2024年中国人工智能产业研究报告

PREFACE

前言

研究背景：

作为新一轮科技革命和产业变革的核心引擎，人工智能产业在2024年被中央及各地政府确立为重点发展方向，陆续出台了一系列针对性强、力度大的政策措施，旨在推动产业创新，提升区域经济的科技竞争力。经过多年持续投资布局，我国人工智能产业体系逐步完善，基础层、模型层及应用层不断升级优化，实现了人工智能、大数据等数据智能技术与实体经济的广泛融合。2025年2月，中共中央总书记、国家主席、中央军委主席习近平在京出席民营企业座谈会并发表重要讲话，强调民营企业的关键角色与发展前景，进一步强调了人工智能产业的战略地位。

2025年初，以DeepSeek为代表的国产开源大模型掀起热潮，其高性能、低成本的特点迅速吸引了国内外开发者和企业的关注，推动了中国AI生态的开放性和竞争力的进一步提升。这一风潮不仅加速了模型层的国产化创新，也为中小企业提供了更易获取的 AI 工具，激发了应用层的创新活力，成为中国AI产业发展的标志性事件。

艾瑞人工智能研究团队延续六年行业研究经验，在第七年聚焦人工智能产业的发展环境、产业进程及产品动态，深入探讨技术驱动、产业机遇、商业模式及挑战等核心议题，为市场提供前瞻性数据与深度洞察。

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

宏观环境

2024年，国家高度重视人工智能发展，将其纳入国家战略，各地政府积极推进科研创新与算力基础设施建设，并因地制宜出台特色政策。尽管GDP增速放缓，AI技术作为新质生产力，凭借其在提升效率和推动产业升级方面的优势，展现出广阔发展前景，政府支持也为其提供了强劲动能。资本市场持续关注AI，投资重点聚焦于语言与多模态模型应用、芯片、算力服务等领域，基础层与应用层协同发展，不断完善产业生态。社会层面，生成式AI的普及加速了市场教育，公众接受度显著提升，但就业替代、隐私安全等问题仍引发一定焦虑。技术方面，Transformer架构依然主导大模型发展，研发侧**通过强化学习、思维链优化提升模型推理能力**，同时加速跨模态融合，并在**推理效率优化和新型注意力机制等方面持续探索**，推动AI产业向更高水平迈进。

产业动态

1）市场规模增速略低于预期：2024年中国AI产业规模为2697亿元，增速26.2%，略低于预期。主要原因为**大模型在实际业务场景的表现未完全满足客户需求，且建设成本较高，较多项目仍处于探索阶段**。

2）算力需求结构性转变：2024年部分地区智算中心出现闲置，但这主要是供需错配导致。随着DeepSeek等开源模型推动推理应用爆发，推理侧算力需求大幅上涨，智算中心利用率有望逐步提高。

3) 工具生态日益完善：分布式AI框架、LLMOps平台和一体机产品等不断发展，深度融合软硬件优势，加速了大模型的训练与部署，有效支撑了产业侧大模型的应用建设。

4）商业化以项目制与订阅制为主流：政企侧客户以项目制为主，C端产品多采用“免费+订阅制”的模式。**新兴商业模式为按应用效果或功能点收费**，创新的模式可在降低客户采购决策成本的同时，倒逼供应商持续优化产品技术与服务。

5）全球化战略：面对国内激烈竞争，众多企业积极出海，布局海外市场，在图像、视频和社交等领域有较多突破。

6）DeepSeek掀起开源开放与应用落地的热潮：DeepSeek刷新了市场对大模型现阶段性能的认知，其开源策略结合高效、低成本的架构显著**加速了中国AI产业向更加高效、开放和自主的方向迈进，并带动产业链上下游的合作与应用落地**。

发展趋势

AI Agent正在重塑大模型的产品应用形态，带领AI产品由简单的对话问答向完成复杂任务的智能代理演进。作为连接数字智能与物理世界的关键技术，**具身智能**是下一代AI竞争的战略高地，其发展需要解决硬件加速和软件优化、跨行业生态协作等一系列挑战。**DeepSeek**的开源开放，推动了大模型技术的普惠与平权，将加速大模型在产业和消费领域的应用普及。构建面向新一代人工智能的**安全治理体系**至关重要，需要在技术、商业、法律、伦理等多个层面协同发力，以确保人工智能的安全发展。

CONTENTS

目录

01/ 中国人工智能产业宏观环境

—— 当下，中国人工智能产业
在经济、政策、认知、技术维度的发展环境如何？

中国人工智能产业政策环境

人工智能新时代的技术引擎，各城市展开地域大模型产业竞速

近年来，国家高度重视大模型产业发展，把“人工智能”纳入国家发展战略，并出台了一系列政策以推动技术创新、资源建设、标准建立与行业应用。随着中央层面人工智能政策的出台，以北京、上海、成都、深圳等代表的各地政府纷纷响应号召，将人工智能及其相关产业发展纳入当地发展规划，以助力新一代人工智能产业生态的形成。2025年初，习近平总书记指出：“中国高度重视人工智能发展，积极推动互联网、大数据、人工智能和实体经济深度融合，培育壮大智能产业，加快发展新质生产力，为高质量发展提供新动能。并在2月召开中央民营企业座谈会，众多与人工智能相关的民营企业家参会，为中国经济转型与产业升级打下重要基调，也进一步反映出未来中国人工智能产业发展的重要战略意义。

中国人工智能产业政策

高维规范建设

数据

- 2024年1月4日，十七部门印发《**“数据要素×”三年行动计划（2024—2026年）**》

算力

- 2024年5月29日，中央网信办等三部门印发**《信息化标准建设行动计划(2024—2027年)》**

规范

- 2024年9月9日，全国网络安全标准化技术委员会发布**《人工智能安全治理框架》1.0版。**

标准

- 2024年6月19日，四部门联合印发**《国家人工智能产业综合标准化体系建设指南（2024版）》**

技术

- 2024年3月18日，市场监管总局等18部门联合印发**《贯彻实施〈国家标准化发展纲要〉行动计划（2024—2025年）》**

全产业级地方鼓励

以**北京、上海**为代表，**成都、重庆、安徽、山东、深圳、湖南**等地方积极响应。

北京

- 北京市政府着重加大对科研机构和高校的资金投入，鼓励开展**人工智能基础理论研究，推动关键技术突破**；计划建设强大的**算力基础设施**，以支持AI模型的研发和应用；积极打造**人工智能产业创新高地**，吸引顶尖人工智能企业和人才汇聚。

上海

- 上海市政府着力于AI大模型的**产业集聚和生态建设**，通过实施大模型创新扶持计划和示范应用推进计划，**推动AI技术在金融、制造、生物医药等领域的应用**；推动**AI开源生态产业集群的建设**，以促进AI技术的创新和应用。

中央

“夯实产业根基，站在高维的视角，对人工智能产业各层次的发展进行标准制定。”

地方

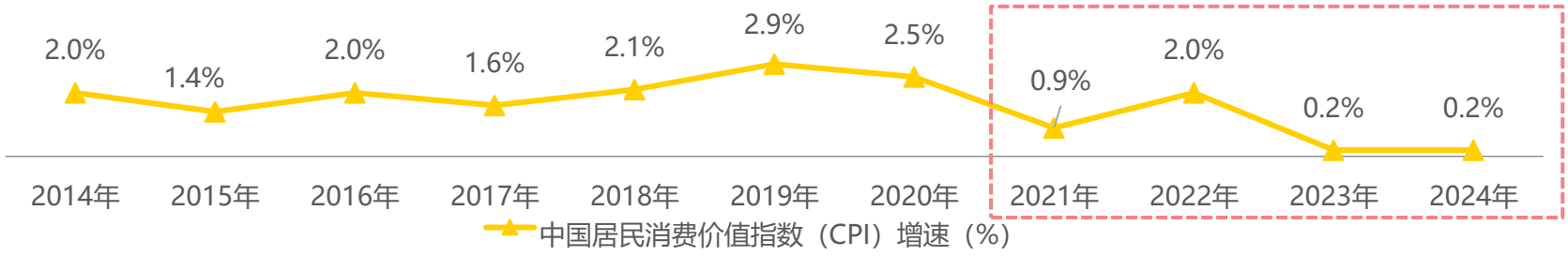
“各地因地制宜，充分发挥自身独特优势，出台具有地方特色的人工智能政策，呈现出百花齐放的发展态势。”

中国人工智能产业经济环境

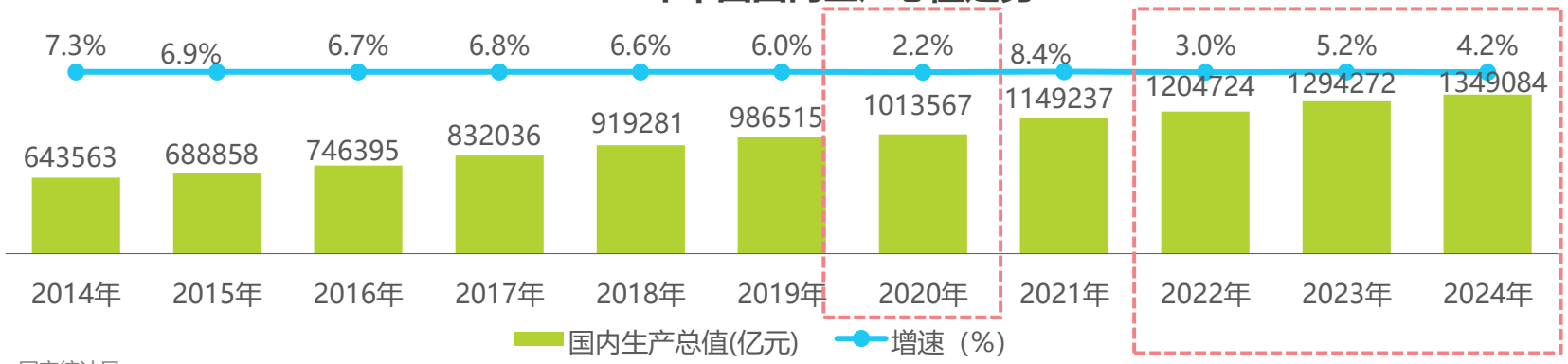
超预期因素反复冲击，GDP 增速放缓，CPI 低位运行

GDP（国内生产总值）是衡量一个国家或地区在特定时期内经济活动总量的重要指标，它代表了该时期内生产的所有最终产品和服务的市场价值总和。2020年，受全球新冠疫情的冲击，我国经济活动的增长速度有所减缓，GDP增速降至2.2%。随着疫情防控措施的有效实施和经济政策的积极支持，经济活动逐渐恢复，GDP增速有所提升，但整体水平仍低于疫情前的增长趋势。CPI（居民消费价格指数）是反映居民消费商品和服务价格水平变动情况的重要指标。近年来，我国CPI指数呈现下降趋势，显示出一定的通缩压力，表明经济环境相对低迷。这一经济形势对中国人工智能产业的发展带来了挑战与机遇。一方面，经济低迷导致投资减少、融资难度增加以及市场需求萎缩，对人工智能产业的增长产生了一定的负面影响。另一方面，人工智能技术在提升生产效率和创新业务模式方面的优势，可以有效促进经济增长和创造就业机会。在国家政策的大力支持引导下，人工智能产业的高质量发展，将进一步推动消费投资增长，助力中国经济的持续回温。

2014-2024年中国居民消费价格指数变化



2014-2024年中国国内生产总值走势



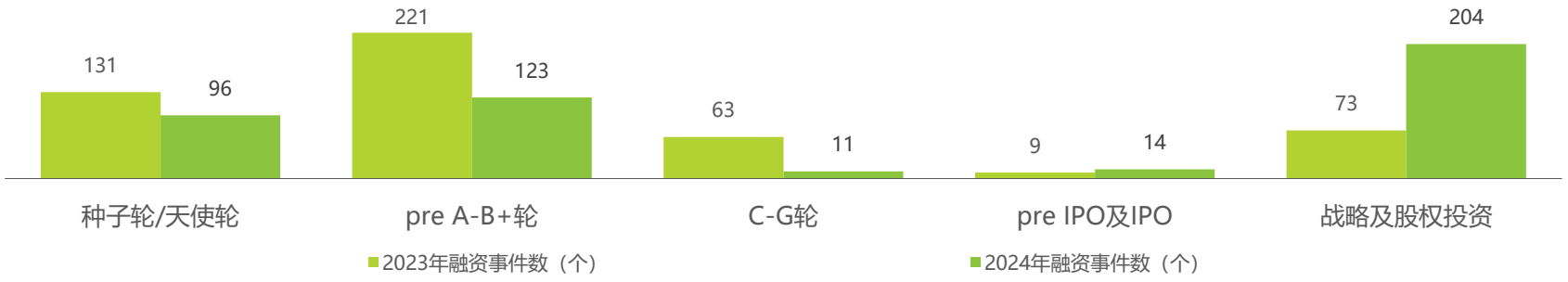
来源：国家统计局。

中国人工智能产业资本环境

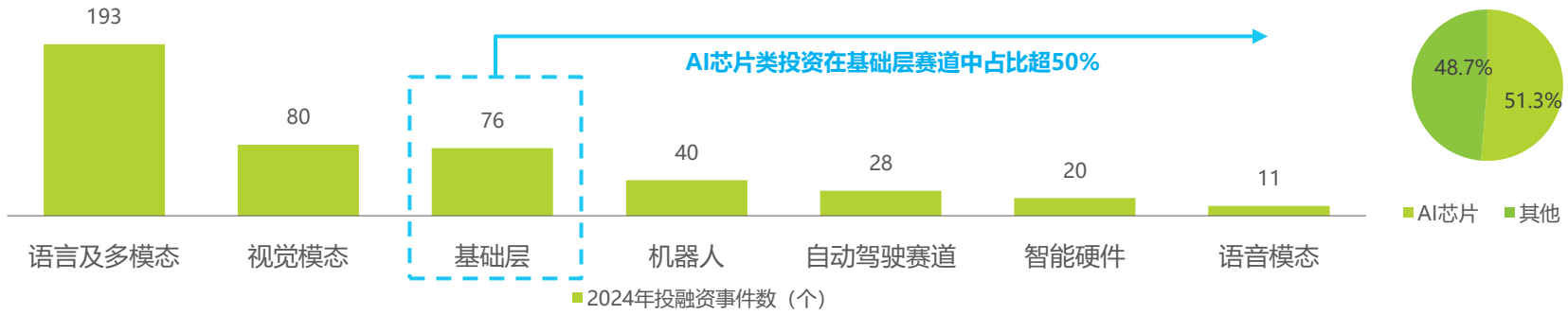
语言及多模态赛道目前最受瞩目，同时基础层厂商积极入局

从投资轮次的分布情况着眼，战略及股权投资的数量及占比均呈现出显著的上升态势，其中股权投资事件占比高达77.9%。而在应用赛道的投资分布方面，语言及多模态赛道目前已成为最受瞩目的投资领域。与此同时，以AI芯片、AI算力解决方案、算法架构等为代表的基础层投融资数量显著上升，其中AI芯片产品的投资占比约为50%，这表明应用层的快速发展正有力地带动基础层的建设，我国人工智能产业生态也因此得到进一步完善。

