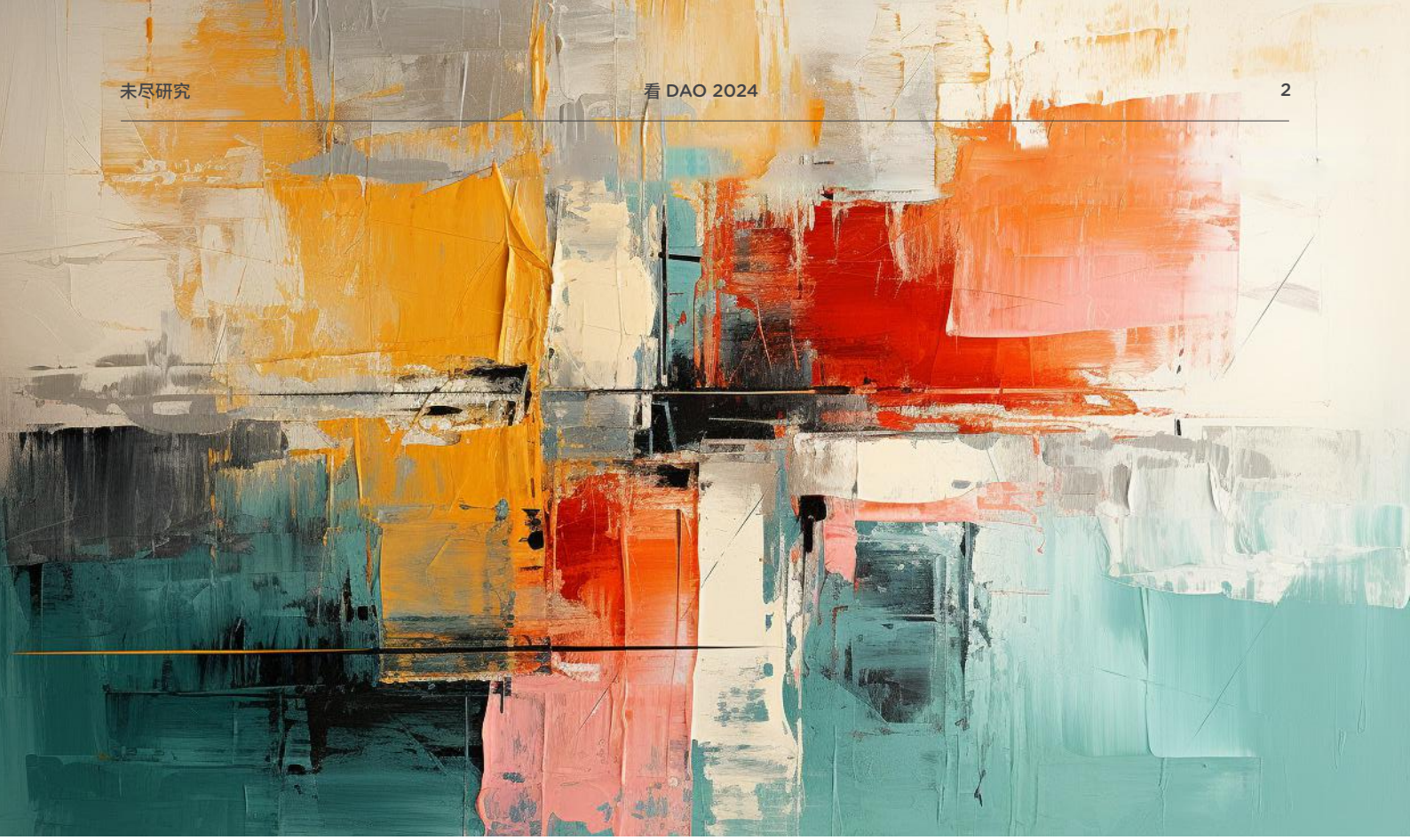
未尽研究
WEIJIN RESEARCH

看 DAO2024



如果没有 ChatGPT 打开的 AI 之光，2023 年的世界 ，将会暗淡很多。但 2024 年，AI 会不会更像是巨头手中的超级工具。与前两年不同的是，《看 DAO2024》选择了 AI 为主题，巨头为主角。

