

AI

全球人工智能发展 研究报告

鼎帷咨询战略创新研究院

2024年9月

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

序言

01

人工智能是第四次科技革命的主导力量。是大国竞争的战略制高点。从学科来讲，它将是新阶段的带头学科，从技术来讲，它将是各领域新兴技术背后的元技术，从能力来讲，它将是所有系统的生成器和物理世界的模拟器，泛化能力不断增强的大模型已经成为一项通用技术，正在引领人工智能走向大一统。

02

当前，中国人工智能人才、投资、技术、AI论文、专利、算力等核心指标第一次与美国一起位列第一梯队，领先日德英法等发达国家。

03

我们研究人工智能的目的是想为中国人工智能行业发展提供绵薄之力，在人工智能发展的关键时期为公众做行业科普，后续我们将展开国际巨头的发展战略研究，为中国人工智能企业创新发展提供研究信息。

01

发展历程及现状

AI的发展将会历经狭义人工智能ANI、通用人工智能AGI和超级人工智能ASI三个阶段，当前尚处于ANI阶段，进入AGI阶段需要的抽象因果演绎能力当前并不具备，在未来20-40年有可能会出现

	 ANI：狭义人工智能	 AGI：通用人工智能	 ASI：超级人工智能
基础定义	机器学习阶段，是在有限领域内执行特定任务或功能的系统	机器智能阶段，能够执行人类能够完成的所有智力任务	机器意识阶段，是指几乎在所有领域都超越人类智能，并且在每项认知任务上都能够超越人类的人工智能系统
阶段特点	擅长某一特定功能，但缺乏人类认知的一般智能和多功能性 机器仍然属于工具和产品的范畴	具有“人格”的基本条件及人类水平的广泛智能 强大自我学习能力和跨领域能力 彻底改变多个行业，如教育、娱乐、交通和健康	快速处理能力：可以在几秒钟内处理和理解大量信息，建立联系并得出人类永远无法实现的见解 知识能力：深刻理解各科学学科，能在物理、医学和太空探索等领域做出突破性发现和进步 解决问题的能力：能应对气候变化、贫困和资源分配等复杂的全球挑战 数据集分析及场景模拟能力：生成创新解决方案解决问题。决策由逻辑和对知识的理解驱动，优先考虑人类和地球的福祉
测试条件	机器学习 自然语言处理 深度 AI NLP	图灵测试 机器人大学生测试 咖啡测试 就业测试	
技术特征	基于规则的系统：遵循一组预先确定的规则来执行特定任务 专家系统：执行人类专业知识任务 决策树系统：根据预先确定的规则做出决策 人工神经网络：可以识别模式并根据该信息做出决策 进化计算：使用进化和自然选择的原理来适应和提高性能 模糊逻辑系统：使用模糊逻辑，允许表示不确定或不精确信息来做出决策的逻辑 贝叶斯网络：使用概率推理根据不确定或不完整的信息做出决策	可解释 AI：决策过程将更加透明、可解释 自主 AI：在没有人工干预时独立运行，这包括自动驾驶汽车、无人机和机器人技术的进步 量子 AI：有望解决目前经典计算机难题，彻底改变密码学、材料科学和复杂优化问题等领域 社会公益AI：预计AI 发展将更多关注社会公益，应对气候变化、贫困和公共卫生等全球挑战 增强智能：将使个人和组织能够做出更好的决策，提高生产力并推动创新。 个性化AI：根据个人偏好和行为提供量身定制的体验和解决方案，为生活的各个方面带来新的个性化水平	仍未知
典型产品	2016年AlphaGo—2023年ChatGPT4	预计未来20-40年可能会出现	

对于AI智能的发展还存在更细致的七阶段论和五阶段论的说法

发展七阶段论

01 起步——基于规则的人工智能系统

- 行动和策略受到预定义的规则（即算法）的控制。擅长执行规则明确的任务，例如诊断机械问题或处理税表。但不存在学习或适应能力，他们的决定只能与他们所获得的规则一样好。

02 青春期——情境感知和保留系统

- 超越基于规则的人工智能，系统能够理解和保留上下文，但离独立思考还很远。

03 专家——特定领域的掌握系统

- 经过训练可以掌握复杂的棋盘游戏“围棋”，甚至能够击败世界冠军。这些人工智能系统在各自的专业领域展示出复杂的能力

04 思考者——思考和推理人工智能系统

- 开始模仿人类的思维和推理能力。他们不只是遵守规则或保留背景，而是模拟人类的思维过程。例如，阅读一本书并理解情节，并根据人物的行为推断他们的动机。

05 AGI——新思维的诞生

- 拥有自我意识、意识以及理解和驾驭世界的能力，能像人类一样理解、学习、适应和实施广泛任务中的知识，目前仍然是一个未实现的概念。

06 ASI——超越人类的人工智能

- 这种人工智能可以解决人类无法解决的问题，以我们无法想象的方式进行创新。它有可能提供对目前我们无法理解的宇宙奥秘的见解。

07 人工智能奇点

- 这个假设的未来点，奇点代表了ASI不仅比人类更聪明，而且能够不断改进自己，远远超出人类的能力。此阶段的技术增长变得不可控和不可逆转，导致人类文明发生不可预见的变化。