2023-2024年中国人工智能产业投融资轮次数量及其分布情况



2024年人工智能产业各应用赛道投资数量及其分布情况

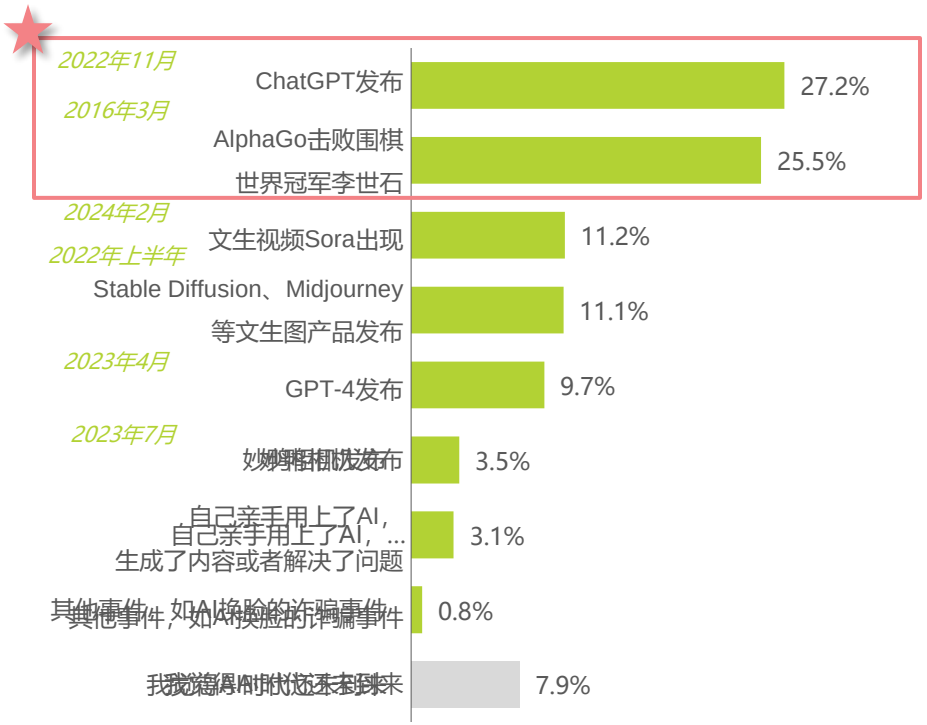


中国人工智能产业社会环境（1/2）

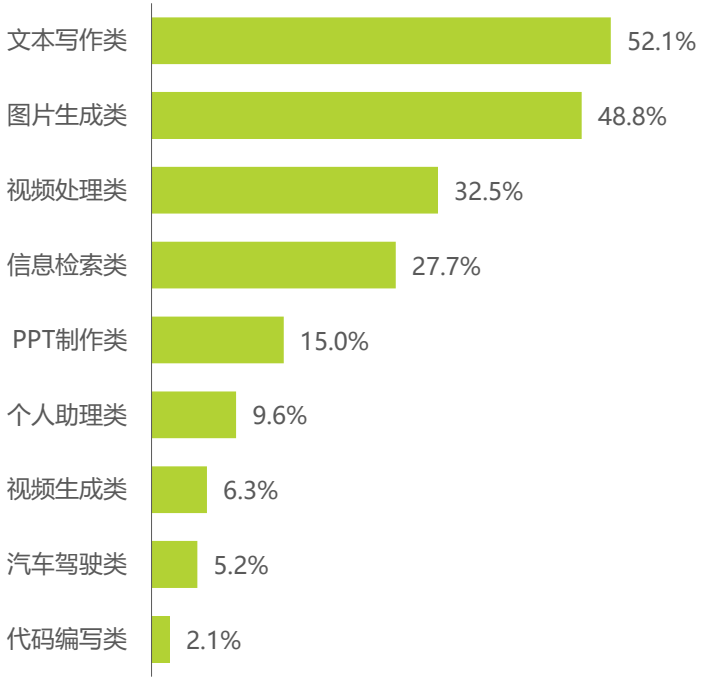
AI观念逐渐“深入人心”，大模型热潮极大助力中国人工智能市场教育

在澎湃新闻·对齐 Lab 对普通人如何使用和看待 AI 的调查《2024年人工智能公众态度调查报告》中， 有27.2%的人认为2022年11月ChatGPT的发布是“AI到来的标志事件”。C端 ChatGPT 产品的出现让公众直观感受到，AI可以理解复杂的语言指令并生成流畅自然文本的强大功能，极大突破了以往人们对AI通常进行简单任务处理的认知。AI、AIGC、大模型快速成为近两年科技产业发展的高频关键词，政府侧、企业侧纷纷加大对AI技术投资以释放大模型生产力，消费者对生成式AI工具产品的兴趣也在增加，其中，文本写作类应用（豆包、Kimi、文心一言等）、图片生成类应用（文心一格、通义万相等）是大家主要尝试的两大AI功能方向。

中国民众认为“AI到来的标志事件”



过去一年大家主要尝试过的AI功能



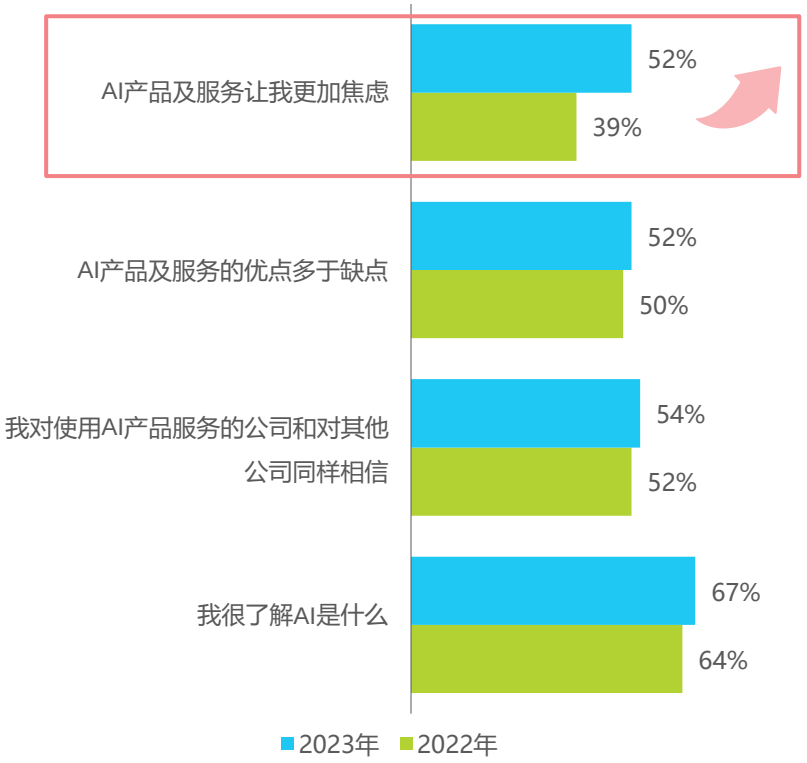
来源：《2024年人工智能公众态度调查报告》，澎湃调研。艾瑞咨询研究院研究绘制。

中国人工智能产业社会环境（2/2）

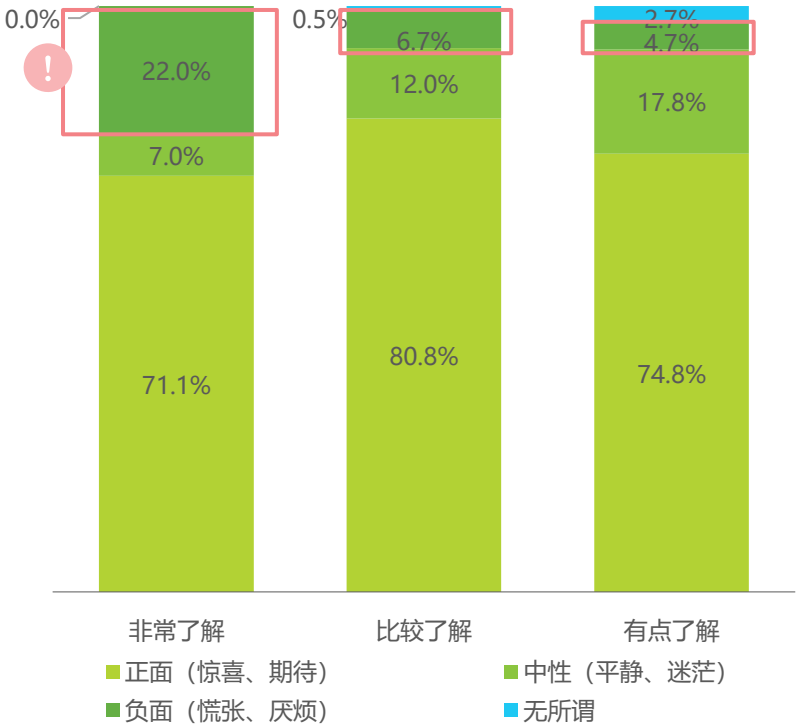
同时，人们对于AI带来的焦虑与不安情绪也在加重

而在调研中，有超过半数的人认为“AI产品及服务让我更加焦虑”。AI 技术的发展应用进一步对社会中重复性、规律性强的工作岗位带来威胁，相较于自动化浪潮对蓝领的冲击，一些初级专业人士和技术人员等职业，如工厂流水线工人、数据分析员、财务法律专员、客户服务等职业可能会被数字员工、AI Agent所取代。此外，人们对AI展示了“慌张”、“厌烦”等负面情绪，且对AI越了解的人，负面情绪占比越高。随着AI进一步广泛应用，相关安全隐私事件频出，大模型能力可能会被恶意利用，用于制造虚假不雅信息、网络攻击、诈骗事件、恐怖活动等，从而对社会安全和稳定造成威胁。

2022年-2023年全球对AI产品及服务的看法



不同AI了解程度的中国民众对AI的情绪分布



来源：《2024年人工智能公众态度调查报告》

中国人工智能产业技术环境（1/2）

CNN与RNN为典型小模型架构，Transformer已奠定当今大模型架构基础

人工智能产业典型模型架构演进历程

相较于大模型，小模型在一些领域会更具应用优势： *1）小模型成熟度高 2）小模型使用成本低 3）对小模型替代成本高*

大模型应用逻辑： *1）**替代逻辑**-小模型既有场景，但大模型的效果更好 2）**可行逻辑**-原本小模型在某些场景能力无法达到，大模型具备可行性 3）**创新逻辑**-大模型发掘了客户需求，在需求侧未提出要求情况下创造新场景需求*

Transformer架构

- 2017年，Google颠覆性地提出了基于**自注意力机制的神经网络结构Transformer架构**，奠定大模型预训练算法架构的基础；
- 2018年，OpenAI发布了GPT-1大模型；Google发布BERT大模型；
- 之后GPT模型持续演进，**2022年11月，GPT3.5的ChatGPT面世，引爆互联网，大模型时代随之到来。**

2022年-2024年，在大语言模型之外，Transformer架构更多融入语音、视觉领域，发展端到端的语音大模型、以DiT、ViT为代表的视觉大模型。

各家积极发展结合强化学习、思维链的“后训练“，推出深度**推理模型。**在效率优化方面，稀疏注意力、线性注意力等相关机制可大幅降低内存和计算成本。

正朝着处理更长序列、更大规模数据和实时应用场景的方向发展，新型高效注意力算法和硬件协同优化可能成为重要突破口。

Diffusion架构

2015年，**扩散概率模型**的基本概念与整体框架被提出，2020-2021年，Diffusion Model在图像生成领域得到广泛应用。

Diffusion Model是一种基于概率生成的深度学习模型，通过模拟数据从有序到无序再到有序的过程，实现从噪声中生成高质量数据样本，**应用于图像生成、图像修复、图像转换、视频生成等方向。**

GAN架构

2014年，**GAN（对抗式生成网络）**诞生，深度学习进入了**生成模型研究**的新阶段。

GAN由两个神经网络，判别器与生成器组成，在生成图像、声音和文本等数据方面表现优异，应用于**样本数据生成、图像生成、图像修复、图像转换、文本生成**等方向。

扩散模型在视觉效果和多样性上表现优异，但计算成本较高；GAN可能存在训练不稳定和模式崩溃的问题，但在一些任务中能实现较快的生成速度。已有研究在尝试融合两者的优点，以在生成效果和效率之间找到更好的平衡。

CNN与RNN架构

- 1980年，**卷积神经网络的雏形CNN**诞生；1998年，**现代卷积神经网络**的基本结构LeNet-5诞生。
- 循环神经网络（RNN）和长短时记忆网络（LSTM）等结构的出现，使得CNN与RNN能够相互融合，形成了更加复杂的模型结构。

里程碑事件：2006年深度神经网络引入；2012年AlexNet ImageNet图像识别大赛让图像领域飞跃式发展

CNN 适用于处理空间结构的数据，如图像识别、目标检测、图像分割等。在这些场景中，CNN能够有效地提取图像的特征，从而实现更好的性能。**而RNN 适用于处理时序关系的数据，广泛应用在自然语言处理、语音识别、机器翻译等领域。**在某些任务中，这两者也可以结合使用，形成更复杂的神经网络结构，**目前 CNN 、RNN 不断演进成熟，以“小模型”架构被广泛应用。**

中国人工智能产业技术环境（2/2）

Scaling Law是否失效？思维链、强化学习、后训练可提升模型训练ROI

自大模型发布以来，Scaling Law成为模型层发展迭代共识，国内大模型基座厂商均通过不断加大参数量级以获得模型能力的优化增强。2024年，随着大模型的训练脚步变缓，人们也开始关注讨论Scaling Law是否存在失效风险。而以国内外头部厂商的技术动态为标杆，我们可以看到大模型的参数规模与数据跨度仍有提升空间，且在多模态能力融合上完成持续突破。但受限于高质量数据、训练资源（算力、电力等）的可获取性及模型资源投入的ROI评估，一些大模型厂商已减缓或停滞了对新一代超大模型的训练投入，此外也在尝试多途技术路径提升大模型能力，如后训练的思维链优化，将Scaling方法由预训练转移到了强化学习推理优化阶段，为大模型能力扩展提供新道路，也对未来大模型的训推参数部署、AI推理算力需求等潜在方案布局带来新变数。

AI技术动态

Scaling Law 演进：Scaling未到尽头，各家仍在积极探索，探索大模型能力边界

- 大模型Scaling Law表示，**增加计算量、模型参数量或数据大小都可能会提升模型性能，但是提升效果会随着这些因素的增加而递减。**虽然Scaling Law原理给大模型能力演进限制了阈值空间，但仍有头部厂商在加大模型参数、数据规模和算力资源的投入，延续大力出奇迹的大模型训练之路。2025年2月，OpenAI推出GPT 4.5系列模型，进一步加大模型参数，主要通过无监督训练提升了模型通用能力，在模型准确率及幻觉率方面达成显著优化。但Altman同样表示这将是最后一代“非思维链”模型，后面GPT 5将采取融合技术路径，纳入推理侧思考。

思维链 CoT 优化：强化学习完成推理侧优化，在复杂计算、科学研究等方向持续加强

- **2024年OpenAI发布GPT o系列，**通过大规模强化学习算法让模型在数据高效训练过程中学会更好应用内部思维链（CoT，Chain of Thoughts），在解决复杂问题时表现更加出色，但彼时尚未公开技术细节。而**2025年初，DeepSeek开源R1系列推理思考模型，将思维链过程开放公开，极大推动全产业推理思考模型的技术进步，**也让人们对AI能力有了更深刻感知。

