式运营服务

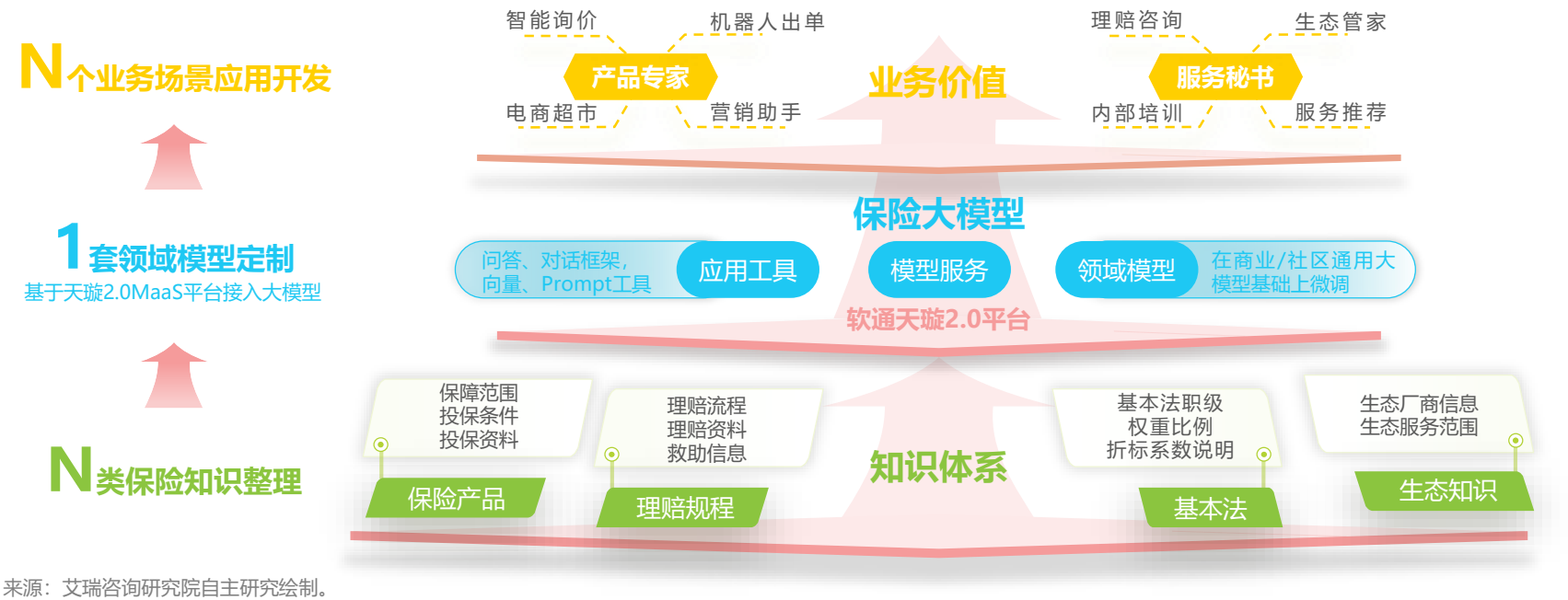


来源：艾瑞咨询研究院自主研究绘制。

保险行业案例：应用业务知识体系数据冲破增长瓶颈

随着AI大模型的产品技术发展，大模型的应用方案正不断向金融、法律、医疗等行业拓展。软通动力正在多个行业领域进行大模型服务探索，并已沉淀数个成功案例。以保险行业为例，保险行业面临着互联网和科技发展带来的客户需求变化和内部效能提升的挑战，基于软通动力的保险解决方案——保险业务中台&保险中台，软通天璇2.0平台接入大模型，为客户提供更为全面、精准、个性化的保险服务，贯穿保险业务的渠道销售、承保、理赔、收付、再保、客服等全业务流程，助力保险企业，实现产品生态化、营销个性化、服务场景化、风控智能化，赋能保险客户业务发展。大模型必将引发新一轮的技术和生产力变革，面对机遇和挑战，软通动力将在产品和服务、业务团队、合作伙伴等方面加大投入，构建完善的人工智能解决方案和服务体系，助力企业客户实现数字化转型，为中国数字经济高质量发展贡献力量。