发展五阶段论

01 系统人工智能

- 根据训练期间建立的概率对提示做出响应：即当前最先进的人工智能。

02 感知人工智能

- 具有典型的好奇心，并利用经验来完善对世界的信念。

03 复杂的人工智能

- 通过制定计划和实验来增加其对世界的了解。

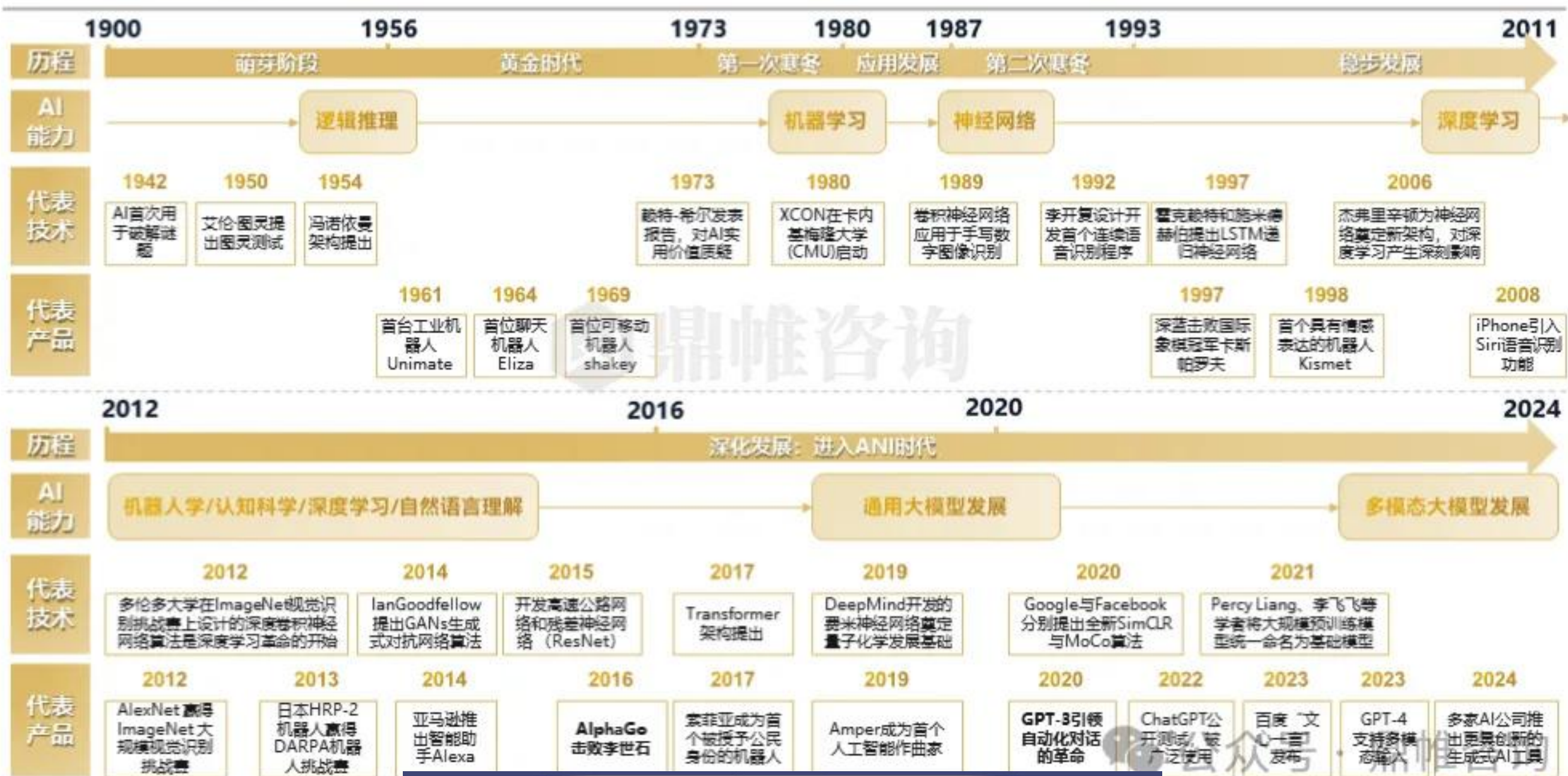
04 富有同情心的人工智能

- 能够识别他人的心理状态，最终识别出自己的心理状态，有自我意识。

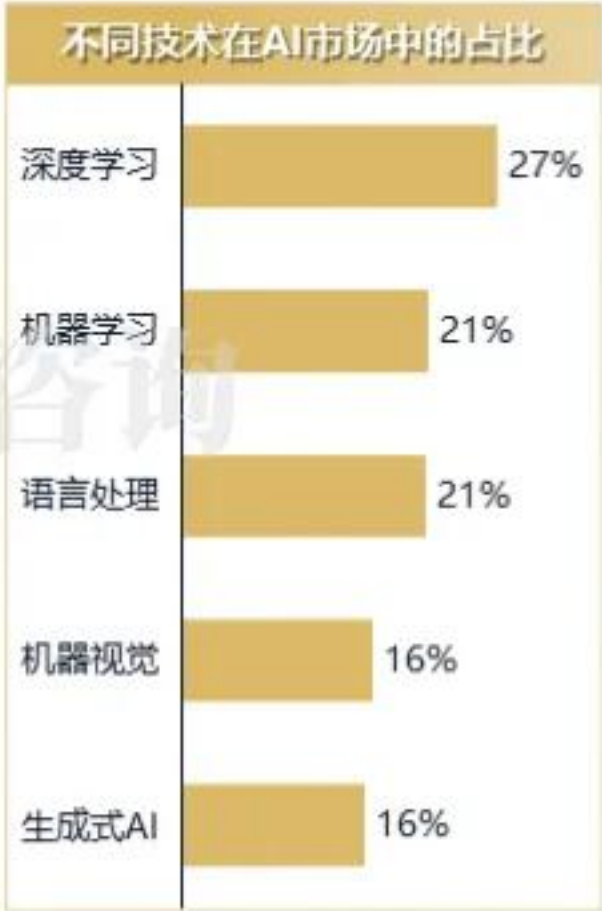
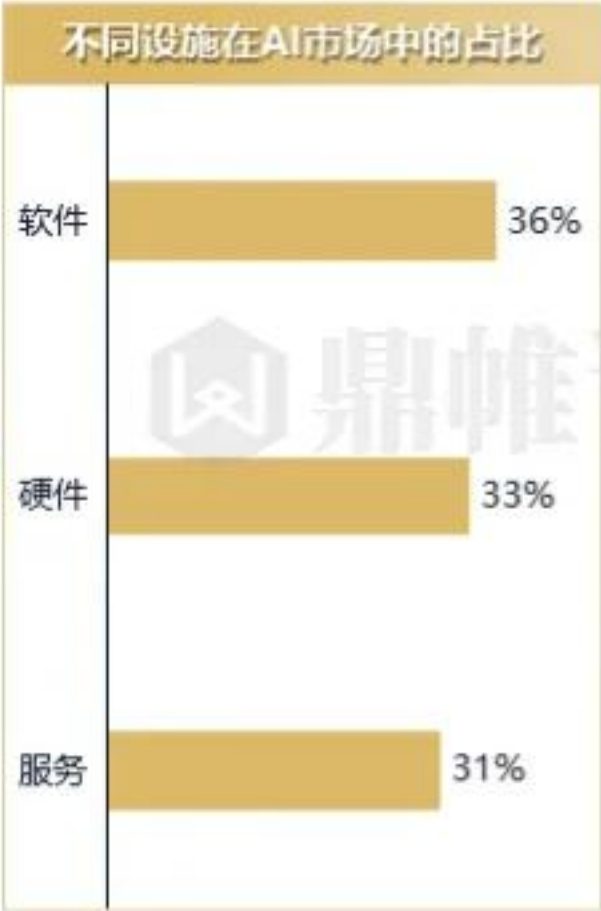
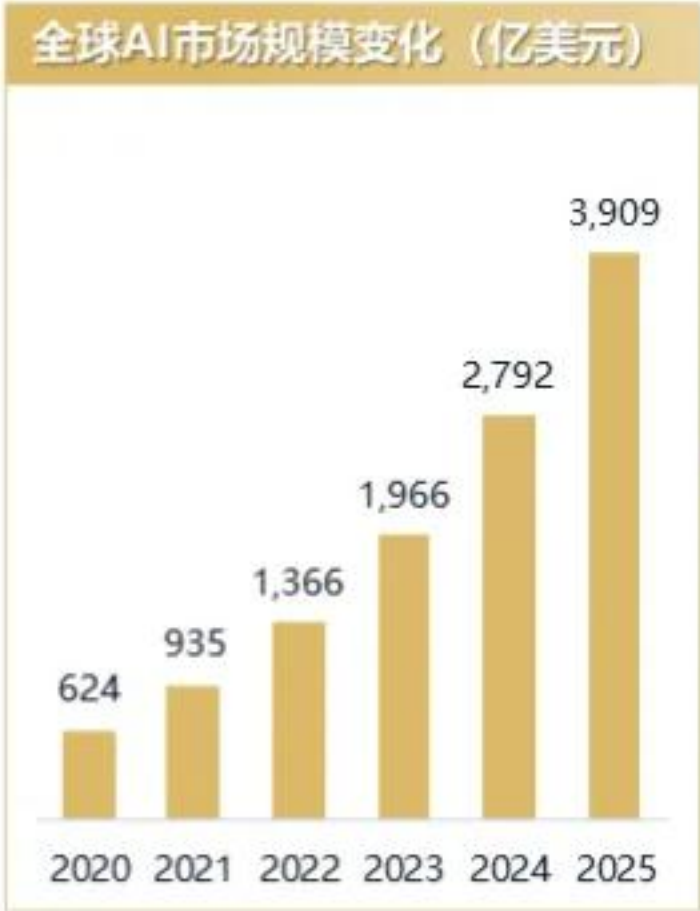
05 共享人工智能或超级人工智能

- 是人工智能与其人类伙伴互动中产生的集体智慧。

上世纪以来，AI已从早期萌芽历经两次寒冬期在2012年迈入深化发展阶段，正迅速向多模态大规模模型的高级形态演进



市场规模：近年来全球AI市场增长强劲，中国和美国以绝对优势占据全绝大多数市场份额，整体来看软件市场营收占比最大，深度学习和机器学习是当下AI市场的主流技术



市场投资：中国AI总体投资规模现居全球第二，近五年投资总额仅为美国的40%，按照目前的投资规模两者间的差距仍在逐步拉大，但仍是唯一有可能追赶并超越美国的国家

2019-2023AI投资最多的10个国家			
国家或地区	2023年投资 (百万美元)	与2019年投资 同比增长 (%)	五年总投资 (百万美元)
美国	67,911	66%	328,548
中国	15,071	-31%	132,665
英国	3,518	0%	25,541
印度	3,808	261%	16,147
德国	1,808	-11%	14,300
加拿大	2,067	40%	12,457
韩国	2,102	239%	10,348
法国	1,853	75%	10,185
瑞典	2,603	2,310%	8,281
新加坡	1,928	192%	7,005

依2023年的投资速度发展AI，中国是最有可能超越美国的国家			
国家或地区	追平2030年美国 所需时间 (年)	追平2040年美国 所需时间 (年)	追平2050年美国 所需时间 (年)
中国	14	38	72
英国	79	195	361
印度	119	291	538
以色列	141	344	636
德国	145	355	656
加拿大	164	400	740
韩国	183	447	825
法国	193	472	872
瑞典	200	486	897
新加坡	274	666	1,231

数据来源：AIPRM网站数据,源自OECD

海内外多位企业家和行业专家也坚持中美差距缩小论，中美人工智能竞争核心在于人才！人才！！人才！！！以及耐心资本的投入

海内外多位企业家和行业专家均认为提出中美整体差距已经缩小到1年或更短

The
Economist

经济学人杂志

“2023年初，专家们估计，确实存在的中国 LLMs 比美国的尖端技术落后十年。现在，同一批专家估计，跨太平洋模式的差距已经缩小到一年或更短。LMSYS Chatbot Arena 对 LLMs，将 5 个中国机器人排在全球前 20 名中。”



李开复
零一万物创始人

零一万物最新千亿参数模型Yi-Large在总榜排名世界模型第7，中国大模型中第一，已经超过Llama-3-70B、Claude 3 Sonnet。

“在一年前，我们（中国）落后OpenAI、Google 7-10年，现在，我觉得差距也就是在6个月左右，大大降低了。”



李萌
全国政协委员、
科学技术部原副部长

“个人认为，中美人工智能差距正逐渐缩小，马斯克、扎克伯格、李开复等企业家和领军学者也坚持中美差距缩小论，而媒体对于这一评价是动态变化的...但我们也要看到，OpenAI旗下的Q-star和Google旗下的Gemini有部分新技术还未公开，不知道还会有什么秘密武器。”



马斯克
领军企业家

“美国拥有最先进的人工智能，中国紧随其后。当然，中国拥有继续延展和优化的资源。中美目前是有差距，这个差距看起来大概有12个月。”

技术竞争核心在于人才

“如果说我们存在什么问题的话，其实就是人才、人才，还是人才。我们总体上人才赶不上发展的需求是个问题，人才被切断与外面的联系也是个问题。”

“吸引人才迫在眉睫，三中全会《决定》已经作了改革人才体制机制的部署，并提出了具体措施，关键是落实到位，要把培养和集聚人才特别是吸引外来人才作为战略举措。”

国内创投基金存续期短及长期投资不足

美国创投基金存续期

10至12年
资本机制相对耐心

中国创投基金存续期

3至5年
资本支持相对乏力

2024年6月，国家发展和改革委员会、国务院国有资产监督管理委员会、国家金融监督管理总局、中国证券监督管理委员会四部门同时发声，呼吁壮大“耐心资本”，推出《促进创业投资高质量发展的若干政策措施》

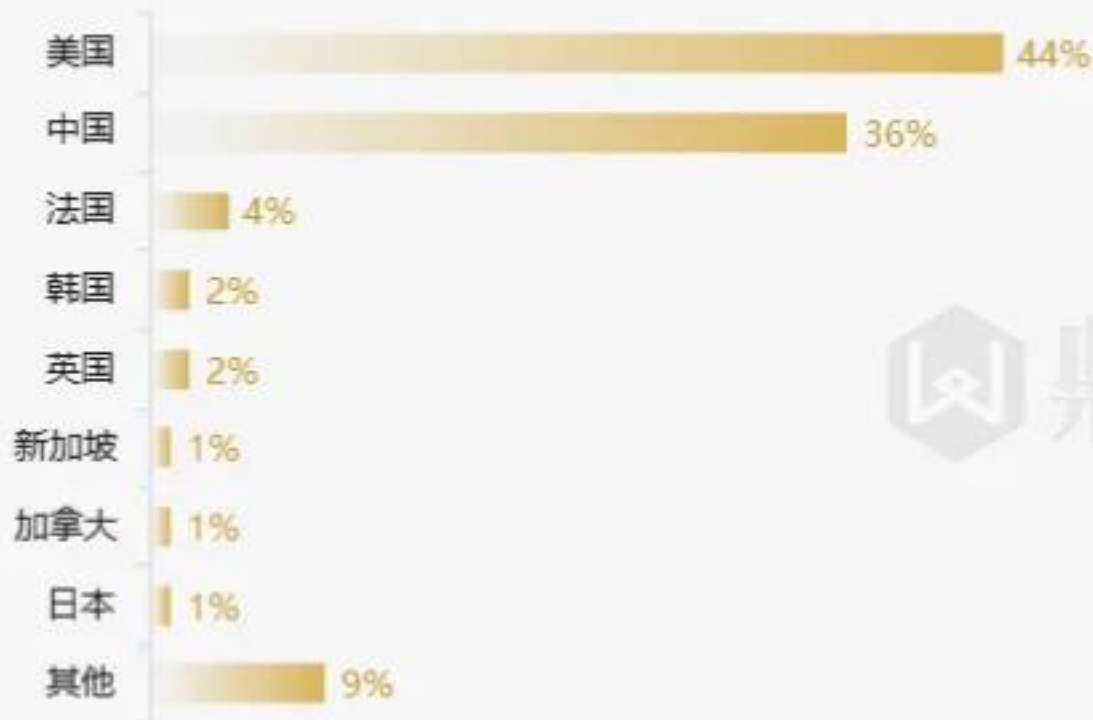
大模型：当前的大型模型仍是以语言大模型为主流，视觉、图像生成及视频生成模型等跨模态应用正在逐步发展和完善



	大型模型		
	语言模型	视觉语言模型	其他专有模型
占比	85% (69/81)	6% (5/81)	9% (7/81)
训练数据与计算	使用自然语言或文本数据进行训练	根据预训练目标在结构各异的大型图像和文本数据集上进行训练，再使用特定于任务的数据集进一步针对各种下游任务进行微调	使用音频、图像、游戏数据训练，收集定制的专有数据集不公开训练数据和硬件信息
应用	以ChatGPT为主的语言建模、聊天、代码生成、蛋白质序列预测等各领域	谷歌最新推出的PaliGemma能够接收图像与文本输入并生成文本输出	- Runway 的 Gen-2 视频生成器 - 英伟达的CALM 机器人系统