系统一
直觉和本能
95%

- 快思考：快速、自动、直觉性、无意识
- 原本GPT系列思考形态更类似于系统一

5%
系统二
理性

- 慢思考：缓慢、需要努力、逻辑性、有意识
- 推理模型加强推理思考能力，思考形态往系统二倾斜

跨模态响应：将大语言模型、视觉理解模型及和视觉生成模型等能力实现高阶融合

- 以OpenAI为例，在GPT 4o模型中将视觉理解模型GPT4V、视觉生成模型Sora、声音模型Whisper等模型模态融合，**通过GPT 4o模型在文本、语音、图像等多维度实现高效交互，可理解视觉、听觉和文本模态，并直接输出音频**，支持灵活的双工交互。未来在直接视频分析及交互领域是跨模态、多模态领域新的突破方向。

02/ 中国人工智能产业价值总览

—— 大模型对AI产业链带来哪些影响？

中国人工智能产业图谱


2024年中国人工智能产业图谱
人工智能应用层
Application for AI
AI+泛安防
计算机视觉 大数据智能
AI+金融
营销客服 信贷风控 视觉产品
AI+政务
便民办公 政务大数据
刑事侦查
AI+零售
运营优化 营销客服 视觉产品
AI+医疗
影像诊断 大数据决策
智慧病案与DRGs AIDD
AI+工业
视觉检测+安全生产 运维决策
AI+交通
智慧管理 自动驾驶
AI+对话交互
智能搜索问答 情感陪伴 企业服务
AI+效率工具
图像/视频处理生成 效率生产工具
AI+编程
AI+传媒影视
内容生成 剪辑特效
AI换脸换声 创意营销
AI+游戏
内容生成、场景建模、策略生成、AI Agent
AI+教育
教育工具 智慧校园
AIoT
产业级 消费级
人工智能技术层
Technology for AI
AI开放平台
AI开发平台
机器学习
知识图谱
自然语言处理
计算机视觉
智能语音
人工智能大模型层与工具层
Models for AI
通用基础大模型
按模型模态
按模型路径
大语言模型
闭源
开源
视觉大模型
语音大模型
多模态大模型
垂直行业/领域大模型
大模型开放平台
智能体开发平台
AI开源社区
模型平台/模型服务
工具层
人工智能基础层
Infrastructure for AI
算力基础
智算中心
企业自建智算中心 城市智算中心
智算软件平台
智能服务器
AI芯片
智能云服务
数据基础
向量数据库
数据集
公共开源 企业私有 高校 政府
AI基础数据服务
数据治理
算法基础
AI模型架构
CNN RNN Transformer
Diffusion Model ……
AI算法框架
下游
中游
上游

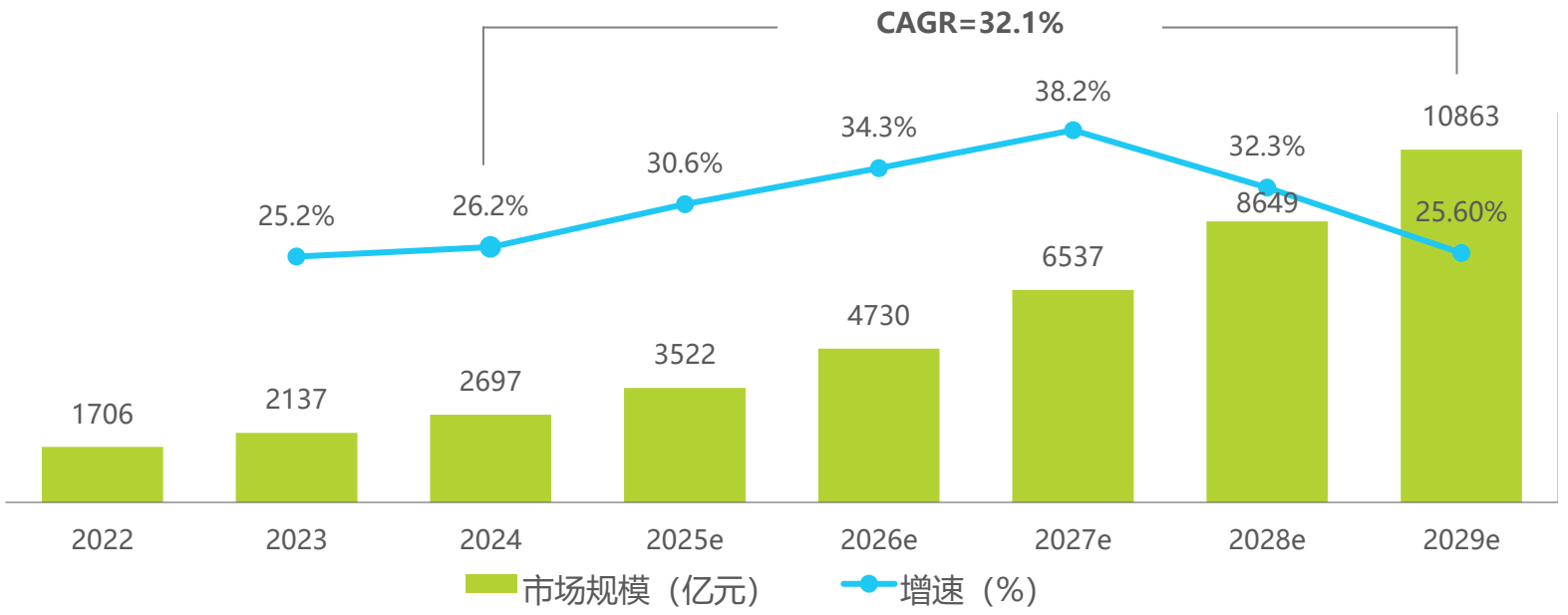
中国人工智能产业规模

2029年中国人工智能产业规模破万亿，未来五年复合增长率32.1%

2024年，大模型驱动的人工智能市场增长低于预期，主要因为大模型在真实业务场景中的表现未能完全满足部分客户的需求，且建设成本较高，企业需在数据基础、算力资源及组织机制等方面投入较多资源，导致多数项目处于尝试探索阶段，难以规模化落地；另一方面，模型计算成本下降叠加供应商间激烈竞争，模型调用的费用持续降低，进一步限制了市场增长。

2025年初，DeepSeek刷新了市场对大模型现阶段性能的认知，其开源策略结合高效、低成本的架构显著加速了中国AI产业向更加高效、开放和自主的方向迈进。各地政府、央国企等机构积极投入，将大模型与自身职能及业务场景深度融合，为2025年中国AI市场的快速增长奠定了基础。与此同时，大模型在推理、多模态等领域的性能持续增强，叠加MCP等智能体开发协议及工具的成熟，使得可自主完成复杂任务的智能体产品的实用性提升，夯实了AI市场增长的潜力；依托大模型的生成式AI产品也推动着传统小模型的落地应用，基于大小模型协同，为客户构建兼具效果与成本优势的理想方案。总结以上分析，艾瑞测算2025至2029年中国AI产业将保持32.1%的年均复合增长率，在2029年突破1万亿的市场规模。

2022-2029年中国人工智能产业规模



大模型对算力产业位置影响分析

“堆算力”不是训练侧的唯一真理位，推理侧算力需求将大幅上涨

近年来中国持续投资智算产业资源，2024年，部分智算中心出现了闲置情况。AI算力是否投资过剩的问题被人们所关注与提出。艾瑞认为，2024年的部分算力闲置现象主要由于当地产业规划错配、供给侧前置布局及需求侧训练需求有所缓解等原因导致，随着模型技术迭代及推理应用爆发，长久来看高性能算力仍处于高需求状态。智算中心建设需协调好地区产业资源规划，从软硬件角度优化算力利用率及平台运行质效，稳健支撑上层AI产业发展。2025年初，随着DeepSeek V3及R1模型的开源及产品破圈，英伟达股价单日下跌幅达到近20%。在美国对中国实施芯片出口管制的背景下，DeepSeek模型通过算法优化，如结构化稀疏注意力、混合专家系统等技术，显著降低了模型训练成本，极大降低对进口高端GPU训练卡的依赖程度，在一定程度上削弱了美国通过芯片出口管制遏制中国AI发展的国际战略。中国基于算法优化与技术创新，进一步突破了模型层性能表现，极大增强了中国在人工智能产业的自主可控能力。而“服务器繁忙，请稍后再试”的AI回复，也象征着模型应用端的需求爆发。未来，如DeepSeek-R1等优质的开源模型及低调用成本将刺激推理算力需求的大幅增长，中国智算中心的利用率也将有望逐步提高。

中国人工智能算力产业发展

2023年-大模型训练任务激增，算力出现短缺

- “百模大战”，训练算力需求激增
- 美国商务部“出口管制条例”限制算力供应，出现囤卡抢卡现象

2024年-短缺问题缓解，甚至部分地区出现过剩情况

- 多方产业提前布局算力储备，部分企业囤积的算力增多，模型训练需求放缓，出现结构性过剩情况

2025年-优质开源模型带动推理落地，算力需求再度攀升

- DeepSeek 系列模型的发布加速了AI渗透扩散，推动大模型普及与应用落地，推理侧算力需求大幅增张

大模型带动基础层工具产品售卖

分布式开发框架、LLMOps平台、一体机等基础层产品热度渐起

在当前人工智能领域，模型参数规模不断扩张，大模型的分布式训练因此变得愈发普遍。在此背景下，算法框架层面涌现出诸如DeepSpeed、Megatron、Colossal-AI 等分布式 AI 开发框架。这些框架基于PyTorch框架生态，提供了深度学习优化库，致力于提升大模型分布式训练的训练与推理效率，助力开发者更高质量、更高效地完成大模型的训练及部署工作。从平台的角度来看，在大模型时代，AI 开发平台也在积极探索与升级。与传统AI模型相比，大模型在开发、应用及部署上对算力支持、数据管理、功能模块及工具库等方面均提出更多要求，MLOps分化出LLMOps，出现面向大模型，提供整个模型生命周期中加速 AI 模型开发、部署和管理的专业平台工具。为了顺应市场热点以及客户需求，各大厂商纷纷推出了各自的一体机产品。一体机作为软硬件集成的大模型实践解决方案，具有显著的优势。它能够降低企业应用大模型的技术门槛，加速大模型在各个行业的落地实施，同时为企业提供安全、高效的 AI 应用开发和部署能力。以DeepSeek为代表的模型，具备开源部署、本地化应用（保障数据隐私）、低成本高质量以及快速定制化交付等优点，精准地满足了政府、金融、医疗以及工业制造等B端行业的特定需求。预计2025年，DeepSeek适配一体机市场将进一步升温，迎来新的市场热潮。随着大模型商业化进程的不断加快，一体机、分布式 AI 开发框架以及LLMOps平台等基础层工具逐渐进入产业视野，成为支撑企业及开发者完成产业端大模型应用建设的重要力量。

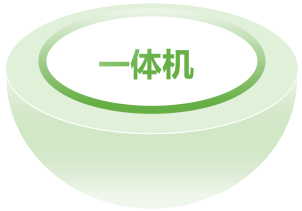
中国大模型产业链带动工具

面向分布式训练的AI框架

- 大模型时代下，分布式训练对面向大模型时代分布式训推的软件栈提出新型框架要求，由此诞生以DeepSpeed、Megatron、Colossal-AI为代表的分布式AI开发框架，提高大规模模型训练的效率和可扩展性，并有效降低训练成本。

大模型AI开发平台

- LLMOps是面向大模型，提供整个模型生命周期中加速 AI 模型开发、部署和管理的专业平台工具。
- 针对大模型的AI开发平台可更好的助力企业、开发者完成大模型应用的开发、构建及部署。



集结硬件算力与软件平台的产品

- 一方面搭载高性能硬件，提供可灵活配置的底层算力，一方面可提供大模型训推平台，或内置开箱即用的大模型场景应用。
- 更多被工业、制造、医药等私有化部署要求高、低应用门槛的需求企业青睐，DeepSeek等开源模型带动一体机销售。

模型层开源创新推动上层商业化实践

降本增效推动大模型落地，选择微调、蒸馏或RAG等路径达到ROI最大化

2024年，“后训练”和“强化学习”成为大模型技术创新的热点。后训练通常由大模型厂商在预训练模型基础上完成，其流程一般包括：监督微调（SFT），即利用特定任务的标注数据对模型进行微调，使其学习任务模式；奖励模型（RM）训练，通过收集人类反馈数据训练奖励模型，评估输出质量；以及强化学习（RL），利用奖励模型反馈优化模型，最终生成更符合人类偏好的输出等。由于代码、数学等领域更适配模型评估与奖励反馈环节，推理模型在这些领域的深度思考能力更强，而在文学、医药、科研等领域，因存在大量实验数据和非唯一最优解等影响，后训练的效果提升相对有限。从落地质效来看，DeepSeek通过创新的模型结构和训练任务优化，如多令牌预测（MTP）、多头潜在注意力机制（MLA）、GRPO（分组相对策略优化）等，在保持高性能的同时，大幅降低了训练和推理成本。这些低成本、高性能的开源模型（如DeepSeek、阿里QwQ系列）极大推动了大模型的商业化实践，吸引更多需求方拥抱大模型能力底座，并进一步采用微调、蒸馏、RAG工程等方式完成定向优化和应用部署。

预训练大模型能力落地实践路径

应用厂商侧“跑马圈地”态势渐起

价格与流量成为应用层核心竞争策略，大模型实践更加定制化及产品化

2024年，大模型能力变现及商业化进程进入关键期，应用层的产品表现成为兵家必争之地。在成本优化与市占竞争的双重驱力下，国内各家大模型厂商纷纷降价，试图通过价格战构建B端竞争策略。2024年5月15日，字节跳动将其大模型的计价单位从分降至厘，声称价格比同行低99%。同月5月21日，阿里云宣布通义千问最高降价97%，百度宣布两款主力大模型免费。在C端，大模型产品也出现大量买量投流的资金竞争策略。根据有关媒体公开信息报道，截至 10 月 29 日，kimi 智能助手、字节跳动豆包、腾讯元宝等所有 AI 应用 10 月全网广告投放（投流）支出超过 3 亿元人民币。由此可见，无论是B端还是C端，大模型厂商“跑马圈地”态势均渐起。从实践落地角度看，大模型落地应用更加定制化及产品化，尤其面向ToB客户，更加开放底层模型能力与定制化程度，为客户提供Post-pretrain、SFT精调、RLHF等成熟丰富的微调方案，将大模型解决方案深度嵌入企业需求与业务流程。

中国应用层厂商市场策略

低价策略/内外动机

低价动机 – 内部：

- AI芯片技术突破，**单位算力对应的成本在下降**
- **MoE架构节约推理成本**，大模型的量化压缩技术越来越先进

低价动机 – 外部：

- 基于低价策略吸引更多B端与C端客源，**尽可能在早期占据市场份额，以此获得业务正循环**，在市占基础上拓展更多业务线及利润点。
- 在吸引客源基础上，**培育更良性的开发者生态，并在产品侧获得更多相关用户数据，**如用户偏好、行为趋势等，更好优化模型技术与应用产品。

¥ 价格　　产品

定制化/产品化

定制化： Post-pretrain | SFT精调 | RLHF | Prompt (Zero-shot/Few-shot) | LoRA

- 面对落地实践需求，尤其是To B产业端需要模型能力与业务需求的深层适配，**大模型往往会通过微调、定制化策略提供产品服务，国内各家大模型厂商的平台方案支持多类微调方案**，Open AI也于8月份开通了最强大模型GPT 4o的微调功能。