目录

■AI 工具：ChatGPT、Dall-E、Claude、Midjourney、Stable Diffusion 等使用入口

■AI 行业研报、书籍、论文（持续更新）

①AI 学社资源目录↓

②AI 学社入口↓

■AI 工具代理副业

①AI 工具代理副业介绍↓

②客服微信↓

看 DAO 2023 回顾

准确 对了一半 不准确

新能源

准确性： ● ● ●

核心判断： 全球化石能源消费将于 2025-2030 年间开始下降。中国碳达峰可能提前至 2025 年。未来 5 年新增可再生能源装机，中国将占一半。

典型事件： IEA 最新预计全球煤炭消费将 2026 年开始下降，中国煤炭需求在 2024 年下降。今年，中国水泥行业提前碳达峰。全球可再生能源新增装机容量刷新历史最高纪录，中国占近 55%。

人形机器人

准确性： ● ● ●

核心判断： 无论与人一样敏捷的双足机器人，与人自如对话的虚拟人，还是用外骨骼或脑机接口实现人体与机器的合体化，2023 年人形机器人将会加快走向商业化。

典型事件： 特斯拉“擎天柱”迅速迭代，小鹏、小米等公司涌入。李飞飞团队试验了 NOIR 脑机接口机器人。福特与亚马逊开始试点人形机器人 Digit 的商业应用。中国出台了全球首部发展人形机器人顶层设计文件。

自动驾驶

准确性： ● ● ●

核心判断： 2023 年，辅助驾驶将更大规模部署到电动车上，新的应用场景得到不断开拓，中国超越美国成为全球自动驾驶技术最大的试车场。

典型事件： 百度近五年来的累计路测里程与累计服务订单超越了 Waymo。中国当前乘用车 L2 及以上智驾渗透率超过了 42%，较去年 30% 大幅提升。换道大模型的特斯拉也准备好让完全版的 FSD 在中国落地。

低轨科技

准确性： ● ● ●

核心判断： 2023 年，在更多的公司和资本的支持下，火箭发射成本进一步下降，太空是经济发展的下一个前沿。

典型事件： 全球全年火箭发射次数首次突破 200 次，主要由 SpaceX 贡献；亚马逊的卫星星座正式升空。中国的朱雀二号成为全球首枚入轨的液氧甲烷火箭。但星舰尚未成功入轨，发射成本大幅下降仍待明年。

合成生物

准确性： ● ● ●

核心判断： 2023 年，人工智能等底层技术将继续创造与优化酶与底盘等生物元件，更多基于合成生物的碳中和技术与新药研发得以验证。

典型事件： Ginkgo 将与辉瑞共同开发多款 RNA 药物，美国能源部也资助这家行业巨头研究抗菌藻类以减少碳排放。但从研究到量产仍然困难。Ginkgo 市值从 280 亿美元跌至 28 亿美元，另一家合成生物巨头 Amyris 申请破产。

半导体

准确性： ● ● ●

核心判断： 中国半导体行业在 2023 年需要与时俱进的新一轮产业政策，聚焦制造薄弱环节，打造中国半导体行业的创新生态。

典型事件： 历经调整后，大基金二期加速投资步伐，集中于芯片制造环节。大基金三期目标筹集 400 亿美元。华为麒麟 9000S 的突破成为业界继续推进国产替代的定心丸，政策加快创新生态的形成。

元宇宙

准确性： ● ● ●

核心判断： 2023 年元宇宙的重点，将是设备的突破，即能为用户提供“基本可用”的 VR 和 AR 等设备，并且性价比易于普及。人工智能在 AIGC 的拓展，包括 AI 生成 3D，将成为 2023 年元宇宙的创新亮点。