软通天璇2.0 MaaS 平台——保险行业案例



来源：艾瑞咨询研究院自主研究绘制。

06/ 中国AIGC产业之发展趋势

Trend

AIGC的技术发展：科研与产业两端突围

中短期基于Transformer算法和结构优化仍是主流，长期可能被替代

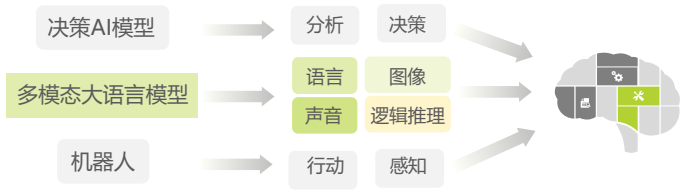
学术界将通过扩大模型参数量、调整模型结构、局部算法优化等方式，进一步探索大模型的能力天花板，触碰AGI可能性；以各大企业为代表的产业侧，一方面从商业化落地角度追求更小模型参数下的高模型能力维持，以及解决大模型出现的知识幻觉问题，一方面也在积极研发探索新模型架构可能性，呈现“对外模型名称为厂商能力代号，但内含技术架构随时可能改变”的发展特征。产业与科研两侧的需求都已经暴露标准Transformer架构的巨大瓶颈，即“不可能三角”。各大机构与开发团队对Transformer架构的成功改进在快速推进，未来极有可能会出现具备推广价值的新Transformer架构。

科研和产业侧大模型技术路径分化，均导向对Transformer的改进与颠覆

科研侧

大参数、多模态通往AGI

通用人工智能（AGI）是指能够像人一样，独立自主地处理各类问题与任务的人工智能技术，需具备全面感知、逻辑推理、自主学习、创作和行动等多种能力。基于Transformer架构的大语言模型已经实现在单模型中表达其中多种能力，即多模态，在不考虑训练推理成本和难度的前提下，这一架构和技术路径是目前多模态生成效果最佳的方式。同时，在大语言模型不具备的复杂分析、决策任务上，也出现了与决策式AI模型进行技术融合的新方向。



产业侧

兼顾成本与能力，减少知识幻觉

成本与能力可控

成本：大模型产业化的一大关键是突破应用成本的门槛，模型参数量需维持在10B到100B之间。通过剪枝、低秩分解、稀疏化等方式能够实现。

能力：基于目前大模型的水平，降解后的模型能力水平必须维持在原大模型的**80%**以上，才具备应用落地价值。

知识幻觉

知识幻觉，是指大模型会对自身不确定或完全无知识储备的内容，进行随机作答，造成答案完全偏离事实的现象。但当前能够改善知识幻觉的技术手段均会显著增加模型成本。

减少大模型知识幻觉的技术手段：

- **扩知识**：在训练时，针对大模型薄弱领域定向补充知识；
- **扩规模**：提升模型参数量级，增加推理的长度和厚度；
- **配合检索与核实**：让大模型在回答问题前尽可能多借助外部信息，同时增加思考过程。

Transformer不会是唯一解

Transformer架构的内生局限性：无法兼顾并行训练、可扩展性与低成本推理

并行训练

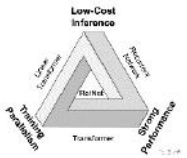
目前，提升模型智能化水平的主要手段仍是扩大参数量，为保证训练效率必须要实现并行训练，同时为了多次训练需求，模型必须具备可扩展性。

可扩展性

同时在模型参数庞大时，需要进行多次训练，同时为与外界知识保持同步，需要对模型知识定期更新，这些都要求模型必须具备可扩展性。

低成本推理

现有模型参数量使得其其推理速度收到较大影响，同时在科研侧，随着参数膨胀，成本也即将逼近单个机构天花板。



不可能三角

两界不断推出突破“不可能三角”的新架构

RetNet：算法优化和架构微调

在标准Transformer基础上，使用retention机制替代了标准的attention机制，大大提升了训练和推理速度。

RWKV：算法重构

用RNN改写的Transformer，同时具备二者优点，具备并行计算和高效推理的特点，但长上下文记忆能力弱于标准Transformer架构。

来源：艾瑞咨询研究院自主研究绘制。

AIGC的应用前景：软硬一体化

大模型低参版本的端侧应用，推动手机、机器人等物联网应用的升级进化

大模型在端侧的应用，软硬一体的结合带来广阔的应用场景。端侧的应用首先需要将大模型进行剪枝、稀疏化等处理，降低参数到十亿级规模，同时根据场景进行专属知识的训练和微调以适配专门的终端设备和软件。这对终端设备的功耗、内存、延迟、成本等都提出了新的要求。具体来看，目前在手机拍照、多终端语音助手、机器人具身智能（指从第一人称视角出发，具备理解、推理、并与物理世界互动的智能系统）等方面表现出应用前景，推动物联网应用的升级与进化。2023年8月，华为推出鸿蒙4引入盘古AI大模型，在消费电子领域赋能；小米官宣13亿参数手机大模型；OPPO预计将与阿里云联合打造OPPO大模型基础设施。手机厂商纷纷入局轻量化手机大模型市场，以期为用户带来全方位智能化体验提升，也许不久将来大模型应用将成为用户体验变革换代的“新触点”。