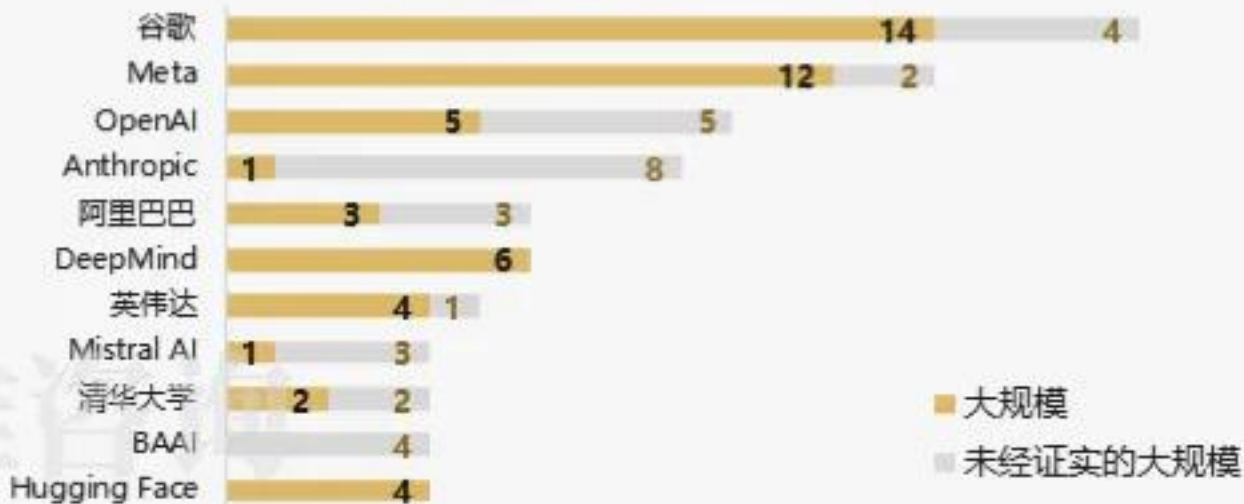
从区域看，美国大模型数量位列世界第一，中国紧随其后，以谷歌、Meta、OpenAI为代表的公司发布模型的数量最多，占据行业领导地位

全球AI大模型分布



截至2024年，全球人工智能大模型有1328个（包含同一企业、同一模型的不同参数版本），美国大模型数量位居全球第一，占44%，位居第二的中国大模型数量占比为36%。

按组织划分的大型模型



大型模型开发组织

全球

中国

主要开发者：谷歌、Meta、DeepMind、Hugging Face、OpenAI

上升组织：Anthropic、阿里巴巴

传统科技公司：阿里巴巴、百度、华为

研究机构：清华大学

新兴AI初创公司：DeepSeek、百川

开发者背景

- 主要来源：工业界
- 少数来源：学术界和政府
- 学术记录：约一半的模型有学术记录，少部分发表在同行评审期刊和会议上

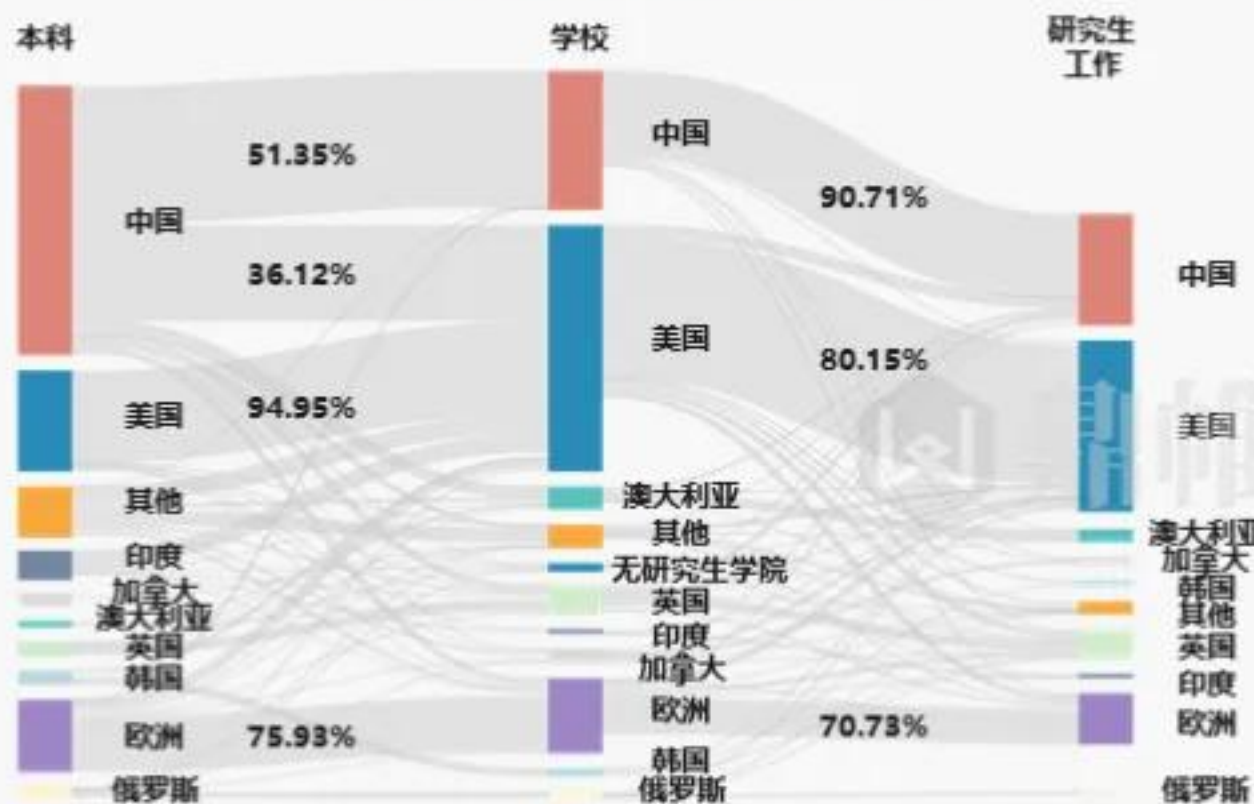
中国的表现

- 可下载大型模型：12个
- 代表性模型：阿里巴巴的Qwen
- 数据：大多训练于包含大量中英文数据的集群。
- 用途：聊天机器人、代码生成

公众号 · 鼎惟咨询

中国面临显著的人才流失问题，大量人才选择前往美国、欧洲等国家深造和工作，其中代表产品ChatGPT的17人核心团队中来自中国的就有6人

2023 年全球AI人才流动情况



- **中国:** 顶尖AI研究人员在国内比例从29%增至47%
- **美国:** 拥有75%顶级AI人才，仍是全球AI人才和顶级AI机构的主要聚集地
- **其他国家:** 英国、韩国、欧洲AI人才流入略增，印度和加拿大减少；印度留才趋势增强，全球AI人才跨国流动性下降

GPT-4o的核心17人团队中，有6人来自中国

负责人Prafulla Dhariwal

Li Jing

- 北京大学物理学学士
- MIT物理学博士
- 曾为Dall-E 3、Sora、GPT-4o都做出过贡献

Casey Chu

- 斯坦福大学获得计算数学硕士
- OpenAI四年元老

Mark Chen

- OpenAI前沿研究主管，五年元老
- MIT数学-计算机专业

Jiahui Yu

- 带领OpenAI感知团队
- 中科大少年班计算机学士
- 伊利诺伊大学香槟分校博士

Huiwen Chang

- 清华大学计算机科学实验班（姚班）学士
- 普林斯顿大学博士

Yu Zhang

- 曾在谷歌DeepMind工作过6年
- 上海交通大学计算机学士，MIT博士

公众号 · 鼎惟咨询

商业模式：未来大模型将通过MaaS模型，即平台的方式从基础层到中间层再到应用层，进而实现商业化

MaaS重塑大模型商业化

	能力	方式
基础层	大算力	提供多模态基模型以及 CV、NLP、多模态等领域大模型
	大数据	输出泛化能力
	强算法	对外开放大模型的调用接口
中间层	大模型精调	付费使用接口，直接调用基础模型
	应用开发	基于行业特色数据与知识进行精调开发行业大模型
	边缘模型部署	精准适配 AI 具体应用需求
		
应用层	模型体验	推出多样化的 SaaS 应用
	垂直领域解决方案	开发新一代智能硬件，出现现象级产品

通用大模型商业模式



02

中美AI关键竞争 能力对比

战略布局：中美均将人工智能发展列入重点战略规划，美国力求通过科技创新制定未来人工智能技术标准，中国则是通过投资教育和赋能产业力求实现人工智能基础理论的重大突破和产业升级

	美国“AI梦”	中国“AI梦”
战略目标	2024年5月美国商务部发布AI战略愿景文件 - 推进人工智能安全科学的发展 - 阐明、展示制定适当的技术标准 - 支持围绕人工智能安全的机构、社群及协调工作	2017年7月国务院发布《新一代人工智能发展规划》 - 到2025年实现人工智能基础理论的重大突破，并建立相关法律框架 - 到2030年，计划将中国打造为全球人工智能创新枢纽，核心产业规模超过1万亿元人民币，相关产业规模超10万亿元人民币
战略举措	三大支柱 重塑国家创新生态体系 从传统创新三角转向多元化“新几何形态”，培育兼顾专有和开源模型的技术生态 重塑国际秩序 制定新外交政策，掌握AI创新威慑力，扩大开源情报应用 重塑经济持久优势 提升美国制造为地缘政治力量，加强技术经济伙伴关系，成立未来工作委员会争夺人才	六大模块 构建开放协同的人工智能科技创新体系 加强人工智能军民融合 培育高端高效的智能经济 构筑安全高效的智能化基础设施体系 建设安全便捷的智能社会 前瞻布局新一代人工智能重大科技项目
实现路径	科技与治理 投资关键技术，建立创新生态和战略合作，推进治理现代化 军事与外交 强化联盟，提升军队建设和军事投入，主导集成化战争，建立新国际论坛 制造与人才 争夺制造业领先，构建数字基础设施，建立融资机制，制定培训方案培养高端劳动力	技术与人才 构建新一代AI基础和关键技术体系，布局创新平台，培养AI高端人才 沟通协作 建立科研、高校、企业、军工单位常态化沟通机制 产业与创新 发展AI新兴行业，推动产业智能化，支持智能企业，打造创新高地 基础设施 优化网络，建设高效计算和分布式能源互联网 社会与治理 推进社会治理智能化，利用AI提升公共安全，促进社会共享互信 项目群建设 围绕重大科技项目支持“1+N”人工智能项目群布局
资金支持	2019财年，美国DARPA计划五年内增加20亿美元投资“AI Next”计划 2020年，美国向国家标准与技术研究院（NITDR）提供6.88亿美元，用于量子计算、人工智能和微电子等尖端研究	- 补贴：通过中国人工智能产业联盟等平台举办竞赛分配现金补贴 - 政府引导基金：公私合营投资基金支持AI和新兴技术。截至2023年，设立2086个引导基金，目标募资约12.19万亿元人民币，已认缴约7.13万亿元人民币