产品化：平台/Agent/APP

AI办公平台	AI编程助手	AI智能对话
AI法律助手	AI Agent助理	AI数字人

大模型能力成熟　　*厂商尝试跑马圈地*

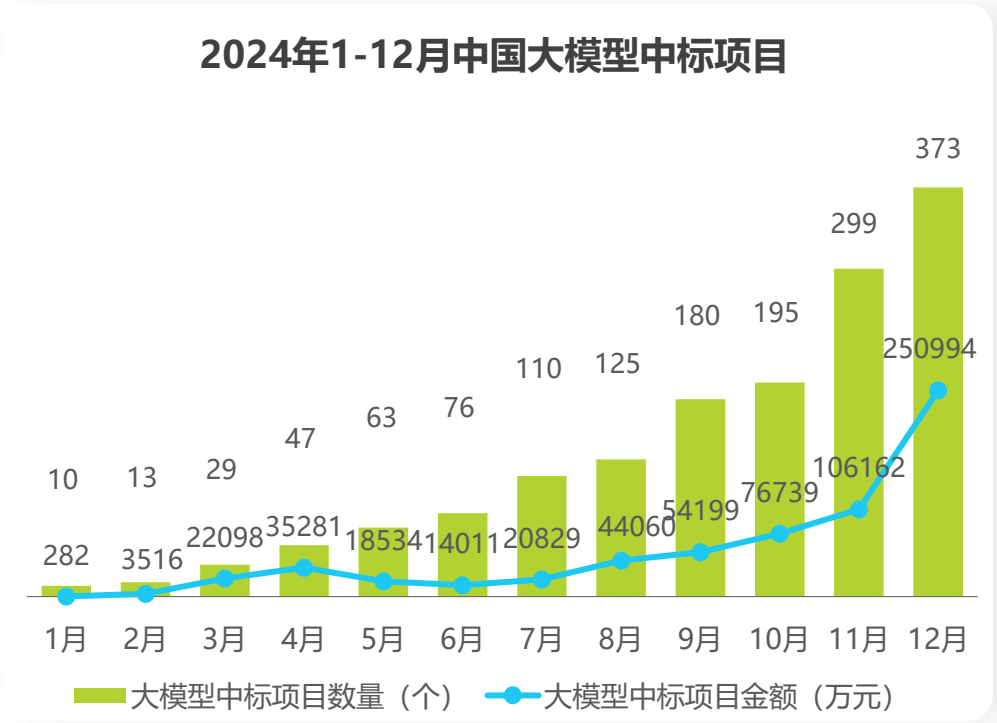
大模型落地产生高频客制化需求

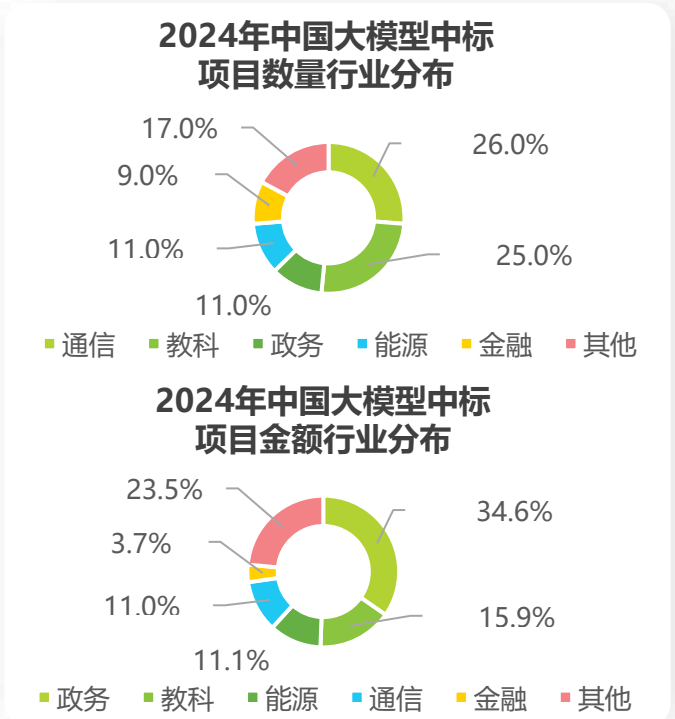
大模型训练/推理成本降低　　*商业化步伐加快*

B端大模型商业化进程

以央国企为需求主力，率先落地在政务、教科、通信、能源等领域

预训练大模型、类Sora模型以及类o1模型的研发训练需要大量的资源投入，大模型厂商面临资源投入与营收压力需探索有效的变现途径。从短期来看，B端项目制落地仍然是大模型基座能力变现的主要方式。 2024年8月，月之暗面发布企业级API加码B端，11月，零一万物推出面向零售、餐饮行业的数字人解决方案。招投标数据显示，2024年中国大模型项目数量与金额呈现快速增长趋势，率先落地在政务、教科、通信与能源等领域。在供给侧，科大讯飞、百度、智谱、火山引擎、阿里云和腾讯云等成为主力中标厂商。2025年初，DeepSeek V3与R1的开源模型发布，进一步掀起B端产业生态的合作浪潮，以沐曦、天数智芯为代表的基础层、以阿里云、华为云为代表的模型平台层、以钉钉、吉利汽车为代表的应用层，纷纷接入DeepSeek生态，借助优质开源模型能力，推动大模型能力在垂直领域的实践落地。





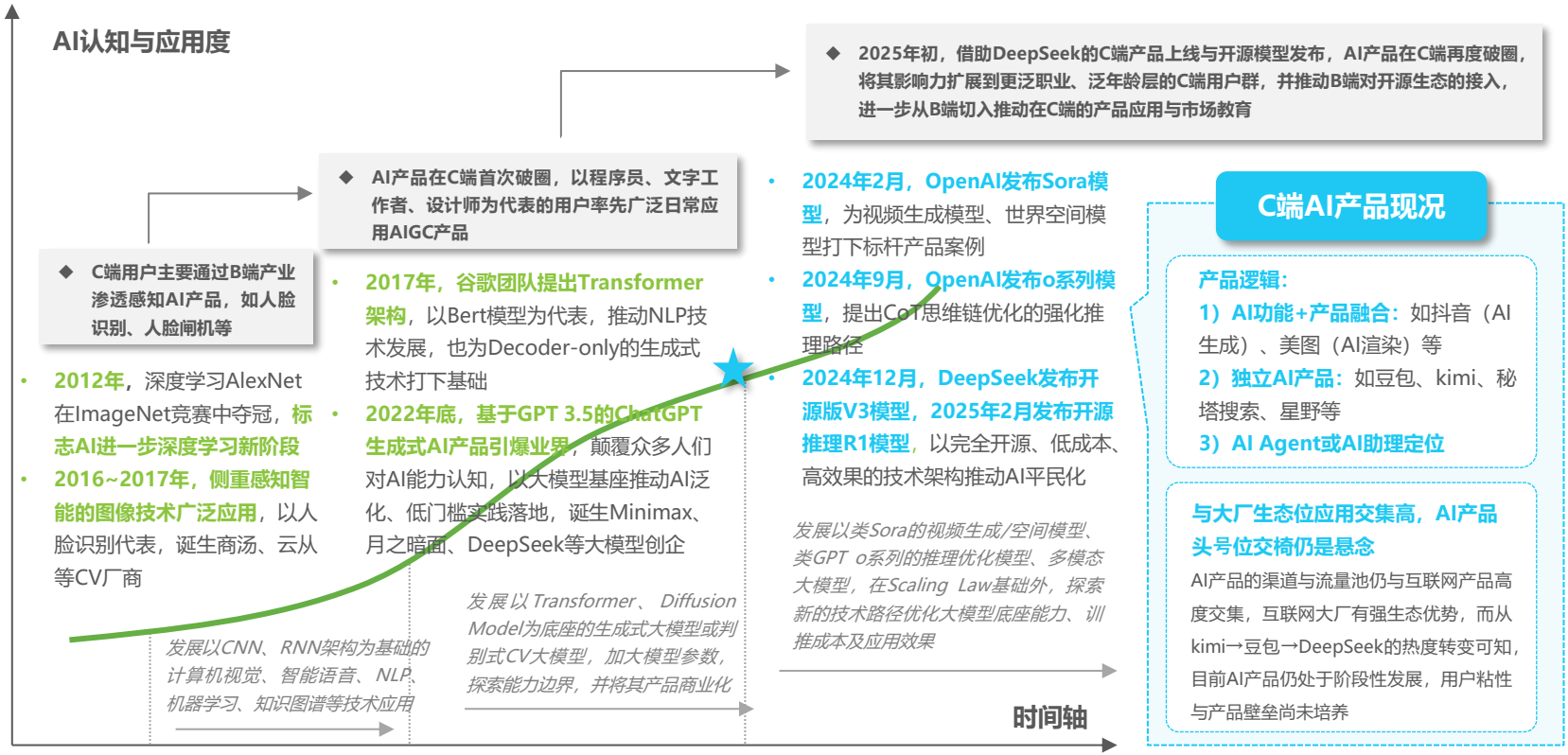
来源：智能超参数，艾瑞咨询研究院自主研究绘制。

C端AI产品生态位分析

AI阶段性产品壁垒仍然非常低，终极产品形态及生态优势尚未形成

2024年，中国AI产品在C端发展迅速，产品类型涵盖内容创作、智能对话、情感陪伴、效率工具及音视频生成等，应用场景广泛。从商业模式来看，中国AI产品在C端产品主要采取“免费+订阅制”的商业模式，视觉模态类产品的商业步伐会稍快于语言类产品，如剪映、美图等产品的会员制AI功能，或无界AI、触手AI等产品的会员制订阅及资源购买。整体来看，大多数C端 AI产品仍然面临用户黏性不足，收费持续性不足的问题，产品形态尚未稳定，生态壁垒尚未建立。相较于互联网较为稳定的生态格局，AI产品的头号位交椅仍是悬念。

近十年中国AI发展阶段及关键性节点



深层产品洞察：AI Coding

强化推理提升可靠coding能力，coding for everyone将改变什么？

AI Coding产品是指利用人工智能和机器学习技术，通过理解人类语言描述来自动生成代码的工具，具备提升编码效率、减少人为错误及简化开发流程等产品优势。2024年6月20日，Anthropic 公司发布 Claude-3.5-Sonnet 模型，该模型在编程能力上取得重大突破，显著推动了业界 Cursor、Devin、Windsurf 等标杆产品的破圈应用。AI Coding 产品的发展可极大地提升专业开发者的编程效率，使其能够将重复性工作交给 AI 处理，把更多精力投入到创造性工作中。此外，随着 Sonnet、GPT o 系列等模型能力的不断提升，AI Coding产品的属性正逐步从辅助性 Copilot 向自主性 Agent 演进。这种演进不仅有望进一步降低编程的门槛，使更多泛开发者、非专业人士能够进入编程行业，推动编程的民主化，为后续的软件开发、产品交互及流量生态带来新的发展可能性。

AI Coding产品方向分析

AI Coding产品

AI Coding for developers

代码审查 | 代码补全 | 调试优化 | 代码生成 | ……

- **技术要求**：需要开发者具备一定的编程基础和技术背景，能够理解生成的代码逻辑，并对代码进行修改和优化。

早期Copilot
以代码补全功能为主

思维链推理优化

Claude Sonnet 3.5等模型加持

Cursor、Devin、Windsurf、Bolt:
对话式编程，侧重代码理解、生成、重构

AI Coding for everyone

图形化界面 | 拖拽组件 | 自然语言描述 | ……

- **技术要求**：用户无需具备专业的编程知识，只需具备基本的计算机操作能力和逻辑思维能力

低代码工具

受限于低代码工具技术路线，目前尚未实现从需求到软件开发应用的端到端实现

思维链推理优化

Claude Sonnet 3.5 等模型加持

新一代开发工具

从自然指令需求到软件开发交付的端到端实现，更多非专业开发者参与到软件开发中，产品生态变得更加多元化且个性化

产品技术方向

- 上下文能力提升
- 复杂代码逻辑、端到端交付能力，0-1门槛跨越
- 产品形态探索，基于Coding以上的PMF

可参考视频剪辑工具演进路线

由专业性要求高的PS剪辑软件 → 拖拽界面、用户友好的剪辑软件

剪辑软件的门槛降低催生大批自媒体工作者，对视频软件生态、用户流量分布带来重构式影响

深层产品洞察：AI Agent

弥合大模型能力与场景应用的鸿沟，多元化厂商生态驱动产品应用创新

AI Agent是一种能够自主感知环境、作出决策并执行行动的智能体产品。受益于强化学习、后训练等技术突破，大模型已展现出优秀的逻辑推理及规划能力，然而，其与应用需求侧之间仍存在一定 “Gap”。Agent 作为“桥梁” 角色，可支撑大模型落地到各类具体应用之中，补足其精准对接业务需求、上下文记忆、主动规划执行以及多任务协作等多方面能力。

当前，Agent市场呈现出厂商生态分化的态势。互联网科技巨头与垂直领域科技厂商常借助Agent能力赋能原有产品，提升其使用体验与智能化水平。同时，各类厂商依据自身优势，针对特定客户群体，推出不同模式的Agent产品，如一站式Agent服务、Agent搭建平台及Agent应用：2024年10月，智谱华章推出自主智能体AutoGLM，覆盖手机、浏览器、电脑等不同场景，可理解超长指令，执行超长任务；2025年3月，蝴蝶效应推出通用型AI助手Manus，可实现企业研究、旅行规划、课程设计等多场景的任务规划与流程操作。

中国AI Agent厂商生态分化现状与当下发展困境

厂商生态分化逻辑：核心能力与市场需求的双向适配

互联网科技大厂 B端与C端

依托底层算力资源与大模型能力，构建"基础设施+平台框架+应用生态"体系

- 提供一站式大模型开发平台
- 推出通用型Agent工具链
- 成熟产品集成Agent能力

垂直领域科技厂商 B端

深耕行业know-how，推动Agent技术场景化落地

- 聚焦特定领域
- 结合原有产品进行智能升级
- 构建领域专属工具链

Agent开发平台厂商 B端

专注Agent技术栈创新，提供跨行业解决方案

- 提供低代码Agent构建平台
- 支持多模型适配
- 提供全生命周期开发管理

原生Agent应用厂商 B端与C端

以终端用户需求为导向，打造场景化智能服务

- C端产品侧重交互体验
- B端产品强调任务闭环

当下AI Agent发展困境

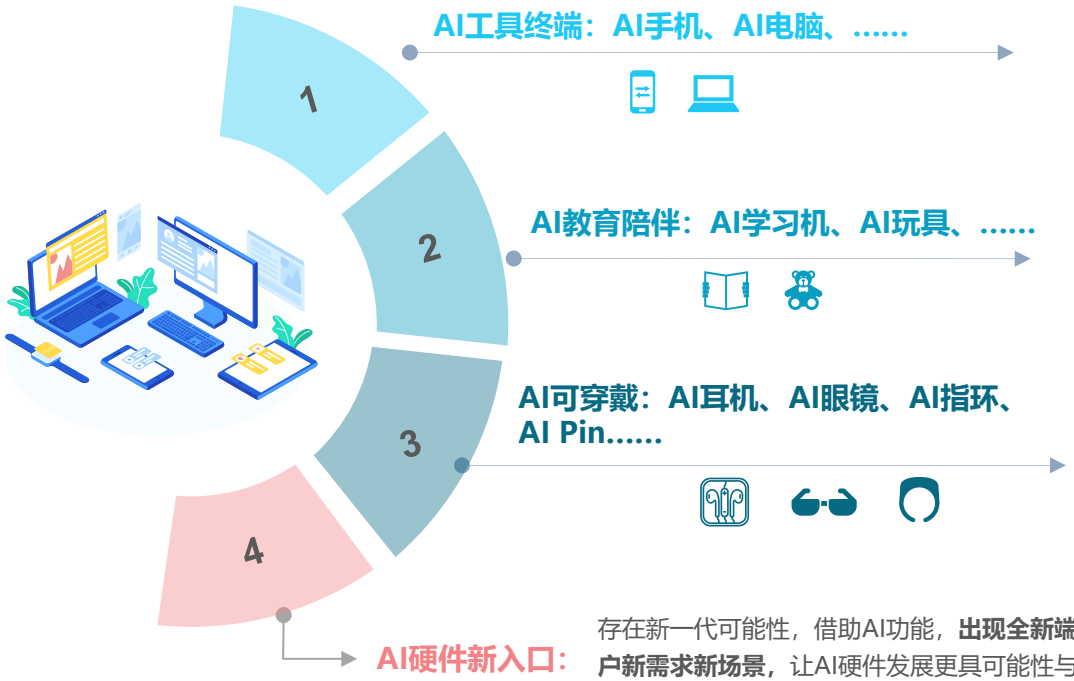
- 侧重垂直领域优化，但仍**囿于对话式沟通，与对话式AI产品逻辑类似**，Agent的强势能力尚未凸显
- 产品同质化严重，**主动性、记忆能力、执行感知能力有待提升**
- 智能体**与生产生活联系紧密度低，**多智能体协作生态尚未培育