AIGC的应用展望

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

AIGC的社会影响：新一波自动化浪潮

AI成为基础设施，部分基础工作被替代，社会人力结构和分配方式重塑

技术的跃迁、生产效率的提升并不会自然带来社会整体福利水平的提升，相反往往以牺牲部分人的利益为前提，进而引发社会结构、分配方式的重塑。AIGC交互界面的用户友好性、大模型开源及API价格的降低、插件服务带来的应用生态繁荣等，都使得AI技术或将成为像水、电、网络一样的基础设施，渗透并改变千行万业。然而，大模型具有认知、分析、推理能力，不同于自动化浪潮下对蓝领的冲击，AIGC时代受AI影响最大的可能是初级专业人士和技术人员，即部分白领。据Challenger报告显示，2023年5月，美国因AI替代造成的失业人数达3900人，且都发生在科技行业。以及据麦肯锡数据预测，到2045年左右，有50%的工作将被AI替代，比此前的估计加速了10年。与此同时，具有创造力、深度思考等高阶智力的人才，将享受到AIGC带来的效率优势，成为AI的驾驭者，相应的工作需求也会增加。智力要素重要性的提升、附加值的提高，都将推动社会资源和财富向顶尖人才和组织聚集，但社会是一个整体，生产效率的提升并不代表着购买力的提升，被替代的普通职工才是购买力的最大来源，为了维持供需平衡，分配制度需要重塑。如美国总统竞选人杨安泽提出向大企业征收增值税，并向公民发放补贴，以及OpenAI创始人Sam Altman提出的世界币均等分配等，都通过反思并调整现有的分配方式，以驱动社会向更美好的方向演进。

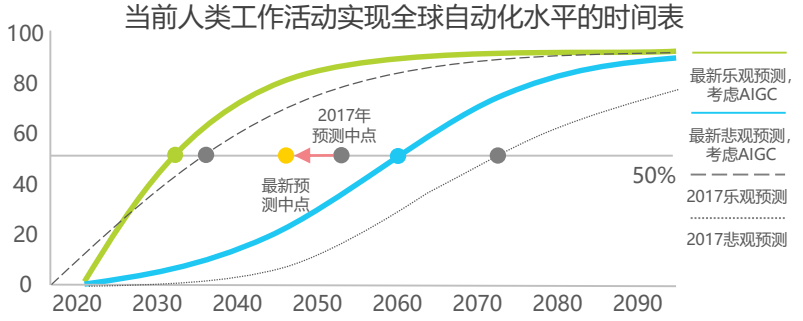
AIGC的社会影响

AI成为基础设施，基础工作被替代

AI大模型具备**认知、分析、推理能力**，冲击替代的对象涉及部分**专业化岗位**：

- **需要一定知识储备的辅助性工作**：法律案件整理等
- **精度及标准化要求不高的图文工作**：如海报设计、物料生成
- **专业性较高，但模型化程度也较高的工作**：如医疗诊断辅助等

当前人类工作活动实现全球自动化水平的时间表



人力结构转变

不同于**自动化浪潮**下，机器人对**蓝领**的替代，AIGC时代，受AI影响最大的是一些**初级专业人士和技术人员**，即部分**白领**的工作

具有**创造力、深度思考、探索能力**的人，将享受AIGC带来的效率优势。

麦肯锡《生成式人工智能和美国工作的未来》报告提出，2030年美国**工资最低**的岗位将**减少110万个**，但**工资最高**的岗位可能**增加380万个**