中国AI发展版图：北京以科研和政策为核心，上海依赖科技创新和市场，深圳靠前沿科技企业和研发资源，杭州则由电商巨头市场需求引领

	AI企业	AI人才储备（实验室及高校）	AI政策	AI投资	AI核心动力
北京	拥有大模型厂商及高等院校122家，居全国首位 百度、字节跳动、智谱华章、百川智能等通用大模型厂商，第四范式、云知声等行业大模型厂商	清华大学AI研究所-深度学习和自然语言处理 北京大学AI实验室-计算机视觉和智能系统 智源研究院-基础研究和应用创新	北京市智能科技发展规划 - 支持智能科技产业发展 建立人工智能投资联盟机制，协同社会资本和国家级基金提供持续保障	风险投资（VC）活跃，如红杉资本、经纬中国、高榕资本；政府和科研机构资金支持	强大的科研基础和政策支持
上海	科大讯飞-语音识别和自然语言处理技术领军企业 深蓝科技- AI芯片和计算平台，提供深度学习加速解决方案	上海交通大学AI实验室 - 深度学习和智能系统 复旦大学AI研究中心 - 计算机视觉和数据分析 同济大学AI研究中心-计算机视觉和智能系统	上海市智能技术应用示范区建设方案 - 促进AI技术应用 上海市智能制造发展规划 - 支持AI在制造业中的应用 上海市人工智能发展行动计划 - 鼓励AI企业和技术研发	资金丰富，包括赛富投资基金和软银中国资本，风险投资如华兴资本和高盛；政府资金支持	科技创新和市场需求
深圳	腾讯AI实验室 优图实验室 华为诺亚方舟实验室 大疆-全球领先AI技术优化无人机图像识别	深圳大学AI实验室 - 专注于智能控制和计算机视觉； 南方科技大学AI研究所 - 机器学习和机器人技术 深圳国际研究院	粤港澳大湾区科技创新规划 - 支持AI技术应用 深圳市科技创新委员会：提供专项资金用于支持AI技术的研发和应用，推动科技创新	腾讯、华为等企业自有资金投资，政府提供前海基金、专项资金支持，另外建设科技园区和高新创新平台	前沿科技企业和研发资源
杭州	阿里巴巴达摩研究院-专注AI、机器学习、大数据	浙江大学AI实验室 - 机器学习和大数据分析 杭州电子科技大学AI研究中心 - 专注于智能系统和机器人技术	浙江省人工智能产业发展规划 - 促进AI产业发展，鼓励企业和研究机构合作； 杭州市智能制造专项资金 - 支持智能制造应用	阿里巴巴的自有资金和风险投资，如真格基金、源码资本	电商巨头和市场需求

美国AI发展版图：波士顿以学术研究和技术创新为核心，旧金山依靠创新技术和资本投入，西雅图由企业驱动和市场需求引领，纽约则结合技术创新与应用场景需求

	AI企业	AI人才储备（实验室及高校）	AI政策	AI投资	AI核心动力
波士顿	50-100家 Nuro（自动驾驶技术） Cerebras Systems（加速深度学习计算硬件） Neurala（AI视觉检测方案）	麻省理工学院传媒实验室 - 前沿AI研究 哈佛AI研究中心 - 聚焦于机器学习和数据科学 东北大学Khoury计算机与信息科学学院及 AI实验室 - 机器人技术	马萨诸塞州科技创新计划 - 提供资金支持和税收优惠，促进科技和AI创业； AI伦理委员会 - 设立AI伦理标准和指南。	风险投资（VC）活跃，政府和州资金支持丰富，例如Massachusetts Technology Collaborative。	强大的学术研究基础与技术创新环境
旧金山	200-300家，AI创业最活跃的地区 OpenAI、英伟达、UiPath（机器人技术）	斯坦福大学 AI实验室 - 提供前沿AI研究和应用 加州大学伯克利分校AI实验室 - 深度学习和计算机视觉研究	旧金山湾区创新激励计划 - 为AI和高科技企业提供财政激励和政策支持； 技术创新促进法案 - 支持AI技术研发与应用，制定相关法规促进产业发展。	风险投资极为活跃，如Andreessen Horowitz和Benchmark Capital；科技公司投资。	创新技术、资本投入和开放的创业生态环境
西雅图	30-60家，以计算机视觉、语音识别为主 亚马逊AI、微软、Xnor.ai、Algorithmia（边缘计算）	华盛顿大学AI研究所 - 进行深度学习和自然语言处理研究 艾伦人工智能研究所 - 重点关注机器学习和人工智能	西雅图科技促进法案 - 为AI初创企业提供支持和资源，推动技术发展； 微软AI伦理政策 - 确保AI技术的安全性和公平性； 西雅图创新支持计划 - 促进技术发展和创新。	亚马逊和微软投资丰富，还有如Madrona Venture Group等风险投资支持。	企业驱动、市场需求和技术创新
纽约	50-80家，以金融科技，解决方案为主 IBM、Clarifai（视觉AI）、K Health（医疗AI）	哥伦比亚大学AI实验室 - 专注于AI和机器学习研究 纽约大学AI实验室 - 提供前沿AI技术和应用研究	纽约市科技支持计划 - 促进科技和AI创业的政策，提供资源和资金支持。	资金来源多样，包括企业投资、风险投资如Union Square Ventures和政府资助。	技术创新与应用场景需求结合

人才储备：中国培养大量AI人才，但顶尖AI人才落后于美国，仍存在不小的缺口

中国仍在大量储备AI人才，顶尖AI人才存在缺口

新增“人工智能”专业院校数量



数据来源：教育部

2023年全球最具影响力人工智能学者



数据来源：清华大学AMiner团队发布的《2023年全球最具影响力人工智能学者》

- 中国新增“人工智能”专业高校数量逐年增加，但增速放缓
- 截至2023年底，中国拥有1014名顶尖AI研究人员和9183家AI相关公司
- 2030年中国AI人才缺口多达400万人，对AI专业人员的需求将增至600万

- 2023年全球AI学者中最具影响力的2000人次中，美国入选人数最多，共有1079人次，占全球54.0%
- 中国，共有280人次入选，美国学者数量是中国近4倍

AI博士生更愿意在工业界中求职

美国与加拿大AI博士毕业生去向



- 2022年，进入学术界与进入工业界的AI博士生差距扩大
- 过去5年，进入政府职位的AI博士比例一直保持在相对较低的水平

数据来源：2023年CRA Taulbee Survey

研发水平：当前中国在AI论文总量和发明专利方面领先美国，质量上仅略次于美国，但并不存在中国论文以量取胜，质量落后”的现象

中美在不同来源及领域AI出版物的数量

2017年至2022年AI研究论文数量



教育学术界贡献最大，中国出版数量超越美国，其次是工业和政府



AI论文数量

国家	数量	占比
中国	24.3万篇	占比25%
美国	17.1万篇	占比18%

计算机视觉贡献最大，发表文章40.5万篇，占比32%



AI论文质量（高频率引用）

国家	数量	占比
中国	2.98万篇	占比31%
美国	3.35万篇	占比35%

企业/组织Gen AI相关专利数量

单位：个



国家Gen AI相关专利数量

单位：个



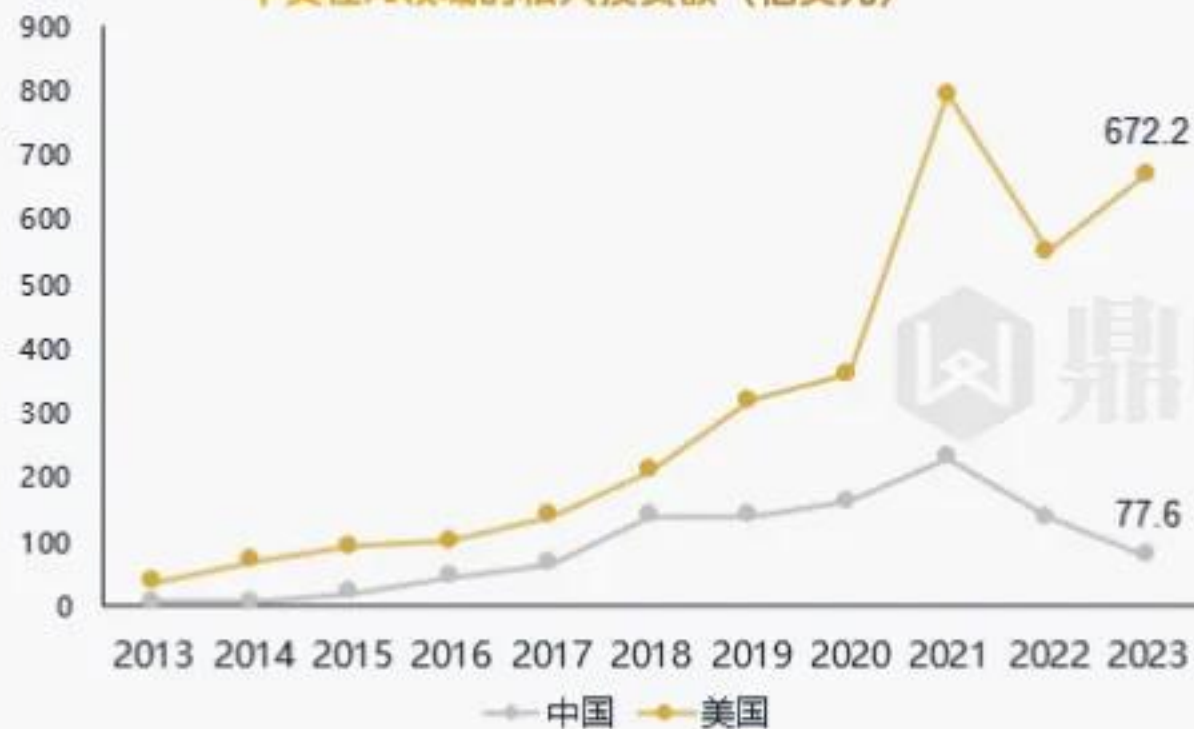
- 自2017年推出大语言模型背后的深度神经网络架构以来，Gen AI专利数量增长八倍，仅2023年公布的Gen AI专利就超过了全部总数的25%
- 2014至2023年间，中国与生成式人工智能（Gen AI）相关的发明申请量达5.4万项，其中Gen AI发明专利超过3.8万项，是排名第二的美国的六倍
- 在Gen AI相关领域，腾讯以2074项发明位于世界第一，与平安保险、百度发表的专利数量总计超过总数的70%