深层产品洞察：AI硬件

AI能力输入到终端硬件，端侧AI寻求规模效应突破，承载更多流量入口

2024年，AI成为手机与电脑的主力卖点，如华为、荣耀、小米、VIVO、OPPO等国产手机纷纷打造手机端侧大模型，在AI消除、AI搜索对话、生活助手等功能已提供较好用户落地体验，10月，荣耀Magic7系列携带全新的YOYO智能体发布，可在手机侧实现一句话点咖啡、一句话取消订阅等功能。字节跳动豆包也在10月推出首款AI智能体耳机Ola Friend，接入豆包大模型进军AI硬件。以科大讯飞为代表的AI学习机产品在大模型技术加持下不断更新迭代，在教育产品内预装大模型能力，精准提升教学能力同时配备AI助手提升引导交互能力。除现有硬件对AI能力的融合，也有大厂及创业厂商探索形态交互更加新颖的新一代硬件设备，如Rabbit R1、AI Pin等新产品，但由于技术不成熟、场景未规模化渗透等原因，尚未出现破圈效应。

中国AI+硬件市场洞察



强功能性角色，也是AI能力注入终端的率先变革力。2024年，手机、电脑等终端厂商已纷纷发力，**基于端侧大模型让AI能力与其更好融合，从软硬生态交互角度完成更好人机协同**，也借此获取更多市占及增长驱力。

语音能力为教育陪伴类产品提供**更好交互入口**，而生成式AI带来更多语义能力增强，各种教育陪伴类硬件的语音助手升级为**智慧助手，基于语音交互的教育、陪伴及亲子类内容同样得以优化，带来产品升级及创业机会。**

结合**生成式AI、大模型技术与可穿戴设备，完成用户数据的实时收集与个性化服务**，基于端侧硬件性能，更多通过云端形式实现，与其他端侧设备达成协同优化。

存在新一代可能性，借助AI功能，**出现全新端侧个人设备，或承接现有端侧设备功能，或打开用户新需求新场景，**让AI硬件发展更具可能性与想象力。

03/ 中国人工智能产业商业进程

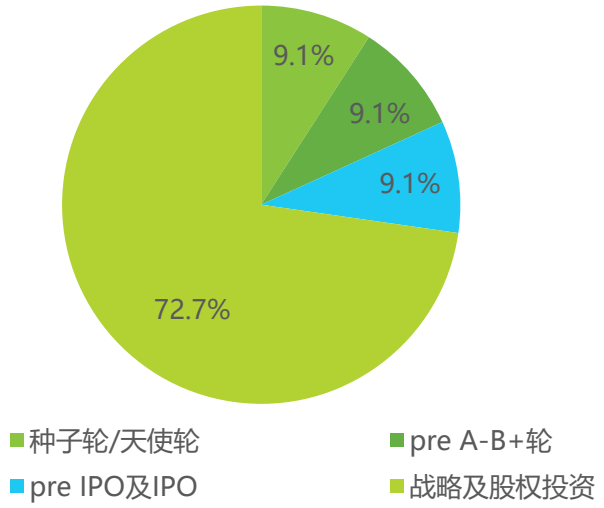
—— 大小模型技术演进下，AI产品表现如何？

语音模态

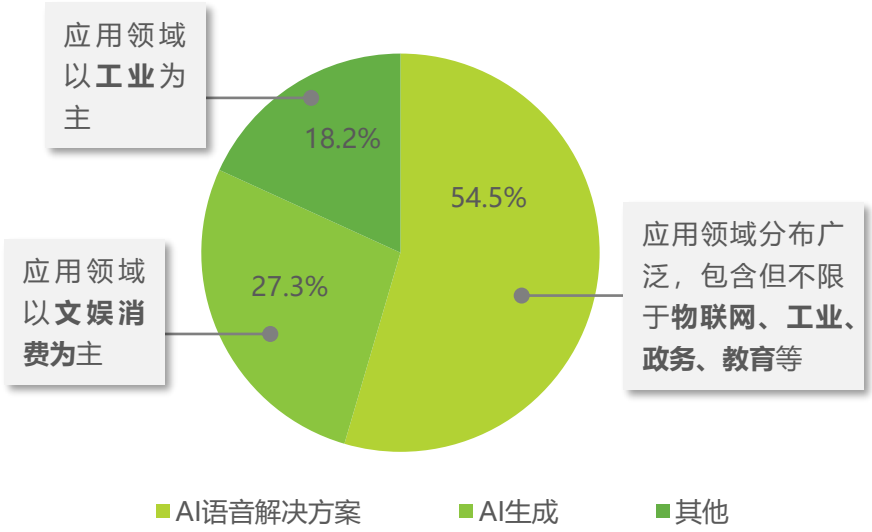
产品形态以AI语音解决方案和AI生成为主，前者应用广泛，后者聚焦文娱

在2024年语音模态赛道的投融资中，超过70%为战略及股权投资。该赛道的产品形态主要分为两大类：AI语音解决方案和AI生成。其中，AI语音解决方案的应用范围极为广泛，涵盖了物联网、工业、政务、教育等多个领域；而AI生成类产品则相对集中于文娱消费领域，如AI音乐生成。

2024年语音模态赛道投融资轮次占比分布



2024年语音模态赛道产品形态占比分布



语音模态

语音识别与生成能力持续增强，重点关注端到端的语音大模型技术架构

在语音识别能力方面，ASR模型数据量和参数量逐步增大，且大模型可为ASR提供上下文内容理解，在识别率、说话人分离、多方言覆盖等方面继续进行有效提升；在语音合成能力上，基于AI语音设计、AI音乐创作等技术方向，音频能力正由被动生成发展到主动创作；在语音交互能力上，更多厂商在语音大模型架构中，可由ASR-LLM-TTS的级联式架构升级为端到端的语音交互大模型，显著提升人机语音交互的响应速度、流畅度、打断性等。2024年5月，OpenAI发布端到端语音架构的GPT 4o系列，在语音交互能力表现优异，8月，科大讯飞更新星火语音大模型，采用端到端语音架构，在响应速度上有了显著提升，使得对话更加自然流畅。而与大小模型应用融合逻辑类似，目前两类语音交互模型架构仍各有优势，级联式架构具备可控性与准确性等优点，端到端语音大模型虽然提升交互效果，但会带来幻觉问题，因此落地实践上仍需根据应用场景选择合适的语音模型架构。

AI语音演进方向

1 语音交互能力

第一代语音大模型：级联式架构

语音识别（ASR）+ 大语言模型（LLM）+ 语音合成（TTS）

- **模型优势**：高准确性与可控性
- **模型劣势**：高延迟

第二代语音大模型：端到端的语音交互模型

语音分词器 + 大语言模型 + 语音合成器

*将连续的音频信号编码为离散的**token*** *对**token**展开自回归建模* *将生成的**token**合成为语音波形*

- **模型优势**：低延迟、泛化能力
- **模型劣势**：可控性与准确性，若与RAG结合则对低延迟优势有损耗

2 AI语音克隆 & 声音设计/转换

由文本生成声音 | AI声音克隆：音色复制

AI声音设计/转换：在语言/声音基础上根据参数设置、Prompt生成对应音色

- **海外代表厂商 – ElevenLabs：**AI音频模型公司，2022年成立，目前可以生成 32 种语言的逼真、多功能且具有上下文感知能力的语音、声音和音效，2024年10月，ElevenLabs发布全新 **AI 语音生成工具 Voice Design，通过简单的文本描述即可创建个性化语音。**
- ElevenLabs在X to voice等项目部分开源。**市面在AI声音克隆、AI音色生成上已有众多优质开源项目**，如ChatTTS、CloneVoice、GPT-SoVITS、Seed-TTS等
- 该类TTS技术成熟度相对较高，已在国内**电商领域（视频制作）、泛娱乐消费领域（有声书、AI配音、歌手音色克隆）**等应用端得到小规模传播应用。

3 AI音频/音乐创作

- **海外代表厂商Suno AI：**AI音乐创作公司，2022年成立，基于Chirp模型为用户生成逼真的音乐和声音效果，目前模型已更新到V4版本，
- 2024年，**Stability AI在音乐生成领域继续开源Stable Audio 2.0和Open等系列**，可生成音频样本、音效、制作素材和歌曲等。
- 2024年11月，**NVIDIA 推出全新生成 AI 音频模型 Fugatto**，能够结合文本和音频输入，生成多种类型的音乐、声音及语音。

语音模态

语音生成类产品涌现更多创新机会，以海外创企为代表进行市场试水

传统市面AI语音产品可分为两类，在过往AI产业浪潮中走在产品化与商业化前列，一类是AI语音转写产品，应用在办公、翻译等领域，一类是AI语音交互产品，以语音机器人为代表，应用在办公、客服、营销等领域。而近两年，以Elevenlabs、Suno AI等为代表的语音生成厂商在TTS技术、扩散模型的创新融合，让2024年海内外生成式语音产品市场出现变革性进步，AI语音产品在原本语音生成的功能属性上，在音色度、内容性、创作性上添加更多生成式的变量元素，AI配音、AI音色克隆、AI有声书阅读、AI音乐生成等产品商业化步伐提速明显。

2024年AI语音典型产品盘点

1 AI语音转写产品

基于语音ASR（Automatic Speech Recognition）能力提供**转写、翻译、总结、提炼等功能**，相较于AI语音交互产品，该类助手及平台产品更强调**对语音进行转化分析的工具属性。**

- **讯飞听见：**提供音视频、录音等转写服务。
- **通义听悟：**记录、转写及分析音视频内容。

2 AI语音交互产品

在对话式AI产品中，在**人机对话系统中提供语音功能**，完成语音式对话交互，常应用于**问答、客服、营销、教育等场景**

- **Realtime API:** OpenAI的语音交互调用服务。
- **豆包：**字节的AI对话APP。
- **智能语音/对话机器人：**百度、阿里、科大讯飞、百融云创等B端产品。

3 AI语音社交陪伴产品

以语音为核心，提供社交陪伴。1）以**语音为输入输出形式**的内容社交平台；2）以**语音为核心交互**的陪伴产品

- **Airchat：**美国创企，以语音为主要内容形态进行异步社交。
- **AI陪伴类产品：**在虚拟角色中添加个性化音色的语音交互，提升情绪价值，如星野、WoW、猫箱等APP

*大模型提升语音产品效果；**主流市场已**被语音技术老牌厂商、国内大厂**占据**；**社交陪伴类有创业机会***

4 AI语音内容生成产品

基于TTS（Text-to-Speech）技术将文本转换为自然逼真的语音。1）着重语音生成的AI语音产品，如**音色克隆、音色设计生成**，主要应用在**AI配音、短视频配音、数字人配音、有声书阅读等**领域；2）着重**内容+语音生成的AI语音产品**：对文本、音视频等信息按需提炼，以语音形式产出，目前在**播客、新闻等领域**有所尝试

- **Elevenlabs：**美国创企，提供基于多语言的多类型声音、风格的语音生成产品。
- **魔音工坊：**出门问问旗下产品，提供文字转语音的生成产品。
- **剪映：**字节跳动的视频剪辑平台，提供语音生成功能模块。
- **NotebookLM：**谷歌推出的AI学习产品，将上传的文本、PDF及音频等各种格式的文档转化为生动的音频播客。
- **PocketPod：**根据客户需求生成音频内容，帮助用户获取新闻、信息等播客内容。
- **Playnote：**美国创企PlayAI的新品，将PDF、文本或视频转换为故事、播客或简报

6 AI音乐生成产品

根据用户的文本提示、旋律提示或音乐元素，**快速生成原创或相似风格的音乐作品。**

- **Suno AI：**美国创企，AI音乐生成软件
- **海绵音乐：**字节跳动推出的AI音乐创作工具，Seed-musix是字节推出的音乐生成大模型
- **SkyMusic：**昆仑万维推出的音乐生成平台产品

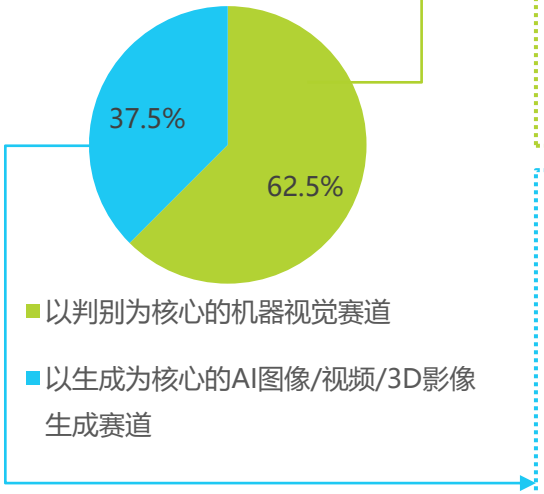
*在**定制化音频内容生成、音乐生成**有产品创新机遇；相较于AI语音转写、AI语音交互**有更多创业机会***

视觉模态

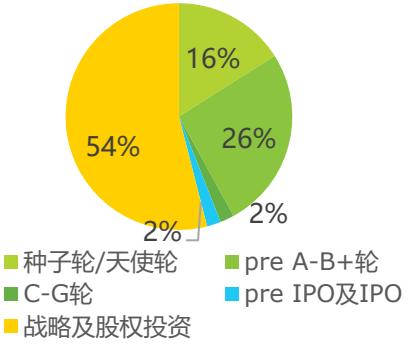
集中于机器视觉与生成类赛道，前者商业化更成熟，后者多为新兴企业

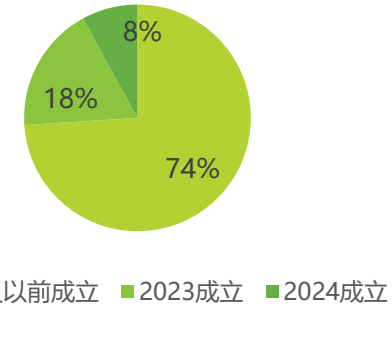
2024年视觉模态赛道的投融资中，超半数为C轮以前的前期融资，投融资方向主要集中于以判别为核心的机器视觉赛道和以生成为核心的AI图像/视频/3D影像生成赛道两大方向。从企业成立年限来看，以生成为核心的AI图像/视频/3D影像生成类企业中，70%的玩家成立于2023年及以后，多数是受AIGC浪潮影响之下成立的新兴企业，说明生成类企业正在成为视觉模态赛道中冉冉升起的投融资新星。但从投融资轮次来看，仅有的2例较为成熟轮次的投融资事件（1例新三板上市，1例D+轮次融资）均发生在以判别为核心的机器视觉赛道中，说明尽管生成类企业势头迅猛，但资本市场目前仍对机器视觉赛道的商业化成熟度抱有更高预期。