典型事件： 苹果发布了 Vision Pro，但尚未正式销售。它的硬件性能远超当前主流消费级设备，但价格昂贵。受限于 VR 等设备全年出货量不及预期，尽管 AIGC 进展迅速，多数创新并非直接面向元宇宙。甚至 Meta 发布会也很少提及“元宇宙”，它的热度被空间计算取代。

人工智能

准确性： 🟢⚪⚪

核心判断： 2023 年值得期待的，不仅仅是 GPT-4，人工智能大模型将会生成视频、3D 建模，多模态的组合，也将会应用于科学研究，包括蛋白质结构预测分析、新材料、新能源。

典型事件： 2023 是大模型军备竞赛的一年，GPT-4、Llama2、Gemini 等相继发布，中美控制了世界上 80% 的大模型。Runway 与 Pika 等初创企业正在颠覆视频行业。人工智能辅助天气预报等研究成果占领了顶级科学期刊封面，预测出 200 多万种晶体结构。

下南洋

准确性： 🟢⚪⚪

核心判断： 2023 年，中国风险资本与华人创新者将在阵痛中寻找机会。跨境电商与金融支付等数字技术在东南亚扩散；先进制造将产能优势复制到更靠近市场的地区。

典型事件： 今年，投资者与创业者热衷于谈论中东主权财富基金。字节跳动收入超过腾讯，优势在于全球扩张；Temu 与 Shein 合计美国用户逼近亚马逊。比亚迪等新能源企业继续在欧洲投产，供应链向周边及海外延伸。中国创业者和企业家出海，走自己的全球化之路。

中美创投

准确性： ⚪⚪🔴

核心判断： 国家、资本与技术，正在中国形成新的创新生态。2023 年，随着走出疫情，以及对平台经济建立常态化管理，消费和互联网的投资有可能再度活跃。

典型事件： 今年，中国“以投促引”进一步下沉，中西部省份新成立基金数量逆势增长，政府引导基金目标规模提升至 13 万亿，硬科技企业上市退出全球最为活跃。但在全球范围内，消费与互联网企业融资都大幅缩水。

人口国运

准确性： 🟢⚪⚪

核心判断： 在长期推动经济增长的人口、资本与全要素生产率中，中国将更加依靠全要素生产率的提升，更加依靠青年人才参与到创业与创新活动中。

典型事件： 年底，官方解读中央工作会议，强调“新质生产力以全要素生产率提升为核心标志”，围绕“技术革命性突破、生产要素创新性配置、产业深度转型升级”展开。

瞧，这些万亿巨头

苹果，微软，字母表（谷歌），亚马逊，英伟达，Meta，特斯拉（马斯克），这些公司之间有什么共同点？

是的，它们都是科技巨头（Big Tech）。它们的市值目前都在万亿美元以上，或者曾经达到过。其中的 5 家，每年研发支出达到了 250 亿美元以上。

科技巨头曾经是指那些通过互联网建立起强大平台经济的企业，拥有十亿用户级别的软件及应用。它们往往赢家通吃（Winner takes all ）。

科技巨头的概念也在随着技术演变。它们实现了软硬件一体，除了软件，它们也设计和制造终端设备，包括 PC、手机、可穿戴、AR/VR、传感器、机器人、智驾汽车，它们也正在渗入到制造业的流程中。

它们还建立起了强大的云计算和企业服务能力，包括芯片、超级计算机和超大规模的数据中心，为第四次技术革命提供最重要的基础设施。

它们已经开始全面竞争一种全新的能力，生成式 AI。在深度学习领域的竞争，从 2012 年视觉计算取得突破时就已经开始，因为大模型的流行而加剧，但只有在 ChatGPT 发布之后，过去的一年，生成式 AI 大模型成为科技巨头之间“军备竞赛”的焦点，迅速成为巨头技术栈上标配的一层。它正在成为所有巨头的业务基础模型。