公众号·鼎帷咨询

公司投资：美国AI领域私人投资规模远超中国，但中美初创企业融资增速均放缓

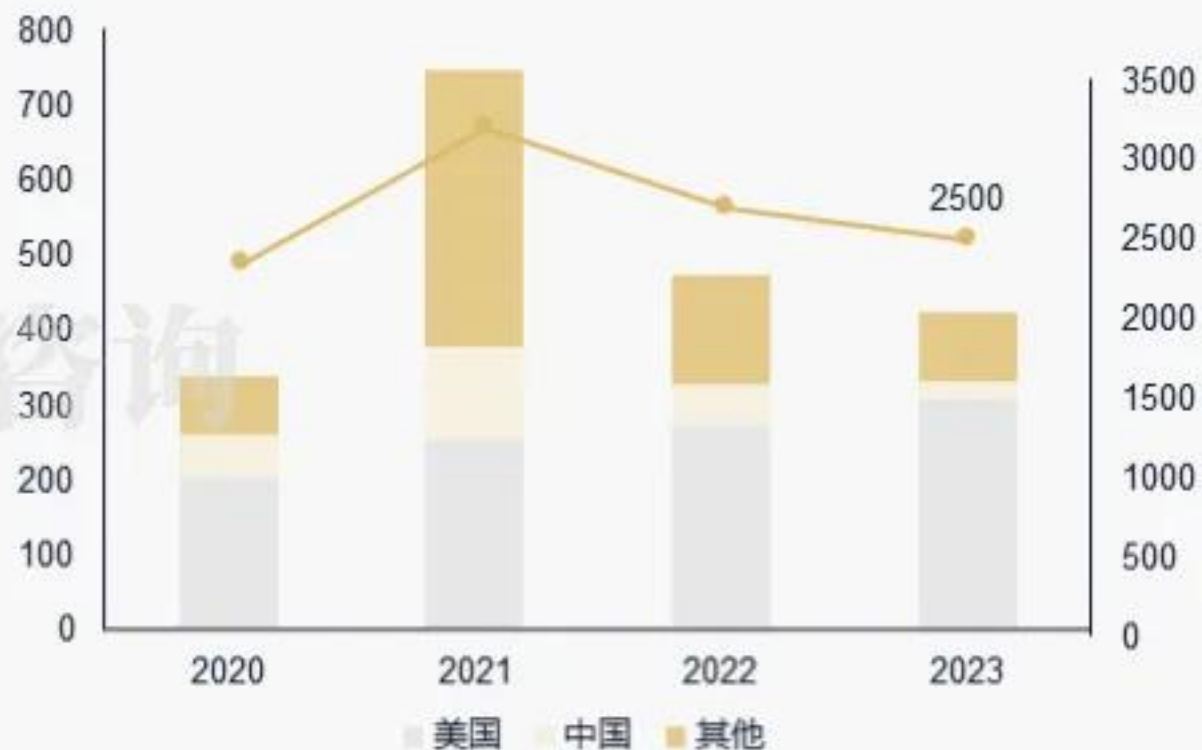
中国在私人投资方面存在巨大资金缺口

中美在AI领域的私人投资额（亿美元）



- 相较于美国，中国在AI领域的私人投资额数量和增速上均存在差异，且这种差异仍然在扩大
- 美国在AI领域的私人投资额仍呈上升趋势，中国自2021年后开始呈下降趋势

中美两国均对初创企业的融资增速放缓



- 全球AI领域的融资交易量约为2500笔，创下了自2017年以来行业新低
- 2023年，中美AI领域投融资数量与投资额均下降38%

03

产业链发展

生成式AI的价值链是基于AI基础设施（计算、专家和数据）作为关键输入，以基础模型和调整模型的开发和服务，以私有化部署、开闭源模型及其应用软件的下游基础模型服务



人工智能产业图谱：人工智能产业包含基础层、技术层、大模型与工具层、应用层



算力基础：算力芯片主要分为以CPU、GPU、DSP为主的通用芯片和以FPGA、ASIC、DPU等为主的专用芯片

		产品	特点	灵活性	性能			
芯片	通用芯片	X86 CPU处理芯片	可以完成多样化且灵活的算力任务，但功耗更高			简称	英文全称	中文全称
		GPU 图形处理器	由大量小处理单元组成，专为并行计算而设计，能够快速完成大数据和图像处理任务，例如在AI训练中加速复杂运算，是高性能计算的关键工具			APU	Accelerated Processing Unit	加速处理器
		DSP 数字信号处理器	专为高速高效的实施数字信号处理任务而设计，具有哈佛结构、硬件乘法器、流水线操作和专用指令集等特性			DPU	Data Processing Unit	数据处理器
	专用芯片	FPGA 可编程集成电路	可以通过硬件编程来改变内部芯片的逻辑结构，但软件是深度定制的，执行专门任务，介于通用芯片和ASIC之间			NPU	Neural Network Processing Unit	神经网络处理器
		ASIC 专用集成电路	为专业用途而定制芯片，绝大部分软件算法都固化于硅片，能完成特定的运算功能，作用单一，不过能耗很低			TPU	Tensor Processing Unit	张量处理器
		APU DPU NPU TPU	近年开始流行的专用芯片					

美国对我国采取出口管制、投资管制和地缘管制，导致我国底层技术获取受限，倒逼我国自主研发

出口管制	
2022年8月	对中国实施半导体和设备出口管制
2023年10月	限制AI芯片及14纳米以下设备出口
2024年3月	限制NVIDIA、AMD等公司AI芯片出口
2024年5月	ENFORCE法案管制AI模型出口

投资管制	
2022年8月	对外投资 《芯片与科学法案》使半导体国际企业不得在中国扩大半导体生产能力
2022年9月	对内投资 严格审查外国对AI技术的投资
2023年8月	限制美国对中国AI和半导体的投资

地缘管制	
2020年9月	启动AI国防伙伴关系，旨在保持对中国和俄罗斯的领先地位
2022年4月	与韩国、日本、中国台湾地区建立“芯片四方联盟”，管制对中国的出口
2023年1月	与日本、荷兰合作限制对华半导体设备出口

推动中国AI自主研发	
投资支持	中国企业自主投资，政府加大政策支持
自主产业链	加快构建涵盖芯片、算法以及数据服务等关键环节的自主可控AI产业链
减少依赖	降低对外部技术的依赖，实现长期稳定和安全

AI底层技术获取受限	
英伟达GPU	对中国销售的低端芯片性能降低
DUV光刻技术	中国未掌握关键光刻技术，美国施压荷兰ASML限制中国高端芯片制造
台积电制造商	美国阻断摩尔线程和壁仞科技与台湾半导体公司及其他制造商的合作

科研人才受阻	
人才流失	美国对华学术交流干预导致中国科研人才流失
成果受阻	科研成果推进困难，阻碍技术交流与产业合作

算力基础：计算能力衡量以FLOPS为主要指标，不同算力平台差异显著，目前全球总规模达1369EFLOPS，我国规模达450EFLOPS，占据全球三分之一

计算能力衡量指标	
计量单位	全称
MIPS	每秒百万条指令
DMIPS	每秒执行百万条指令
OPS	每秒操作数
FLOPS (以数量级划分，浮点数有FP16、FP32和FP64规格)	MFLOPS 每秒浮点运算数
	GFLOPS megaFLOPS百万次浮点运算
	TFLOPS gigaFLOPS千兆次浮点运算
	PFLOPS teraFLOPS万亿次浮点运算
	EFLOPS petaFLOPS千万亿次浮点运算
	ZFLOPS exaFLOPS亿亿次浮点运算
	MFLOPS zettaFLOPS泽塔浮点运算
Hash/s哈希/秒	每秒哈希值

全球算力总规模		
2023	2025E	2030E
1369EFlops	3ZFlops	20ZFlops
中国:450EFlops 从2018年到2023年5年间，我国智能算力占比由11%增长至65%，过去5年平均增速117%，高于全球增速。		其他算力预计规模 FP16:增长500倍至105ZFlops FP32:增长10倍至3.3ZFlops

不同算力载体之间的算力差异对比	
算力平台	算力大小
ENIAC	3×10^2 FLOPS
Intel8086CPU	7.1×10^5 FLOPS
IntelPentium4HT3.6Ghz	7×10^9 FLOPS
IntelGeforceGTX1080Ti	1.08×10^{13} FLOPS
QualcommSnapdragon888processor	2.6×10^{13} FLOPS
MilkyWayOne	2.566×10^{15} FLOPS
SunwayTaihuLight(Supercomputer)	1.25×10^{17} FLOPS

算力基础：综合评估体系下，中美位于全球算力水平第一梯队，日德新英法等国位于第二梯队

全球算力指数评估体系

计算能力

通用计算能力:服务器支出规模及占比

科学计算能力:全球Top500超级计算机数量及排名

AI计算能力: AI服务器支出规模

终端计算能力:智能手机及PC支出规模及保有量

边缘计算力:边缘计算硬件支出规模

应用水平

大数据:相关软件、硬件、服务整体支出占比

人工智能:相关软件、硬件、服务整体支出占比

物联网:相关软件、硬件、服务整体支出占比

区块链:相关软件、硬件、服务整体支出占比

机器人:相关软件、硬件、服务整体支出占比

计算效率

云计算渗透率:云计算支出规模占比

CPU利用率:服务器CPU平均使用率

内存利用率:服务器内存平均使用率

存储利用率:服务器存储设备平均使用率

新技术使用率:采用SSD,加速技术的服务器等平均使用率

基础设施支持

数据中心软件和服务:致数据中心软件和服务支持规模占比

数据中心规模:致数据中心数量

数据中心能效:数据中心平均PUE

网络基础设施:每年度企业网络设备出货量,每年运营商网络支出占比

存储基础设施:每年度存储出货量(TB)和支出占比

算力评估结果

依据：算力指数的分布，算力指数的提升对经济增长的影响

第一梯队：领跑者

美国、中国

评分在60分以上

第二梯队：追赶者

日本、德国、新加坡、英国、法国、印度、加拿大、韩国、爱尔兰和澳大利亚

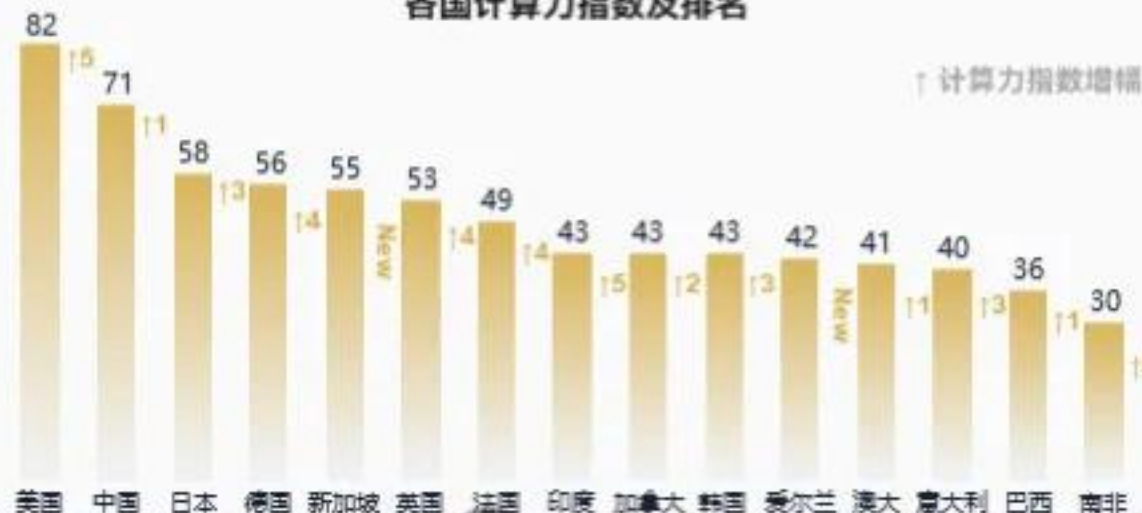
评分在40—60分之间

第三梯队：起步者

意大利、巴西和南非

评分在40分以下

各国算力指数及排名



算力基础：数据中心主要分为五大类型，全球市场规模持续增长，美国容量占据全球第一，中国位列第二

三大基础设施

网络基础设施	存储基础设施	计算资源
连接服务器、数据中心服务、存储系统以及用户位置	用于存储业务关键数据，是数据中心的重要资源	提供处理能力、内存、存储和网络连接，是数据中心的运算核心