2024年视觉模态赛道投融资产品应用方向占比分布

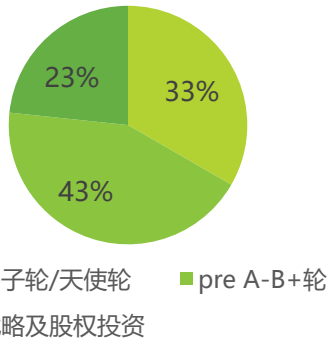


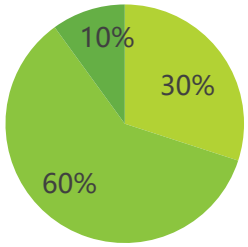
2024年机器视觉赛道投资轮次与企业成立年限占比分布





2024年生成类赛道投资轮次与企业成立年限占比分布





来源：烯牛数据。艾瑞咨询研究院自主研究绘制。

视觉模态

Transformer架构为技术主旋律，持续演进ViT与DiT两类技术路线

Transformer架构最早在2017年由Google研究团队Vaswani等人提出，而后在语言领域取得了革命性突破。借鉴语言领域的成功经验，Transformer架构同样可将图像分割为多个小块（patch），应用自注意力机制，从而大幅提升视觉CV大模型的泛化能力、理解能力及处理模糊复杂影像能力。而在生成领域，融入Transformer架构的DiT模型为扩散模型带来新思路，相较于U-Net传统卷积神经网络，DiT架构模型能够更好地处理图像的潜在表示，并捕捉图像的长距离依赖关系，以生成高质量的图像。2024年初，在Sora产品验证DiT路线的涌现能力之后，主流SD模型、Flux模型及国内视频生成模型多延续此技术架构，并在生成模型的真实性、可控性、可编辑性上持续发力，更大程度释放模型链接需求的技术生产力。

AI视觉演进方向

CV领域：图像分类、目标检测等

小模型
以CNN为代表

- 人脸识别、车辆识别
- 违停违放等行为检测
- 抽烟检测、安全帽检测
- 明厨亮灶检测等

小模型在这些领域仍具优势

1）政策、技术升级驱动
2）商业驱动：前端摄像头路数及图像体量足够多以分摊大模型成本，提升项目ROI

大模型
以ViT为代表
(Vision Transformer)

- 泛化能力增强，可处理更多图像类目，且标注工作变少
- 可识别分析图像质量不高、像素较少的影像，可处理复杂影像，具备进一步推理分析能力
- 如意图识别、微表情识别、人员追踪、流量统计分析等

图像/视频生成领域

Diffusion Transformer（DiT）架构将为主流，图像生成领域同样也有Scaling Law

- 以Stability AI为例，2024年6月Stable Diffusion 3正式开源，**相较于过去版本的U-net + Diffusion 模型架构，Stable Diffusion更新采用DiT架构**，提升图像生成可控性，并让模型消化更多的图像及视频数据，提升模型在生成领域的涌现能力。
- 目前主要以DiT为技术架构的模型有**Stable Diffusion模型、Sora模型、Flux模型**等

内容可控是生成领域核心关注方向之一

ControlNet
条件控制

对大扩散模型做微调，控制扩散生成走向

动作姿势	深度	边缘检测
柔和边缘	涂鸦乱画	……

- 与LoRA风格预设类似，ControlNet可以通过插件预设模型，**进一步精细控制人物姿势动作、面部表情、手部动作等，**“有条件方向”的创作让模型生产力大幅跃升。

视觉模态

产品定位清晰，以功能为出发点分别面向G端、大B、中小B、C端市场

艾瑞将主流视觉模态产品分为AI图像/视频分析产品、AI图像/视频生成编辑产品、AI视觉搜索问答产品三类。其中AI图像/视频分析产品定位为早期计算机视觉、CV判别产品，以商汤、云从、海康及华为等厂商为代表，率先进军大B端与G端市场，掀起第一波人工智能产业浪潮，在安防、金融、医疗及工业等领域开展落地。2024年，大模型Transformer与传统小模型CNN的架构融合为该类产品注入强心剂，在厂商格局保持稳定的情况下，重点在央国企市场实现项目拓展与架构升级。而AI图像/视频生成编辑产品、AI视觉搜索问答产品更多定位在生成式AI路径，在AIGC浪潮中涌入云厂商、AI厂商、业务厂商及创业厂商等诸多市场参与者，面向电商、零售、设计、游戏等行业，设计师、博主及C端等消费者开展产品试水升级与商业化实践。

2024年AI视觉典型产品盘点

1 AI图像/视频分析产品

AI图像视频分析产品是以机器视觉为基础，基于深度学习算法对图像中的对象进行识别分类、检测跟踪及定位分析等操作，提供强大视觉分析和处理能力的产品。

- 人脸识别
- 人证核验
- 车牌识别
- 物体识别
- 商品识别
- 缺陷检测
- 证件识别
- ……

硬件产品：AI相机、面板机、门控机、传感器等端侧设备

软件产品：与设备、数据融合的视觉算法平台产品等

- ✓ **代表厂商：**华为、商汤科技、云从科技、海康科技、创新奇智等，市场格局相对稳定。
- ✓ 作为AI视觉产品的先锋者，**AI图像视频分析产品率先落地于安防、金融、医疗及工业等领域，以To B与To G市场居多。**
- ✓ 2024年，大模型Transformer架构AI图像视频分析产品的**识别理解能力，视频结构化动态分析能力大幅提升**，此外更多融合大语言模型能力，实现多模态检索、交互及分析，以项目制落地于**智慧矿山、智慧园区、智慧交通、智慧城市等领域**，以央国企为代表，实现治理管理水平的升级建设。

2 AI图像/视频生成编辑产品

AI图像/视频生成编辑产品是指基于AI技术来生成、编辑和优化图像与视频内容的工具和平台，目前主要以剪辑软件与内容生成软件两类为主，面向中小B及C端市场。

- **Remini：**以粘土风再度出圈的AI图片编辑产品，主打AI高清、AI风格化、AI生图等。
- **剪映：**字节跳动的剪辑平台，融合AI技术提供AI特效、AI编辑、AI生成等功能。
- **美图：**美图旗下美图秀秀、WHEE、Wink、美图工作室等多产品。
- **Sora：**OpenAI的视频生成产品，于2025年初正式对公众开放。
- **可灵AI：**快手视频生成类产品，支持图像生成、视频生成等功能。
- **妙鸭相机：**创造AI分身，生成各类风格的AI照片。

3 AI视觉搜索问答产品

AI视觉搜索问答产品依托于大语言模型，添加视觉理解模块，实现视觉问答、以图搜图、图像搜索等功能，进一步提升人机交互体验。

+ 图像理解模型

- 编码器：ViT、Clip等
- 适配器
- 大语言模型 LLM

自然语言与视觉内容的映射视觉语言模型

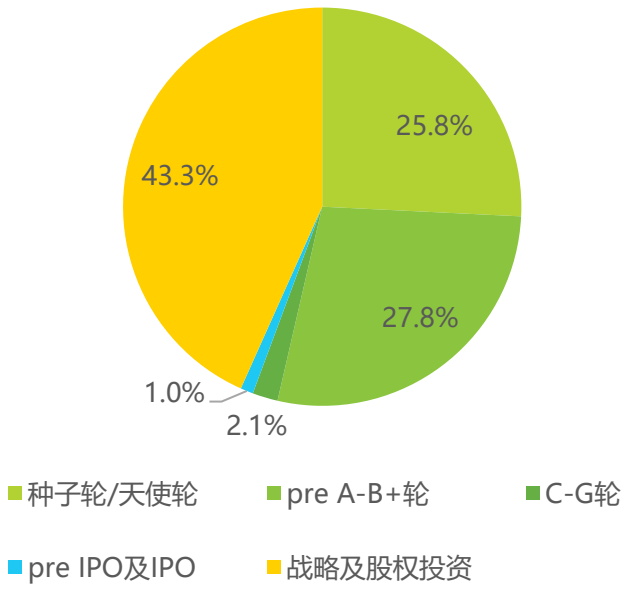
- **豆包：**字节跳动的AI问答产品，已开放图像理解推理能力
- **AI秘塔搜索：**AI搜索产品，提供图片搜索功能
- **支付宝：**开通“探一下”，提供视觉搜索功能

语言模态及多模态

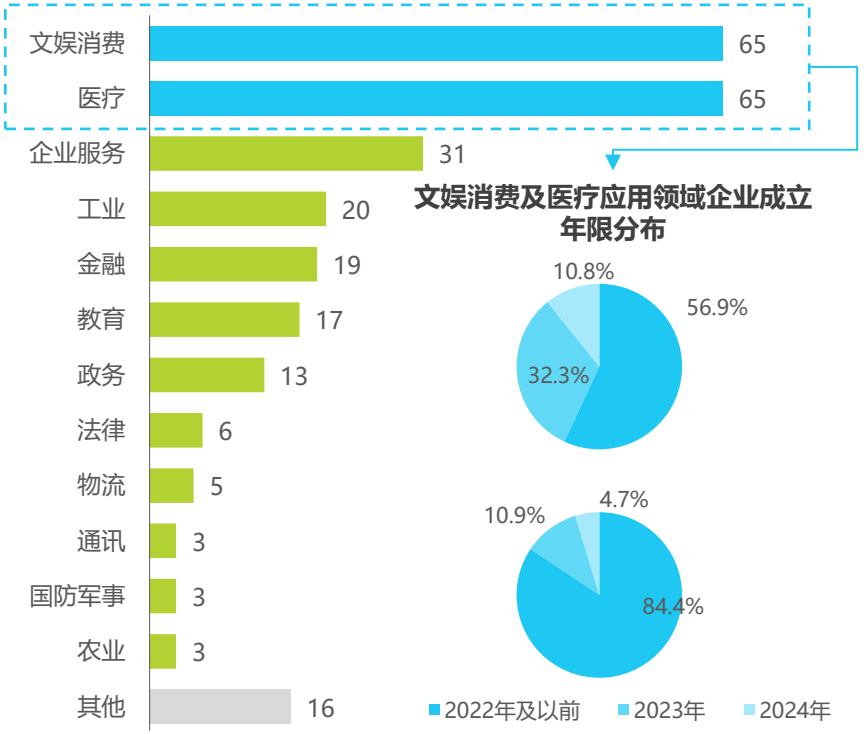
早期投资占比高，文娱消费和医疗为最热门领域

在2024年语言及多模态赛道的投融资事件中，早期投资仍占据主导地位，其中53.6%的投融资集中在C轮以前。仅有2家企业进入pre-IPO及IPO环节，其应用领域分别为文娱消费和医疗，这与热门应用领域的分布情况高度契合，文娱消费和医疗领域也是2024年语言及多模态赛道中最受关注的应用领域。在文娱消费领域，近半数企业为2023年及以后成立的新兴力量，其产品形态丰富多样，涵盖通用/垂类领域的大模型、AI Agent以及AI对话等。反观医疗领域，超八成企业成立于2022年及以前，产品形态也相对集中，主要以AI创新研发平台为主，例如药物分子设计平台等。

2024年语言及多模态模态赛道投融资轮次占比分布



2024年语言及多模态赛道应用领域频次分布



语言模态及多模态

多模态架构仍会侧重在生成或理解的单一路径，未来期待技术深度融合

作为大模型架构的先行模态，大语言模型的技术进展主要聚焦于预训练的scaling law、后训练思维链优化的推理模型等，在第一章的技术环境中已做展开。本页主要探讨多模态大模型的定义范围、技术进展及发展趋势。市面上，多模态大模型的定位范围多有不一。从定义来看，多模态大模型是能够处理和理解多种不同类型数据（文本、图像及音视频等）输入输出，实现跨模态的理解和生成任务的人工智能模型。从分类来看，艾瑞认为，多模态大模型可分为“以DiT架构为代表的生成向”和“以MLLM架构为代表的理解向“两类。其中，DiT架构的生成向多模态大模型同样为视觉模态中提到的生成路径模型，如Sora、通义万相、VIDU、可灵等模型产品。MLLM则广泛应用于AI搜索问答对话产品，以支持多模态信息的输入输出，如ChatGPT、豆包、kimi等C端产品。然而目前，生成向产品通常无法提供多模态信息的处理分析，理解类产品无法提供视频生成功能，两类产品的技术路径与模型能力尚未达到底层融合。未来，基于AGI角度出发，如何更好地整合不同模态的生成和理解，将是多模态模型发展的关键方向。多模态大模型能力的技术路径达到融合，兼具多模态的生成向与理解向能力，将进一步扩大多模态大模型能力的应用范围与场景价值。

多模态大模型技术路径

语言模态及多模态

以大语言模型为技术基础的AI产品多在大厂射程内，尤其是C端领域生态

因多模态生成产品已放在视觉产品去讲，本页产品主要聚焦语言类产品及以MLLM架构为主的多模态产品。而从LLM及MLLM技术基础的产品类型来看，中国AI语言及多模态的典型产品大致包括AI搜索问答翻译产品、AI情绪陪伴产品、AI Coding产品、To B的企业员工及To C的个人助手几类。从市场厂商参与者角度来看，大厂凭借技术、数据、资金资源和市场理解等多方面的优势，在多类型产品占据了主导地位，尤其是To C领域。艾瑞认为，未来更多AI创业机会或诞生出小而美的To B市场，面向垂域需求，深耕业务理解与工作流优化，打造基于AI产品能力的一体化解决方案，在前期实现市场开拓与资源积累。

2024年中国AI语言及多模态典型产品

1 AI搜索问答翻译产品

基于人工智能技术，向用户提供智能搜索、即时问答和多语言翻译等功能的智能工具产品。目前广泛应用于C端用户，厂商产品多以免费形式更多占领C端用户心智。

- **秘塔AI搜索：**提供系统性内容搜索及问答交互。
- **豆包：**字节跳动旗下产品，支持文本、语音及图片交互。
- **Kimi：**月之暗面旗下C端产品，擅长长文本分析，支持深度思考。
- **DeepSeek：**提供开源模型，并在C端开放产品，支持深度思考。

2 AI情绪陪伴产品

基于人工智能技术，向用户提供情感支持、陪伴互动等功能的智能工具产品。相较于功能性对话，情绪陪伴产品聚焦情感方向，现已和语音模态紧密结合，未来将进一步结合多模态交互、硬件产品提供更优质体验、情感连接的AI产品。

- **星野：**Minimax旗下产品，提供个性化定制人设、剧情等智能体交互。
- **WoW：**美团旗下产品，提供角色扮演、虚拟朋友等产品功能。
- **猫箱：**字节跳动旗下产品，由内部团队Flow开发，豆包大模型驱动。

3 AI Coding产品

利用人工智能和机器学习技术，通过理解人类语言描述来自动生成代码的工具产品。受益于思维链优化、强化学习对coding能力的加强，AI Codng产品正逐步由早期辅助的Copilot产品向主动性更高、编程能力更强的AI Agent产品形式演进。