科技巨头的核心能力，越来越区别于其他非科技企业和非巨头企业。它们能把技术和应用端到端地垂直整合到一起，不断扩张业务范围，形成一种科技巨头所独有的、以计算和智能为核心的创新能力。它们都是全球化的企业。

这样看来，华为也是这样一家科技巨头，它也已经建立起了从芯片到软硬件应用的全栈技术能力，以计算和智能为核心展开业务范围。它在 2022 年的研发投入达到了 240 亿美元。但华为是其中几家巨头的挑战者。

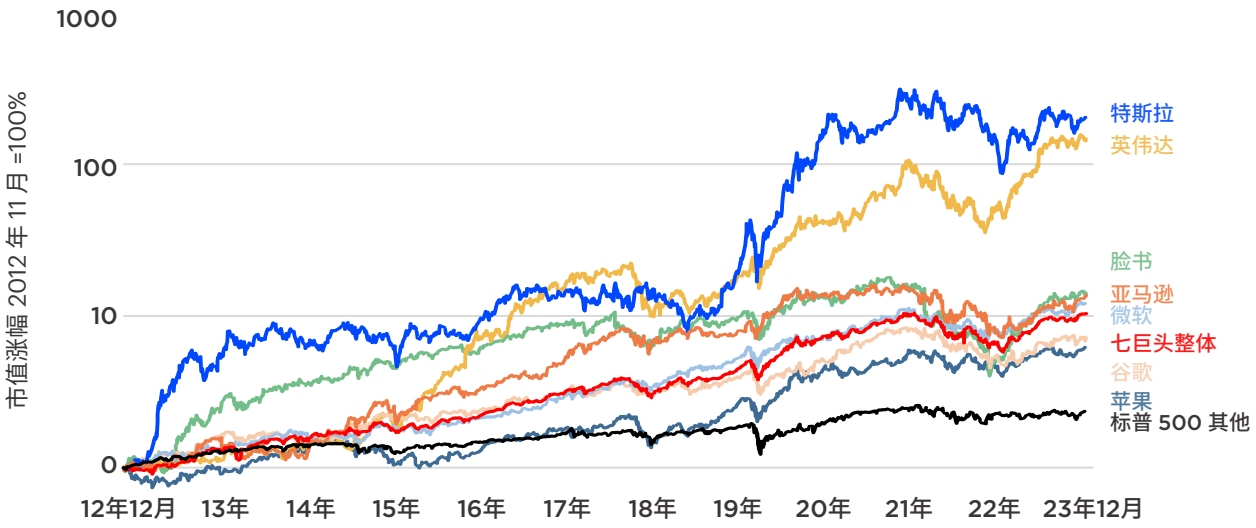
巨头公司的研发投资覆盖了基础技术研究和具体产品的开发。除此之外，它们每年还把研发和投资的 20% 左右，投入企业内部 IT 系统、效率软件、先进的技术平台、以及数据中心等基础设施，并持续地提升员工技能，不断把那些复杂和重复的业务和流程进行自动化和简化。

巨头企业保持内部技术的先进性，是其创新和竞争力的一个重要来源。巨头用最先进技术围绕用户数据建立起了正向的反馈机制，形成了飞轮效应。

人工智能正在加快这一飞轮的运转，放大它们在各自领域的核心能力，也在生成新的能力。生成式 AI 更像是这些科技巨头为自己发明的新工具，一种放大器和加速器。

AI 最早的受益者是科技巨头自己。英伟达从 2012 年起就开始为深度学习提供 GPU，营造 CUDA 生态，10 年之内把自己变成了科技巨头。巨头最早受益于 AI 的，都是其核心业务，如云计算和广告，目前是企业服务和生产力软件，接下来还有硬件和消费智能产品，以及新的“赢家通吃”的领域。

深度学习的黄金十年



来源：Wind，未尽研究
说明：市值涨跌幅，对数轴。标普 500 其他指不包含七巨头的标普 500 成分股的总市值，此外，特斯拉等部分企业部分时段不属于 SP500 成分股。七巨头整体指七巨头总市值的整体变化。以 2012 年 11 月 19 日为基准 100%。