五大数据类型

企业数据中心	多租户/主机托管数据中心	超大规模数据中心	边缘/微型数据中心	集装箱/模块化数据中心
专为单个企业定制，满足其独特的网络需求和业务流程	提供异地托管硬件服务，适合无空间或管理团队的企业	支持大规模IT基础设施，主要由云计算巨头如亚马逊、微软和Google拥有	靠近用户，提供低延迟的实时数据处理和分析	即插即用的数据中心，包含完整的计算和支持设备，灵活部署

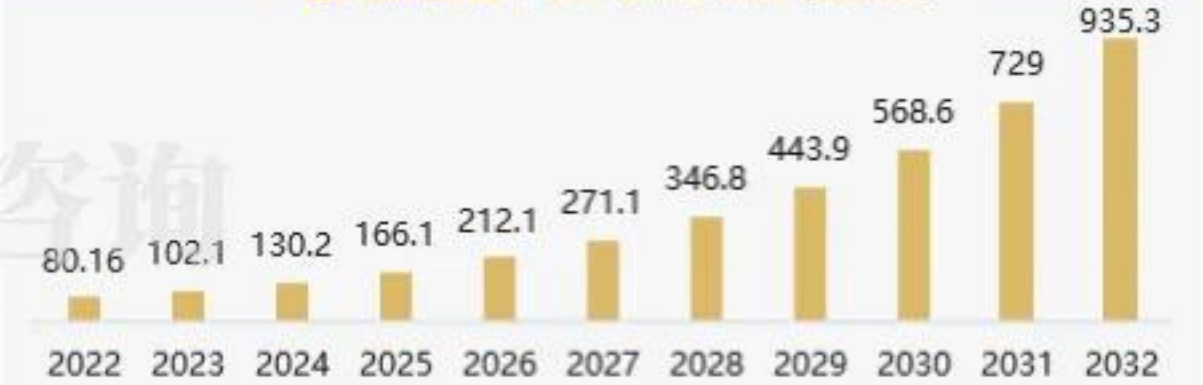
数据中心核心组件

路由器	交换机	防火墙	服务器
存储系统		应用程序交付控制器	

全球超大规模数据中心现状

截至2023年底	未来预测
全球活跃的超大规模设施增至992个，现已超过1,000个	每年将新增120-130个超大规模数据中心，约440个正在筹备中

超大规模数据中心市场规模预测（亿美元）



全球容量占比



主导公司

主要公司: Amazon、Microsoft、Google，占有超大规模数据中心容量的60%

其他公司: Meta、阿里巴巴、腾讯、苹果、字节跳动

算力基础：智算中心面向产业发展、科学研究、公共服务等，以政府、企业、科研机构为投资主体，为AI研究及应用提供强大计算资源，目前很多科技企业均在布局人工智能与算力

全球智算中心

现状	全球的人工智能计算中心正在快速发展，这些中心通常由政府、企业和学术机构共同建立		
目标	为 AI 的研究和应用提供强大的计算资源和专业知识		
服务类型	产业发展	科学研究	公共服务
服务内容	数据服务 为服务购买方提供数据存储、分析、查询等基础数据服务	算力服务 提供高性能计算资源，如芯片、服务器等，用于复杂计算任务	
	算法服务 提供算法模型与开发能力，帮助服务方优化业务流程和决策	生态服务 提供智能计算生态支持，促进各方协作和资源共享，提升计算中心效率	

智算中心建设模式

独立投资	政府投资	由政府或园区管委会直接投资建设，所有权归政府，资金来源为财政资金和专项债券，旨在促进产业发展
	企业投资	主要由大型科技企业投资建设，服务于特定产业和应用场景，企业可单独建设或与专业建设企业合作
	高校或科研机构	以智能计算平台为主，成本较低，主要用于科研和教育，提供算力支持助力基础和前沿研究
第三方出资		■ 主要由国有控股企业出资建设，政府全程把控并对接资源。 ■ 建成后中心归第三方公司，政府通过项目补贴或置换，投资方式包括新设或授权现有国有控股公司。
特殊项目公司 (SPV)		■ 政府与企业合资成立项目公司，独立负责智算中心的建设和运营，为政府和企业提供服务。 ■ 此模式通过专业团队高效管理，降低成本，灵活应对设计和运营需求。

智算中心技术供应商图谱

应用层	智慧城市	电力能源	智能交通	政府服务	智能制造	医疗健康	金融科技	服务和运营	
技术层	人工智能	百度 阿里云 百川智能 软通智慧 商汤科技 腾讯 科大讯飞 云从科技			大数据	第四范式 星环科技 东方国信 明略科技 浩鲸科技		中国移动 中国联通 中国电信 华为 商汤科技 阿里云 鸿博股份	
基础设施	云平台	华为 中国电子云 阿里云 百度 浪潮云 天翼云 优刻得 青云科技							
	AI服务器	华为 浪潮 新华三 宁畅 安擎 曙光 联想 戴尔							
	智能芯片	英伟达 寒武纪 华为 壁仞科技 燧原科技 海光 天数智芯 云天励飞				海康威视 科大讯飞			

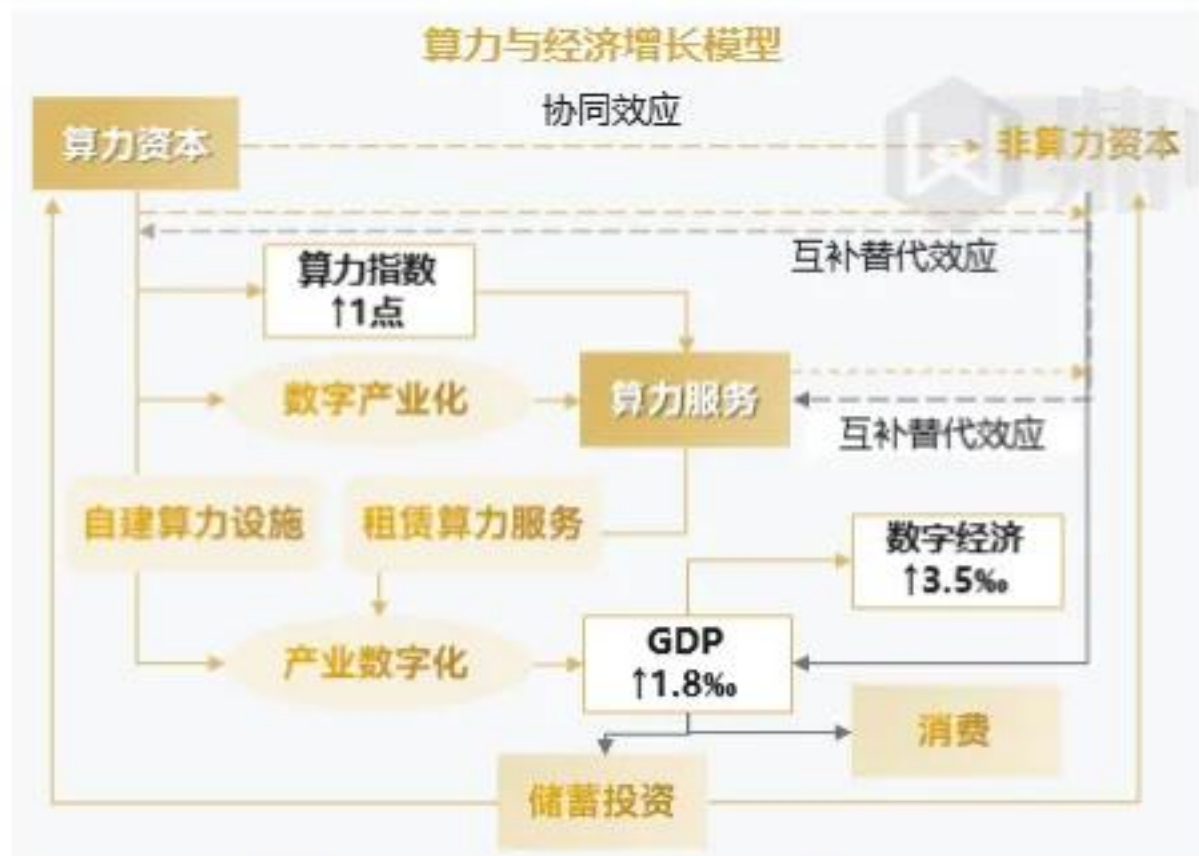
算力基础：我国正分层次分类别加速布局智算中心，美国位居智算中心发展前列，欧盟聚焦数字主权超算及量子计算