- **通义灵码：**基于通义大模型的智能编码辅助工具。
- **文心快码：**基于文心大模型的编码辅助工具。
- **Trae：**字节跳动推出的免费AI IDE产品。

4 AI企业员工/个人助手

1）B端：实现企业员工角色，结合企业业务系统，打造各类AI Agent产品，高效处理企业业务问题

2）C端：实现个人助手角色，提供个人日程安排、知识管理、工作生活辅助等功能

- **企业级Agent产品方案：**提供Agent平台、或一体化Agent产品方案。
- **手机智能体助手：**如智谱AutoGLM、荣耀的YOYO助手。
- **知识管理助手：**如微信-IMA知识库，提供个人知识管理功能。

AI产品商业模式解析

AI产品变现路径暂以项目制与订阅制为主流，新产品或伴随新商业模式

当下，大模型在G端与B端落地仍以招投标的项目制部署为主，由运营商、云厂商、AI厂商等几股供给侧力量主导，在国央企率先开展AI大模型的实践落地。虽然AI产品使用成本是按量/次调用的模式，但目前除项目制外，AI产品较少会以按量付费方式（部分AI图像生成、视频生成产品会开启按生成次数付费的窗口）向用户收费，而是更多采用用户更熟悉接受的SaaS订阅制，初期以免费体验为渗透点，慢慢引导用户接受AI产品的付费订阅。而AI产品的订阅制收费逻辑可分为三类：1）对AI功能模块的额外收费；2）添加AI功能的整体产品价格上调；3）AI原生应用的产品收费。展望未来，AI产品或更多主张提供按效果付费的模式，并以AI Agent形式为企业提供个性化产品服务，在用户订阅数量减少、客户IT支出缩减的大环境下完成更多AI产品拓展与用户付费转化，当然这种新商业模式对企业现金流、AI产品效果、价值定价评估均提出了更高要求。此外，AI让C端产品玩法更加多样，尝试添加更多AI功能点，以单点形式花样吸引用户付费，提供更多产品付费点与转化变现渠道。

AI产品商业模式探索

AI主流商业模式

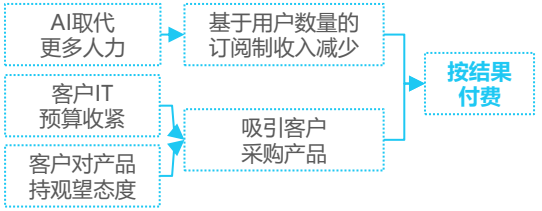
项目制收费	提供定制化开发项目收费，由软硬件+服务构成，常用于**to G和to B领域**	**如：XX大模型服务平台建设招标项目**
MaaS收费	Model As a Service，基于平台按量提供AI能力及产品服务，常用于**to B领域**	**如：火山方舟、百炼等MaaS平台**
SaaS收费	按订阅制收费，按一定使用周期收费，目前在**B端与C端**广泛应用	**如：B端chatbot；WPS会员订阅；**
按量收费	按使用量收费，**B端**主要为API调用次数，**C端**主要是产品/工具使用次数	**如：AI开放平台；触手AI积分充值**
流量变现	**针对C端需求**推出单点工具吸引流量积累，通过广告、平台抽成等方式将流量变现	**如：抖音、快手等短视频平台**

AI时代浪潮对B端与C端的产品生态格局带来影响，用户市场载体与流量时长分布发生改变，如网站浏览时长变短，用户转向AI搜索等AI原生应用。

AI创新商业模式

按对话结果/应用效果收费

根据对话结果、产品应用效果付费。



如：海外CRM厂商Zendesk在2024年9月推出基于结果的定价模型；AI客服创企Sierra同样采用“效果计费模式”

按功能点收费

AI功能为产品带来更多玩法，从商业变现角度提供更多单功能点付费的可能性。

B端客户在采购产品时看重ROI，ROI是推进企业数智化转型的重要衡量指标。大模型加持下，AI产品质效得到显著提升，采用新商业模式-按产品效果付费，客户只需为成功效果买单，可有效提升其采购意愿。

AI产品出海化尝试

出海成为企业扩市场扩营收的关键性策略，产品方向与生存法则为何？

2024年，在国内厂商卷价格、卷应用、卷生态的同时，不少企业将目光放向海外市场。以阿里云、字节跳动、Minimax为代表等大模型厂商出海动作频频，其中，阿里云主要聚焦2B电商的AI赋能与数据中心建设，字节跳动在AI图片生成、AI视频编辑、AI社交、AI教育等领域进行了多方产品布局，而Minimax与零一万物等大模型创企也下注在图像视频与AI社交等赛道。此外，一些垂直业务厂商及出海创企表现同样亮眼，如图像编辑厂商恒图科技、以插件工具切入的Monica.AI、以NSFW（Not Safe For Work）切入的Crushon.AI等。从产品方向来看，目前AI对话、问答、搜索等用户流量仍把控在ChatGPT、Perplexity等主流产品当中，AI出海产品大多落在AI图像/视频、AI社交/情感陪伴两大赛道，在下载量、活跃度、留存时长均处于榜单前列。然而这两类产品也是面临支付渠道限制与监管合规风险的重要品类，出海厂商需在版权水印、违禁词审查、风控警戒等方面做严格管理，确保AI产品在海外的长久健康运营。

AI产品出海市场总览

出海背景

- 海外更广阔的**市场空间**
- 海外用户的**产品心智与付费习惯**更加成熟，付费转化率高
- 一些海外市场**竞争激烈程度较低**，且容易获得**先发优势**
- AI加持下，**语言文化壁垒降低**，能以更低成本实现内容本地化
- 部分海外市场**监管较为宽松**，利于产品上线、扩张及循环发展

出海策略

地区选择

- 美国、英国、澳大利亚等英语国家市场
- 中东、东南亚、非洲等新兴国家市场
- 日本等垂直文化国家市场
- ……

团队情况

语言 | 文化 | 用户偏好/习惯 | 本地配置

产品策略

- PMF：产品市场匹配度，产品方向
- GTM：市场策略，SEO营销、SEM营销、KOL等

AI产品方向与出海企业矩阵 –产品列举

	大厂	大模型创企	垂直业务企业	
AI图像/视频编辑生成类	阿里-Pic Copilot 字节跳动-Capcut 快手-KLING	Minimax-Hailuo AI 爱诗科技-PixVerse HeyGen-HeyGen	万兴-Media.io等 稿定-insMind LibAI Lab-Cutout Pro	恒图科技-Fotor 网旭科技-Picwish
情感陪伴类	字节跳动-AnyDoor 百度-SynClub	Minimax-Talkie 西湖心辰-Joyland 昆仑万维-Linky	作业帮-Poly.AI 逐鹿行-Museland	Crushon.AI-Crush on Language Power-Hlwaifu
教育、营销、写作、搜索对话等工具	字节跳动-Gauth 字节跳动-Cici AI	零一万物-PopAI	作业帮-Question AI	蝴蝶效应-Monica AI 个人开发者-FlowGPT

出海挑战

- **海外市场认知：**出海厂商面临人群种族、文化差异、行为习惯等认知差异，需从组织、人才、团队及运营等方面做好本地化产品与市场策略
- **海外需求认可：**出海面临欧美等海外厂商竞争，需建立消费者对中国品牌信任度
- **数据跨境隐私：**出海厂商在数据上需关注训练数据来源、数据跨境传输及用户数据隐私等多方面
- **监管合法合规：**出海厂商要了解当地外资政策、投资限制等信息，且严格遵守版权产权保护、运营合法合规等监管要求

04/ 中国人工智能产业标杆案例

AI - Coming

提供企业一站式大模型与AI原生应用开发及服务平台，顶层应用种类丰富

字节跳动的人工智能产品布局主要由四层构成。底层是基础大模型，以自研豆包系列为主，含多种类型模型并提供第三方模型，为上层提供核心能力。其上是大模型开发平台火山方舟，具备多种功能，方便开发者定制优化。再上是智能体开发平台，扣子支持零代码搭建 AI 应用，使用群体广泛，HiAgent 更专注于搭建企业应用。最上层大模型应用层，既面向普通用户提供写作等功能，也为专业开发者和企业提供各类助手工具，形成完整生态，满足多样需求。

字节跳动人工智能产品矩阵

大模型应用层

豆包　即梦AI　即创　TRAE　飞书
剪映　Ola Friend　河马爱学　…

行业应用场景
智能座舱/智能终端/在线教育/社交娱乐/智能客服/营销提效/消费零售…

大模型开发平台 — 火山方舟 *一站式大模型开发平台*

体验中心	模型精调	模型评测
模型推理	Prompt优化	智能体广场

智能体开发平台

扣子
- 零代码快速搭建AI应用
- 1w+插件繁荣生态

HiAgent
- 海量企业应用模板/插件
- 提供更强安全防护能力

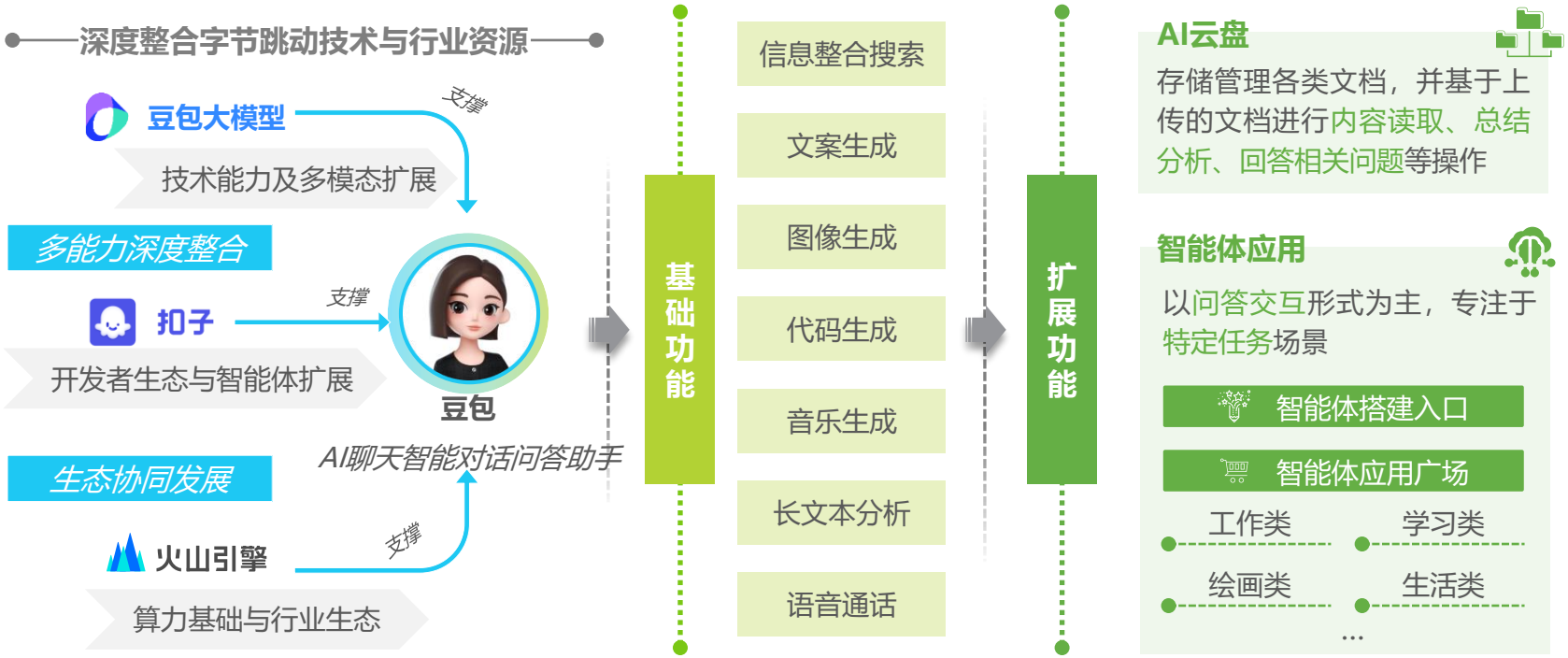
基础大模型 — 以自研豆包系列为主，包含大语言模型、多模态模型、视觉大模型、语音大模型等，同时提供多个第三方模型

豆包大模型1.5 pro 大语言模型	豆包1.5 视觉理解模型 多模态大模型	豆包·语音合成模型 语音大模型	…
…	豆包·视频生成模型 视觉大模型	豆包·声音复刻模型 语音大模型	豆包·角色扮演模型 大语言模型

融合多模态交互与生态协同技术，提供个性化智能服务与高效体验

豆包是字节跳动推出的多模态 AI 助手，凭借其独特的生态优势与丰富的应用场景，迅速成为 AI 领域的标杆产品。其生态布局深度融合了豆包大模型系列的技术能力、扣子的开发者生态以及火山引擎的强大算力，为上层丰富功能的快速运行提供了有力支撑。这些功能涵盖文案创作、PDF 问答、长文本分析、学习辅助、图像生成、信息搜索与整合以及智能体等多个方面，全方位满足了人们在工作、学习和生活等多场景下的多样化需求。

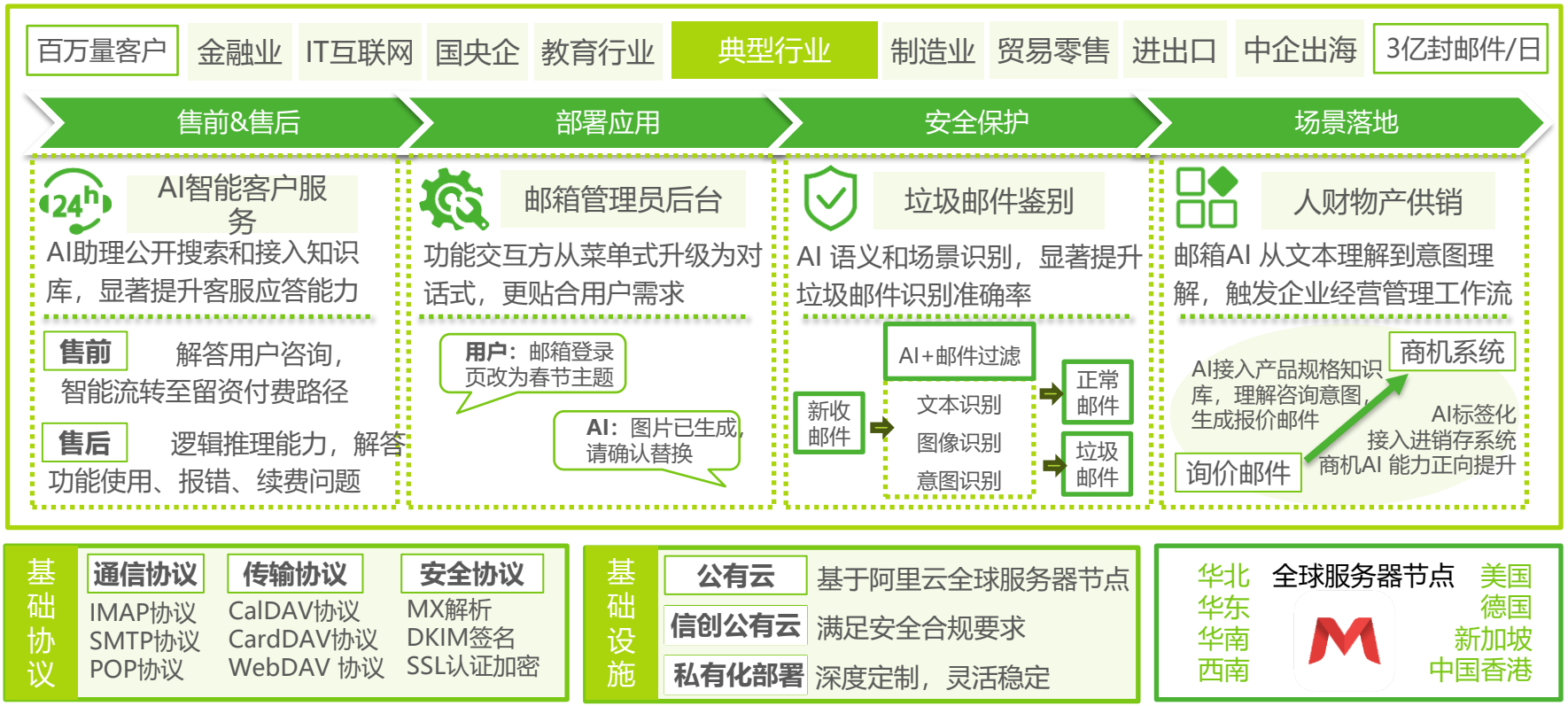
豆包生态及功能架构展示



服务百万企业，国内领先的企业邮箱产品

阿里邮箱在2009年发布上线，目前是钉钉旗下的产品。企业邮箱具有品牌形象、安全合规、国际通用的差异化价值，而阿里邮箱与钉钉的融合互通，也探索出一套即时沟通+异步沟通的全新协同办公方式。阿里邮箱经过多年产品积淀，已实现对Outlook等同类产品的国产化替代。人工智能技术在企业邮箱场景的成熟落地，不仅提升了阿里邮箱在反垃圾、数据安全、内容生成和多语言翻译等场景的服务能力，也将电子邮件作为AI工作流的信息输入端，调用AI能力解决客户的业务问题。