分配方式调整

智力要素的重要性提高、附加值提高，推动**社会资源向头部汇聚**

生产效率提升并不代表着购买力的提升，普通民众收入有限，**社会供需失衡**

再分配方式调整：向大企业征收增值税，向公民发放补贴，世界币计划等

来源：麦肯锡《生成式人工智能的经济潜力:下一个生产力前沿》，艾瑞咨询研究院根据公开资料自主研究绘制。

AIGC的监管展望：宽松鼓励与整顿规范

鼓励AIGC研究，放宽内容容错率，强调AI生成标识，推动公开数据建设

自AIGC逐渐应用以来，引发了知识幻觉、数据安全、个人隐私、道德伦理等诸多问题和讨论，新生的行业亟需监管措施的跟进和健康发展引导。2023年7月，网信办等七个部门正式发布了《生成式人工智能服务管理暂行办法》（以下简称《办法》），距离征求意见稿发布仅隔三个月，且监管要求更为宽松，反复强调了鼓励发展的态度。具体来看，《办法》主要规范公共服务环节，不包含有关专业机构的研发和应用环节，鼓励企业在自研自用范围加强技术攻关；其次，《办法》不强求生成内容的真实、准确性，放宽了容错率，对前期探索的企业带来一定利好，但同时也提高了用户辨别的时间和成本。同时，《办法》要求提供者对AI生成内容进行显著标识，有望从根本上杜绝AI生成内容难以辨别的问题，但也可能影响用户对内容的价值判断，对企业带来负面影响。最后，国家以立法的形式打造数据和算力协同共享的平台，最大化促进资源利用，有利于为中小型企业减负，降低研发成本。《办法》发布后，即引发了苹果应用商店对ChatGPT、讯飞星火等AIGC相关App的下架整改行动，行业整顿步伐进一步加速。

中国AIGC监管带来的机遇和挑战

监管政策		《生成式人工智能服务管理暂行办法》、《互联网信息服务深度合成管理规定》、《互联网信息服务算法推荐管理规定》	
监管维度		监管内容	机遇和挑战
涉及环节		行业组织、企业、教育和科研机构、公共文化机构、有关专业机构等研发、应用生成式人工智能技术，未向境内公众提供生成式人工智能服务的，不属于监管环节。	《办法》监管主要在于用户触达环节，**鼓励AIGC相关研究，鼓励企业在自研自用范围加强技术攻关，为研究发展与创新留足空间。**
内容监管	真实性	基于服务类型特点，采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性。	相比于征求意见稿中“保证数据的真实性、准确性、客观性、多样性”，《办法》缓和了表述，**对前期探索企业的容错率有所放宽**，但同时提高了用户辨别的时间和成本。
	生成标识	对可能导致公众混淆或者误认的，应当在生成或者编辑的信息内容的合理位置、区域进行显著标识，向公众提示深度合成情况，包含智能对话、智能写作、合成人声、人脸生成等服务。	AI生成的内容标识**可能影响用户对内容的价值判断，从而影响企业引流获客、产品单价等。**
	内容审核	提供者发现违法内容的，应当及时采取停止生成、停止传输、消除等处置措施，采取模型优化训练等措施进行整改，并向有关主管部门报告。	相比于征求意见稿，**取消了“3个月”整改的时间限制，对企业要求更加宽松。**但企业仍然需要建立内容审核与问题响应机制，**或对内容审核员、内容质检产品有较大需求。**
公开数据		推动生成式人工智能基础设施和公共训练数据资源平台建设。促进算力资源协同共享，提升算力资源利用效能。推动公共数据分类分级有序开放，扩展高质量的公共训练数据资源。	国家以立法的形式打造**数据和算力协同共享的平台**，最大化促进资源利用。有利于**为中小型研发企业减负，降低研发成本。**
境外服务		对来源于中华人民共和国境外向境内提供生成式人工智能服务不符合法律、行政法规和本办法规定的，国家网信部门应当通知有关机构采取技术措施和其他必要措施予以处置。	**调用境外API**向中国境内公众提供服务的，也**属于监管范畴**，企业**需注意合规问题。**
特定行业		• 国家对利用生成式人工智能服务从事新闻出版、影视制作、文艺创作等活动另有规定的，从其规定。 • 提供具有舆论属性或者社会动员能力的生成式人工智能服务的，应当按照国家有关规定开展安全评估，并履行算法备案和变更、注销备案手续。	**新闻出版、影视制作、文艺创作等领域**或有更具针对性的规范要求，**不确定性较大，需等待新规出台。** **“具有舆论属性或者社会动员能力的生成式人工智能服务”范围待界定**，需要细化的规范指导，相关行业及产品不确定性较大。

来源：艾瑞咨询研究院根据公开资料自主研究绘制。

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