算力基础：算力投资极大促进经济及各产业增长，IT每投入以1美元，就会产生29美元的GDP和15美元的数字经济

算力决定了数字经济发展的速度和社会智能化发展的高度

- 根据IDC、浪潮信息、清华大学全球产业研究院联合发布的数据显示，计算力指数平均每提高1点，数字经济和GDP将分别增长3.5‰和1.8‰



三部门形成从生产到消费的经济闭环

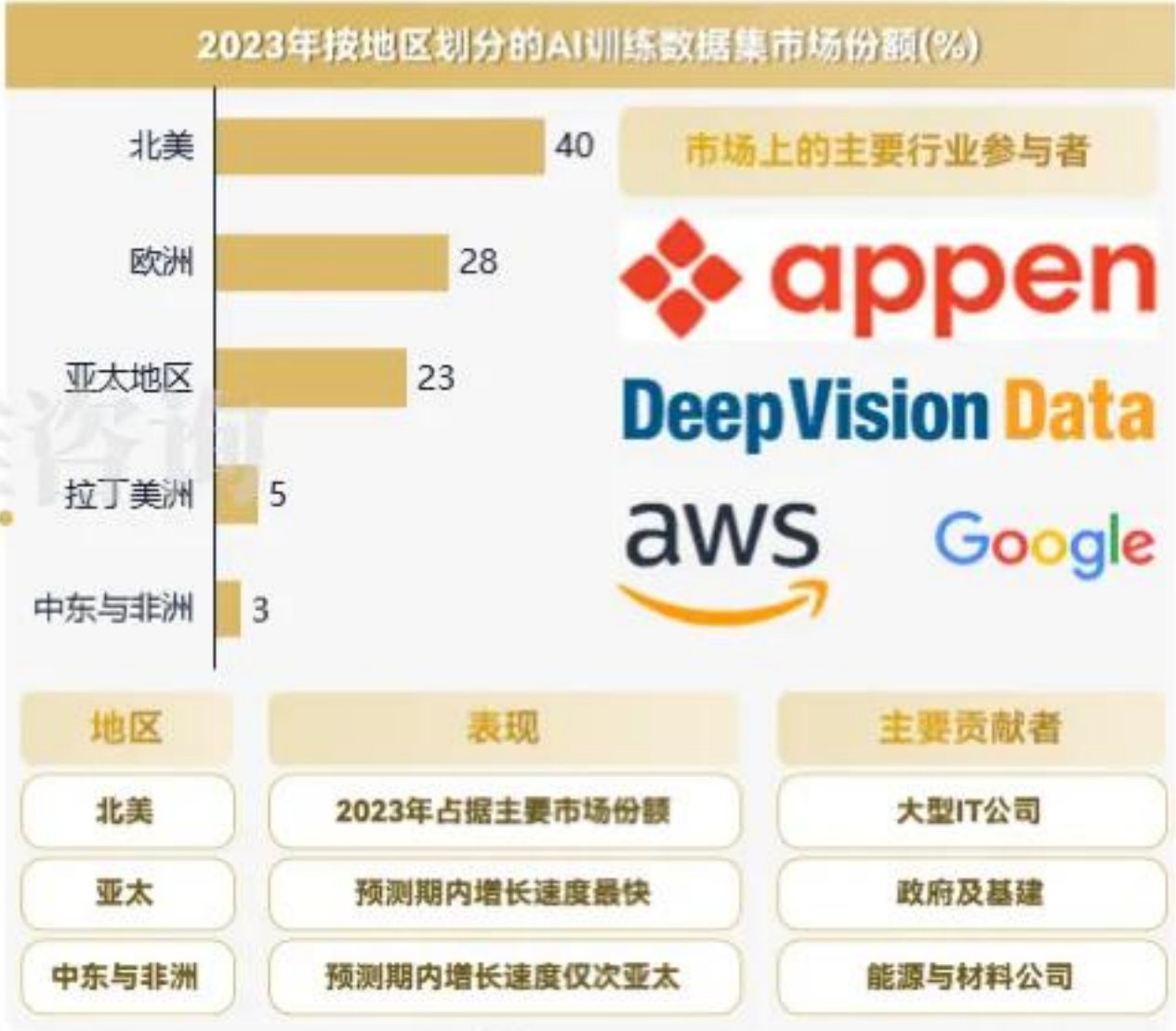
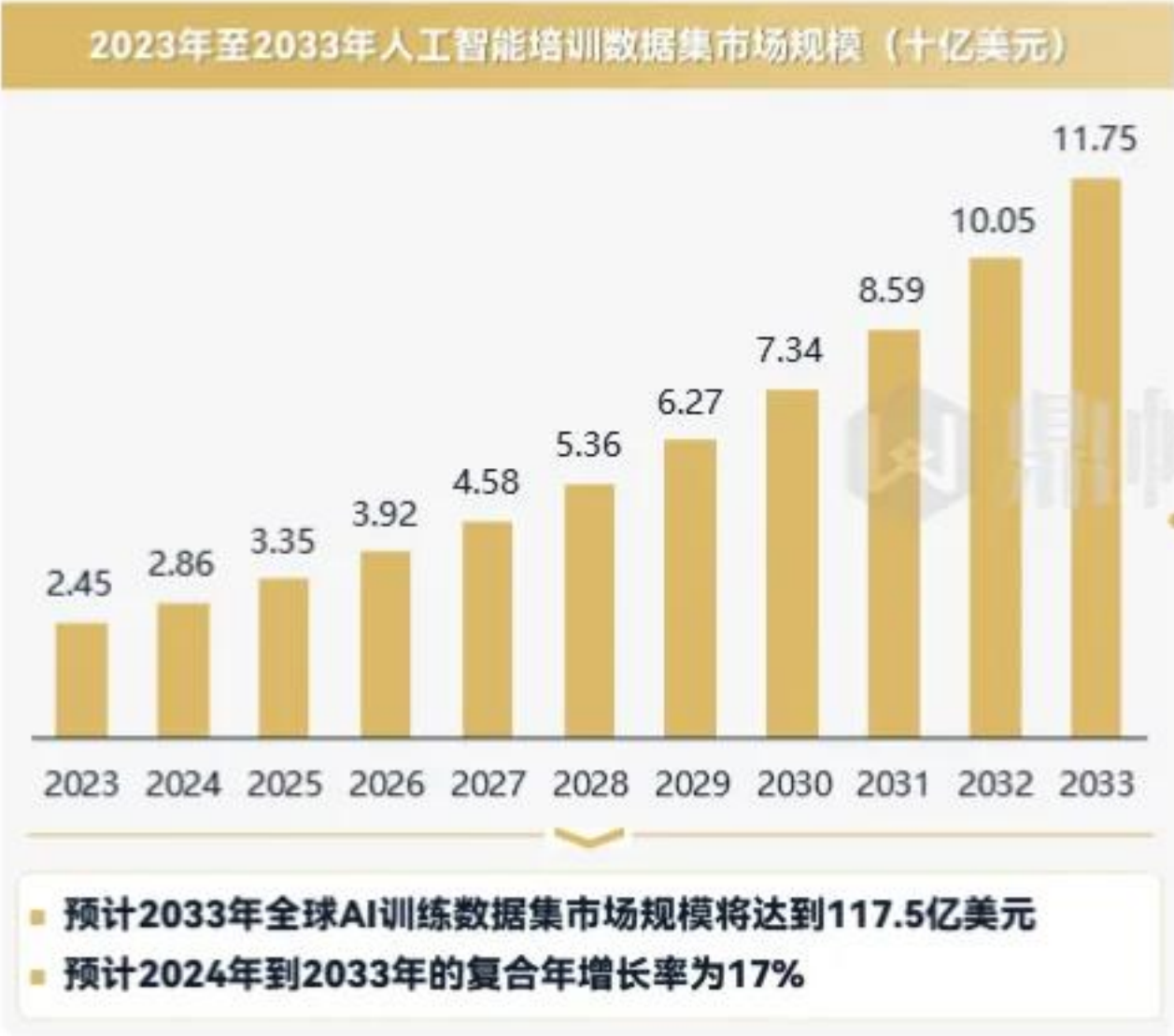
- 中间品厂商投资算力资本生产算力服务，供最终品厂商使用。
- 最终品厂商通过投资算力和非算力资本生产商品，商品销售给居民或用于再投资。
- 家庭部门提供生产要素，获得收入后消费或投资。



数据基础：从免费数据集进入市场通过本地及云端部署方式触达最终用户



数据基础：全球AI训练数据集市场规模不断扩大，其中北美占据主要份额，以亚马逊为首的公司提供AI数据解决方案



公众号 · 鼎惟咨询

算法基础：传统架构以神经网络架构为主，目前发展出生成对抗网络、胶囊网络、迁移学习、自编码器等新型架构，从传统语言向图像、音频等数据类型转换

	类型	领域	特征	优势	用途
传统架构	卷积神经网络 (CNN)	图像、语音	善于处理具有网格状拓扑结构的数据 (如图像)，通过卷积操作提取局部特征，具有平移不变性	高效提取空间特征，擅长处理图像等具有网格结构的数据	图像识别、物体检测、视频分析、医学影像处理
	循环神经网络 (RNN) 长短时记忆网络 (LSTM)	自然语言、语音	适用于序列数据处理，具有记忆历史信息的能力，通过循环机制保存序列间的依赖关系	能够捕获序列中的时间依赖关系，适合处理序列数据	自然语言处理、时间序列预测、语音识别、机器翻译
	Transformer BERT模型	自然语言、 图像、语音	专为并行计算设计，利用自注意力机制处理序列数据，无需依赖序列的顺序	擅长处理长距离依赖，解决了RNN的长期记忆问题	机器翻译、文本生成、语言模型 (如GPT)、图像处理
	Wide & Deep 深度学习推荐模型DLRM	推荐系统	结合用户历史行为和产品特征，通过深度神经网络捕获用户兴趣的多样性与复杂性	能够处理大规模用户和产品数据，提供高度个性化的推荐	电商推荐、个性化推荐、视频推荐、广告推荐
新型架构	胶囊网络	图像	捕捉对象及其各部分的关系，能够保留物体的位置信息和方向信息，增强对平移和旋转变换的鲁棒性	保留空间信息，能够识别物体的姿态和结构关系	图像识别、3D目标检测、姿态估计、物体检测
	生成对抗网络 (GAN)	图像、文本、 音频	通过生成器和判别器相互对抗训练，能够生成与真实数据相似的合成数据，提升了模型的生成能力	生成逼真数据，图像生成能力强	图像生成、图像修复、超分辨率重建、文本到图像生成
	迁移学习	图像、自然语言、 语音	通过利用预训练模型的特征减少训练时间，提高在数据不足情况下的模型效果	在数据不足的情况快速构建高性能模型，减少计算资源需求	图像分类、情感分析、语音识别、小样本学习
	自编码器	图像、数据降维	通过编码器-解码器结构学习数据的低维表示，能够从输入数据中提取重要特征	自动发现数据中的特征并进行降维，擅长无监督学习任务	数据降维、图像去噪、异常检测、特征学习

大模型层：大模型历经了从垂直单模态模型向通用单模态模型到多模态大模型的转变

	垂直单模态模型	通用单模态模型	多模态大模型
定义	针对特定任务（如图像分类、语音识别、文本生成等）设计的机器学习或深度学习模型，且仅处理一种类型的数据	只处理一种类型的数据，但可应用于该模态下的多个不同任务（如图像模态下的分类、检测、分割等）	从不同模态（如图像、文本、音频等）的数据中提取和整合信息，以完成复杂任务
特点	■ 专用性：模型为特定任务设计，在该任务上具有较高的性能 ■ 数据类型单一：只处理一种数据模态	■ 通用性：能够应用于该模态下的多种任务 ■ 灵活性：具有更广泛的适用性 ■ 扩展性：可通过微调或增量学习扩展到新任务或新数据集	■ 数据融合：更好地理解 and 生成复杂信息 ■ 广泛应用：在如视觉问答、图文生成、视频分析等需要融合多种信息源的任务中表现出色 ■ 复杂性：在训练和推理过程更具挑战性
典型模型	■ ResNet：用于图像分类的卷积神经网络 ■ BERT：用于自然语言处理任务的预训练模型 ■ DeepSpeech：用于语音识别的深度学习模型	■ GPT-3.5：可用于文本生成、翻译、问答等多种自然语言处理任务 ■ Vision Transformer (ViT)：大模型首次进入视觉领域 ■ Midjourney：人工智能绘图工具	■ PaLM-E：全球首个多模态具身视觉语音模型 ■ GPT-4：用于图文匹配、图像生成等任务。 ■ DALL-E：根据文本描述生成相应图像 ■ CLIP：多模态经典模型，用于图文匹配、图像生成等任务
代表产品			