巨头的无限游戏

我们正处于当年个人电脑开始的同样时期。1980 年代初，信息技术革命发轫于英特尔发明的 CPU。乔布斯创办了苹果电脑，比尔盖茨创办了微软，近 50 年后，它们跨越了 PC、互联网、移动、云计算，直到人工智能，至今是世界上市值最大的两家公司。

同样，2023 年，AI 开始真正大规模走向消费者。AI 时代真正开启，人们称之为苹果时刻、人机交互的范式转移时刻，寻找这个时代的苹果和微软。

但是，科技巨头似乎从源头就控制着这一切。

通用人工智能最具颠覆性的两家初创企业，DeepMind 和 OpenAI，前者被谷歌收购，后者技术被微软买断和控制。

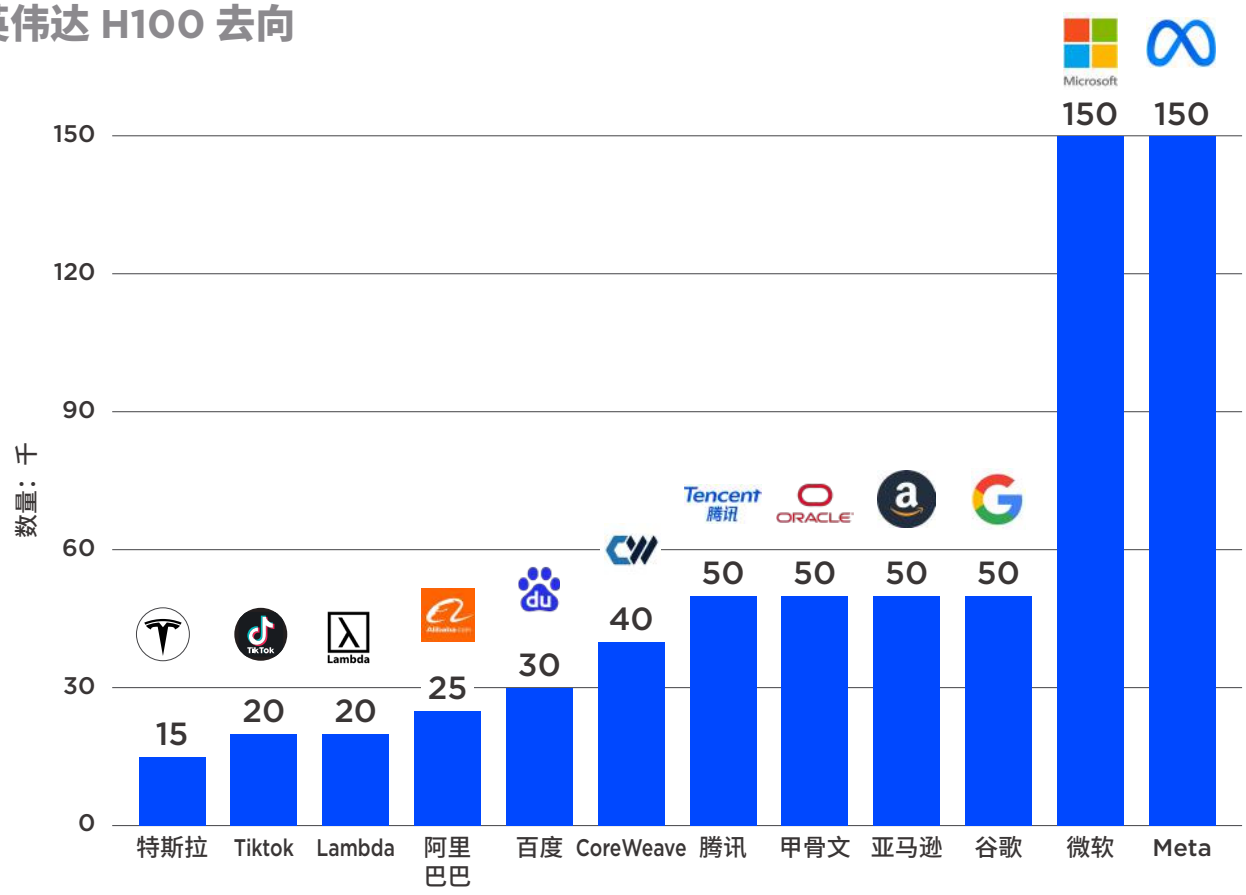
生成式人工智能，从训练最先进的大模型，如 GPT-4、Gemini、Claude2，到向数亿用户部署和应用，多数都要依靠微软、亚马逊和谷歌这三家的云计算。英伟达成为新晋的科技巨头，除了用于 AI 加速的 GPU 芯片，它也在构建其人工智能的基础设施，还成为过去一年投资人工智能初创企业最多的公司。