阿里企业邮箱产品架构



细分场景接入AI，自动化能力发挥邮箱业务价值

阿里邮箱每日服务企业收发3亿封邮件，不仅是重要的通信工具、协同办公工具，也承载了大量的企业工作任务流转。例如企业工单邮件、业务审批邮件、商务谈判邮件、营销信息邮件等。借助AI的多模态内容生成、数据分析处理、逻辑推理能力，阿里邮箱通过接入AI模型，可以将邮箱数据格式化，成为自动化工作流的“元数据”，让邮箱数据接入企业的人、财、物、事等业务系统。

基于邮箱的智能业务场景

高阶：场景邮件流转

邮件对接企业人财物产供销存系统
AI识别业务信息，触发RPA工作流

进阶：三方应用接入

Webmail插件　邮件客户端插件
钉钉AI助理　宜搭　审批　多维表格

- Windows
- macOS/iOS
- Android
- Harmony

- 面试助手
- 账款小管家
- AI法律查询
- 宜搭工作流

- 网页邮箱+主流浏览器

初阶：邮件文本处理

AI多模态+逻辑推理+联网搜索
高效处理邮件读、写、搜

- 写
  - 总结邮件/感谢邮件/谈判邮件/工单邮件/营销邮件
  - 多语言/图片/架构/流程图
- 搜
  - 搜索知识库和互联网，智能推断邮件内容和意图
  - 理解邮件信息，文本搜索升级为知识搜索
- 通用AI大模型
- 读
  - **单封**邮件文本理解
  - **多邮件**上下文场景理解
  - 多语言/多图片理解
  - 系统业务数据理解
- 推理判断
  - 提炼邮件内容，自动调用RPA工作流和OA系统
  - 上下文推断邮件关联的人员、业务归属、决策流程

基础：多种大模型接入

通义千问　DeepSeek　Minimax
零一万物　月之暗面　ChatGLM

AI = 模型 X 场景 X 数据 X 算力

- 阿里邮箱
- 钉钉邮箱
- 邮件群发

邮箱数据
钉钉数据
私有数据
联网数据

阿里云
灵骏集群
百炼平台

专注于大模型底层技术研发，其通用模型和推理模型达到业界领先水平

DeepSeek成立于2023年，由幻方量化创始人梁文锋创立，定位为专注于人工智能基础技术研究的科技公司，致力于探索AGI的实现路径。截至目前，DeepSeek已经推出了包括通用模型V系列、推理模型R系列以及代码模型、数学推理模型和多模态模型等在内多款开源模型。其最新版本通用模型V3-0324实现660B参数下的轻量化部署，推理模型R1在数学、代码、自然语言推理等任务上，性能比肩OpenAI o1正式版，引发产业界高度关注。DeepSeek官方APP自上线以来至2月9日，累计下载量超1.1亿次，周活跃用户规模峰值接近9700万。

2023年7月
DeepSeek公司成立

2023年11月
相继发布开源代码模型DeepSeek Coder和通用大语言模型DeepSeek LLM

2024年5月
发布开源MoE模型DeepSeek V2

2024年8月
合并DeepSeek Coder V2和DeepSeek Chat模型，发布融合通用与代码能力的DeepSeek V2.5

2024年11月
发布推理模型预览版DeepSeek-R1-Lite

2024年12月
DeepSeek V3上线并开源，性能比肩领先闭源模型

2025年1月
发布官方APP，支持联网搜索与深度思考模式
推理模型DeepSeek R1正式发布

2025年3月
发布新版本V3模型DeepSeek-V3-0324，推理、前端开发、中文写作、中文搜索、工具调用等能力有所提升

DeepSeek发展历程及模型家族

通用模型

DeepSeek LLM

DeepSeek V2

DeepSeek V3

MoE架构的大语言模型，具备多任务泛化能力，在知识问答、长文本处理、代码生成、数学问题求解等方面性能领先

代码模型

DeepSeek Coder

DeepSeek Coder V2

沿用DeepSeek V2模型结构，具备全球顶尖的代码能力和数学能力

推理模型

DeepSeek R1

在后训练阶段大规模使用强化学习技术，在极少标注数据的情况下极大提升模型推理能力

数学推理模型

DeepSeek Math

针对数学相关数据进行预训练强化，提升模型在复杂数学问题求解上的可靠性与精准度

多模态模型

DeepSeek VL

使用集成视觉和语言数据的方式进行预训练，在不丢失语言能力的情况下融入多模态能力

来源：艾瑞咨询研究院自主研究及绘制。

采取开源策略，通过工程优化与算法创新突破模型的性能与成本瓶颈

DeepSeek在工程和算法层面的持续创新，不断推动大模型突破性能和成本的瓶颈。其中，DeepSeek V3通过工程优化在资源有限条件下实现低成本、高效能的模型训练与推理，而DeepSeek R1 Zero（Deepseek R1的初始版本）则证明了纯强化学习路径的可行性，跳过了传统大模型训练流程中的监督微调，无需大量人工标注数据，依靠简单的奖惩机制实现模型推理能力的自我提升。与此同时，DeepSeek坚持开源理念，将模型权重、技术论文和训练细节及部分工程代码开放出来，吸引开源社区的参与者形成正向循环，未来有机会构建更强大的生态壁垒。

DeepSeek开源策略与关键创新

DeepSeek主导模型开发

DeepSeek开源内容

- **模型权重**
- **技术论文和训练细节**
- **部分工程代码**

开源社区参与推理、再训练和微调

开放生态

吸引开发者贡献代码，利用社区反馈改进模型性能，形成技术反哺

DeepSeek V3：工程优化

混合专家架构MoE	多头潜在注意力MLA	多token预测MTP	混合精度框架FP8
通过多个专家网络协同工作，提高模型的多样性和效率	通过优化KV缓存的使用，减少内存占用，提高推理效率	一次预测多个token，训练更长更多数据，提升训练和推理效率	将部分参数压缩到FP8，在资源有限的情况下保证模型质量

DeepSeek R1：范式创新

第一阶段：训练DeepSeek R1 Zero

DeepSeek V3 Base

采用**GRPO（组相对策略优化）**进行强化学习训练

- **准确性奖励**：评估模型输出内容是否正确
- **格式奖励**：评估模型是否使用标准化格式输出推理过程和最终答案

DeepSeek R1 Zero

第二阶段：训练DeepSeek R1

DeepSeek V3 Base

- 冷启动：由人类注释者和R1 Zero生成的高质量链式思考数据进行双重验证，**提升推理链的语义连贯性和可读性**
- 推理为中心的强化学习训练：**提升模型推理能力**，同时引入语言一致性奖励，**减少语言混合问题**
- 拒绝采样和全领域监督微调：**将模型能力泛化至全领域**
- 全领域强化学习训练：进一步**提升模型帮助性和安全性**

DeepSeek R1

来源：艾瑞咨询研究院自主研究及绘制。

05/ 中国人工智能产业趋势洞察

AI - Coming

AI Agent的进阶

模型能力、工具生态、市场需求协同共振，持续推动Agent的通用性演进

随着人工智能技术的不断发展，大模型推理能力与工具调用生态的迭代升级，叠加市场对可自主决策、规划并执行多步骤任务的智能体（Agent）的迫切需求，AI Agent迈入从认知到执行的突破性拐点。当前，AI Agent已初步实现跨域任务整合（如单指令触发多领域任务协同）、任务链自主拆解及深度决策支持等场景应用，但执行成果仍面临挑战，可能存在大模型幻觉、规划逻辑稳健性不足、工具调用能力有限等问题。值得期待的是，迈入拐点已激活市场创新动能：开源生态蓬勃发展，DeepSeek、Qwen等优质开源模型释放技术红利；推动大模型与外部数据源、工具和服务进行高效标准化连接的MCP等工具协议的生态加速扩容，自2024年11月Anthropic发布MCP（模型上下文协议）以来，已超1100个社区服务器与官方集成落地。未来，在模型能力、工具生态、市场需求的协同共振下，AI Agent将向复杂任务持续演进，加速走向“决策-执行-反思”的自主闭环能力顶点。

AI Agent进阶：向着通用场景升级

物理AI的演进

作为融合数字智能与物理世界的桥梁，物理AI正成为下一代AI竞争高地

物理AI作为融合数字智能与物理世界的创新范式，除硬件设备和软件系统的发展外，其AI发展核心在于构建具备多模态感知与具身行动能力的智能系统，以保障物理实体行为的反应速度、操作精度、决策智能度等。这一领域发展需解决传统AI在具身交互中的瓶颈——软硬件技术升级、跨行业生态协作、伦理规范等问题，以推动人工智能从虚拟助手向实体协作伙伴演进。目前，物理AI已经在机器人动作表现、工业机器操控等场景初步体现，物理AI甚至可在老年陪护、家庭管家等场景进一步发挥普惠作用。

物理AI演进路线及升级挑战

DeepSeek的产业价值

推动技术普惠与平权，加速大模型向产业端和消费端的应用渗透

DeepSeek让开源模型以低成本、高性能的方式进入公众视野，其影响不仅在于工程、算法层面的突破，更重要的是重构了技术扩散的路径，促进人工智能在应用层面的落地。在技术路径方面，DeepSeek进一步推动了对于AGI的探索，有可能带来AI技术范式从监督微调向自我推理的演进，以及竞争范式向“以效率换规模、以创新换算力”的转变。DeepSeek不仅通过开源的方式加速了技术迭代，其高效的技术架构还显著降低了大模型的训练和推理成本，使得中小企业和个人开发者能够更容易地应用大模型技术。技术侧与应用侧的双螺旋进化将促进人工智能的普惠与平权，推动行业迈向开放、繁荣的应用生态。

DeepSeek对中国人工智能产业的影响

技术侧：*对AGI实现路径的启示* ››

强化学习

- 重新审视传统大模型训练方法，可能过度依赖监督学习，强调模拟人类思维方式
- 证明无需预设推理框架的可行性，为不受人类先验约束的人工智能提供新的可能性

合成数据

- 通过GRPO快速迭代高质量链式思考数据，在特定场景下可以达到人类标注数据的效果
- 展示了模型生成数据到数据反哺模型的正向循环，有望提升行业整体的数据使用效率

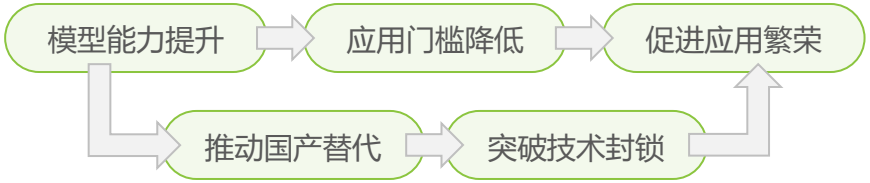
模型蒸馏

- 证明模型蒸馏可以有效将大型模型的知识迁移到小型模型中，显著提升小型模型的性能，为行业提供了经济可行的应用路径

技术创新

- 利用工程和算法创新可降低对Scaling Law的依赖，推动行业扭转堆砌算力的竞争范式，避免盲目追求模型的参数规模

应用侧：*对B端+C端应用落地的作用* ››



B端应用

- 政府、金融、能源等敏感领域可以利用国产芯片+微调版开源模型，实现大模型**本地部署**
- 医疗、法律等专业场景将出现更多的**垂直模型**，软件厂商需强化工程平台和应用能力

C端应用

- DeepSeek的出圈提升公众对AI的认知，实现全民**市场教育**
- 新一轮应用开发浪潮来临，AI原生的**杀手级应用**将可能出现
- 模型蒸馏技术将促进端侧推理的爆发，推动**AI硬件**的落地

来源：艾瑞咨询研究院自主研究及绘制。

人工智能安全治理体系的构建

安全是产业发展的红线，需构建面向新一代人工智能的治理框架

随着人工智能技术的快速迭代，其潜在风险已从理论威胁演变为现实挑战。相关治理体系在应对生成式AI、自动驾驶等高复杂度场景时，仍存在标准缺失、监管滞后等诸多问题。人工智能的安全风险包括内生安全和应用安全两部分，其中内生安全聚焦技术层的数据与算法风险，应用安全覆盖更广泛的社会影响与伦理挑战。构建动态、前瞻的治理框架，可以考虑从技术、商业、法律、伦理等多角度协同发力，通过政策引导、行业自律与国际协作，确保在释放人工智能技术红利的同时，守护安全发展的底线。

人工智能面临的主要安全问题

内生安全

数据安全风险

数据隐私保护	提示注入风险	模型投毒风险
使用公司机密或个人隐私等未经授权的训练数据	通过刻意设计的提示词诱导模型输出违规内容	攻击者篡改训练数据或标注，导致模型偏离预期

算法安全风险

模型幻觉问题	可解释性缺陷	算法偏见问题
模型生成看似合理但与现实世界事实不符的内容	黑箱机制下模型内部运作机制难以被理解和审查	因数据偏差或设计缺陷导致模型出现歧视性输出

应用安全

内容安全风险

虚假信息传播	信息茧房加剧	违法犯罪滥用
滥用模型生成内容，造成网络环境虚假信息泛滥	根据用户反馈断优化推荐内容，加剧认知固化	人工智能被应用于欺诈、攻击等违法犯罪活动

责任归属风险

模型生成内容的知识产权归属不明确，在高风险场景中如模型出现错误决策，事故追责存在困难

情感伦理风险

过度依赖人工智能，可能导致人机交互中的情感依附，人类权力的让渡可能引发根本性道德伦理问题

来源：艾瑞咨询研究院自主研究及绘制。