公众号·鼎帷咨询

大模型层：依托于Transformer架构产生的DALLE、ChatGPT、文心一言等代表产品，也随着算力和技术的升级历经了从单模态大模型向多模态大模型转变



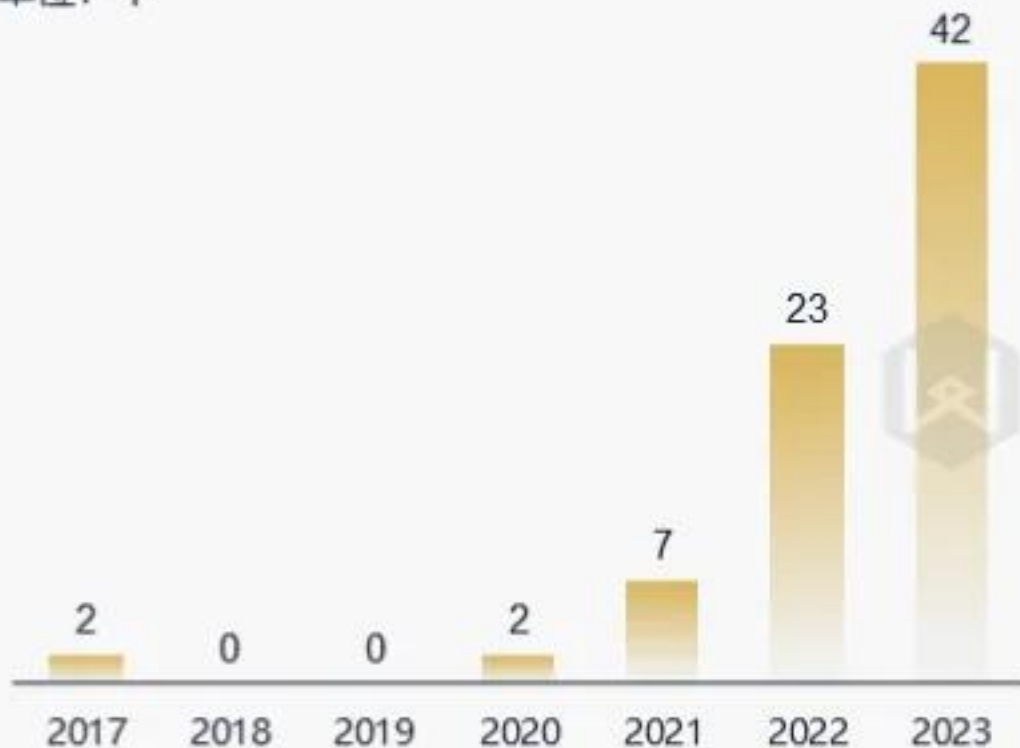
大模型层：构建大模型需要通过训练（识别和生成规律）并推理（输出预测结果）两个步骤，其中训练步骤经历自我监督学习、监督学习、强化学习三个主要阶段



大模型层：大模型算力普遍达到1亿 petaFLOPS，随着训练计算的增加，算力需求持续增长

大于100万petaFLOP的模型数量

单位：个



- 目前大多数大模型的算力处于1亿petaFLOP
- 算力超过 100万petaFLOP的模型数量从2020 年开始急剧上升
- 算力的快速发展与机器学习研发投资和硬件性能的指数级增长是一致的
- 如果训练计算继续每年增加 4 倍，顶级模型将在2024年超10亿petaFLOP

随着训练计算的增加，AI系统变得更加强大



公众号·鼎帷咨询

大模型层：高昂的训练、推理及储存成本成为是阻碍大模型发展的关键因素

训练

推理

储存

成本量级

- 训练成本非常高
- 大语言模型单次训练成本从**140万美元至1200万美元**

- 随着模型的复杂性增加和推理频率提升，成本会显著增加
- GPT-4 Turbo日推理成本为**2.94亿美元**

- 通常较为稳定
- 随着模型大小的增长，储存成本会显著增加

成本来源

- 强大的**计算资源**消耗，特别是GPU的使用
- 大量的**电力消耗**

- 使用**计算资源**，尤其是GPU和其他加速器
- 需要**优化模型和硬件的配合**，以减少延迟和提高性能

- 需要大量的**存储空间**，特别是在使用高性能存储设备或云存储服务时
- 包括设备购买、维护和云存储的使用费

成本估算

AI模型训练及推理成本估算（亿美元）



- 1PB的可用存储空间需要60多万元，加上安全冗余和网络设备、安全设备，成本会接近**70万元/PB**
- GPT-4以及Gemini 1.0 的原始数据量为50~100PB，则储存成本预计为7000万

大模型层：目前大模型以开源为主，然而闭源模型依靠其技术壁垒和商业资本优势，表现普遍优于开源模型

	开源模型	闭源模型	基础模型的访问方式
定义	源代码对公众开放，允许用户查看、使用、修改和分发的软件	源代码不对公众开放的软件	2023年，开源模型仍占据主导地位，访问受限模型数量逐步增加
形式	社群或组织开发和维护，可免费使用	公司或组织私下开发、销售和使用，通常需要付费购买或订阅，受到知识产权保护	
优势	- 有高度透明性及灵活性，广发的社区参与度，形成了强大的生态系统和产业链 - 使用开源模型可降低成本，提高可定制性和可移植性	- 由专业团队负责开发和维护，具有较高的稳定性和性能。 - 可提供更全面的技术支持和服务 - 受知识产权保护，防止源代码被恶意使用或篡改	



年份	闭源	访问受限	开源
2019	4	0	0
2020	2	0	0
2021	10	9	0
2022	28	12	32
2023	28	23	98

闭源模型无论在技术水平还是融资情况均优于开源模型

基准测试	任务类别	开源	闭源
GPQA	一般推理	29.1%	41%
GSM8K	数学推理	93.3%	97%
HELM	一般语言	82%	96%
MMLU	一般语言	70.6%	90.4%
SWE-Bench	编码	3.97%	4.8%
Chatbot Arena Leaderboard	一般语言	1149	1252



应用层：中国AI发展方向集中在AI技术的创新和应用，涵盖自动驾驶、AI芯片、边缘计算、电商智能化、金融科技、社交平台智能化、医疗健康、计算机视觉和智能安防等多个领域

	AI研发及投资方向	代表AI技术	代表AI产品	落地应用场景
百度	百度研究院、百度大脑AI实验室、Apollo实验室 自动驾驶、自然语言处理、智能云计算、智能设备	ERNIE语言模型（先进的语言理解模型，增强了多任务学习能力） Deep Speech（语音识别技术）	文心大模型 Apollo（自动驾驶平台） 百度大脑（AI开放平台）	自动驾驶：提供智能驾驶解决方案 语音识别和自然语言处理：应用于搜索引擎、语音助手 AI云服务：用于企业数据分析和智能应用
华为	诺亚方舟实验室、昇腾实验室、华为云AI实验室 AI芯片、云计算和边缘计算、智能硬件	昇腾芯片（自研AI处理器，用于深度学习加速） MindSpore（自研深度学习框架）	昇腾AI芯片 Atlas AI计算平台 华为云EI NLP 大模型“鹏程·盘古”	智能制造：利用AI优化生产流程 智能家居：通过AI提升设备智能 企业云服务：提供AI计算和数据分析服务
阿里巴巴	阿里巴巴达摩院 城市大脑实验室 电商智能化 云计算、金融科技	城市大脑（集成城市管理和数据分析的自研技术） AliMe（智能客服技术，利用NLP技术提升客户服务体验）	通义千问 阿里云AI服务 ET大脑（智能城市解决方案）	电商推荐系统：优化商品推荐和搜索 智慧城市：应用于交通、安防、环境监测 金融服务：用于风险控制和客户服务
腾讯	腾讯AI Lab、腾讯云AI实验室、腾讯优图实验室 社交平台智能化、云计算、医疗健康	腾讯优图（自研图像识别技术，应用于人脸识别和图像处理） 腾讯AI Lab的深度学习平台	腾讯云AI（包括语音识别、图像识别服务） 微信智能助手 手语数智人“聆语”	社交应用：提供智能对话和内容推荐 医疗健康：用于医学影像分析和健康管理 AI开放平台：支持企业应用AI技术
商场	SenseTime AI+实验室 计算机视觉、智能安防、自动驾驶	SenseFace（人脸识别技术） SenseDrive（自动驾驶方案） SenseID（多模态身份验证技术）	SenseTime Smart Surveillance（智能视频监控） SenseTime Smart Retail（智能零售解决方案）	安防监控：提供智能视频监控和人脸识别 智能零售：应用于顾客行为分析和店内优化 金融：用于身份验证和风险管理

公众号 · 鼎帷咨询

应用层：国外AI战略集中在通用性研究，以及推荐系统、GPU优化AI计算、社交媒体与AI结合等多个方向，旨在推动技术创新，拓展应用场景，提升企业竞争力和市场影响力

	AI研发及投资方向	代表AI技术	代表AI产品	落地应用场景
OpenAI	人工智能通用性研究、自然语言处理、强化学习、生成模型	GPT系列 DALL-E Codex	ChatGPT DALL-E Copilot	聊天机器人、内容生成、编程助手
微软	AI与云计算结合、企业AI解决方案、自然语言处理	Azure AI Turing模型	Azure Cognitive Services Copilot	企业AI服务、生产力工具
谷歌	搜索引擎优化、AI与量子计算、AI伦理研究	BERT、Transformer、LaMDA	Google Assistant DeepMind AlphaGo、Gemini	语音助手、自动驾驶、搜索引擎
苹果	AI与硬件结合、隐私保护AI、人机交互	Siri、Face ID Core ML	Siri Apple Neural Engine	智能助手、人脸识别、健康监测
亚马逊	智能家居、推荐系统、云计算AI	Alexa、Rekognition SageMaker	Alexa AWS AI Services	智能家居、电子商务、云计算
英伟达	GPU优化AI计算、AI与边缘计算、深度学习硬件	CUDA、NVIDIA Tensor Core GPU	NVIDIA DGX Jetson、Omniverse	AI计算平台、自动驾驶、边缘计算
Meta	社交媒体与AI结合、虚拟现实、内容生成	LLaMA Computer Vision	Meta AI、Instagram推荐算法、Horizon Workrooms、Ray-Ban Meta AI眼镜	社交平台、广告推荐、虚拟现实

未来AI技术发展方向



2023年10月17日，美国加强了面向中国市场的AI芯片禁令。其中明确将性能、密度作为出口管制标准，将单芯片超过300teraflops算力，以及性能密度超过每平方毫米370 gigaflops的芯片都纳入了禁止出口行列。

虽然在这一禁令下，AMD、英特尔等公司提供的高端AI芯片也受到影响。但由于英伟达在全球AI芯片领域一家独大，本次禁令涉及A100、H100等主流AI训练用英伟达GPU，因此这次的芯片禁令也被称为“**英伟达禁令**”。

2024年6月19日，英伟达的市值达到了3.34万亿美元，首次成为全球市值最高的上市公司

英伟达人工智能战略研究报告

敬请期待 . . .

鼎帷咨询战略创新研究院

2024年9月

更多免费行业报告 并购家 www.ipoipo.cn