它们控制了数据。用来训练生成式人工智能的自然语言数据、视觉数据、代码符号数据、知识图谱，包括合成数据，海量地来自并存储在这些巨头的业务、平台和基础设施中。更重要的是，大数据创造了对自动化和人工智能的需求。

生成式人工智能重新定义了大数据。由于人类已经开始掌握以大型语言模型（LLM）技术以产生智能，网络上的公开数据、企业数据、个人数据、自然界中的数据，以及人工合成的数据，都可以成为训练智能的原料。科技巨头本身拥有巨大的数据库，它们还在获得更多和更好的数据。它们建立起了联盟，为用于人工智能训练的数据建立标准。它们可以合成数据，成为新的模型的训练的来源。目前许多人工智能的研究和模型训练，开始用 GPT-4 等先进大模型生成或者标注的数据来训练。除了公开数据，OpenAI 还开始与各机构展开私有数据的合作。

它们掌握了算力。巨头们已经在全球各地建立了数据中心，拥有最先进的 AI 加速算力；或者已经囤积的 AI 芯片，已经超过了世界上许多中等国家所拥有的数量。它们除了用来实现自身业务的 AI 化之外，还去进行科学探索：AI 用于创新药研发、医疗服务、芯片设计、材料发现、能源转型和应对气候变化。

英伟达 H100 去向



来源：Omdia，未尽研究
说明：预估 2023 年 H100 主要买家购买的 H100 芯片数量。

巨头所拥有的强大的算力，实际上是人工智能所引领的第四次工业革命的基础设施。微软以后每年将在数据中心投入 500 亿美元，包括自行研发的芯片的支出，这已经相当于一个科技大国的 AI 基础设施的投入。亚马逊和谷歌，都在为数据中心更高的计算效率研发芯片。苹果研发的手机和个人电脑的芯片，已经超过了专业的芯片设计公司。

美国及许多国家的政府、大学、研究机构，会日益依赖这些巨头的算力基础设施。发展人工智能成为许多国家优先考虑的事项，这些巨头的云中心已经遍布世界各地，它们可以在这些数据中心的基础上，轻松地与当地政府合作。

拥有数据和算力，加上资本的力量，这些巨头可以吸引世界上最好的人才。它们建立起了最先进的研究部门，吸引了世界上最优秀的图灵奖获得者和理工科博士，也掌握了最好的算法——Transformer 论文就出自谷歌，而谷歌一直是最高质量 AI 论文的来源地。Meta 也建立了超级算力集群，研发出最流行的开源机器学习库 PyTorch，不仅支持自己在社交媒体上的推荐算法，支持它建立起最大的线上广告系统，而且推出的大模型 Llama 引领了开源大模型潮流。

这些巨头公司还有一个重要特点，都是从初创企业成长起来的，除了苹果公司之外，其创始人依然在管理企业，或者对企业的方向与战略发挥着影响力，其 CEO 依然能让日益庞大的企业保持敏捷。企业体内部活跃着技术基因，工程师思维主导了企业文化。

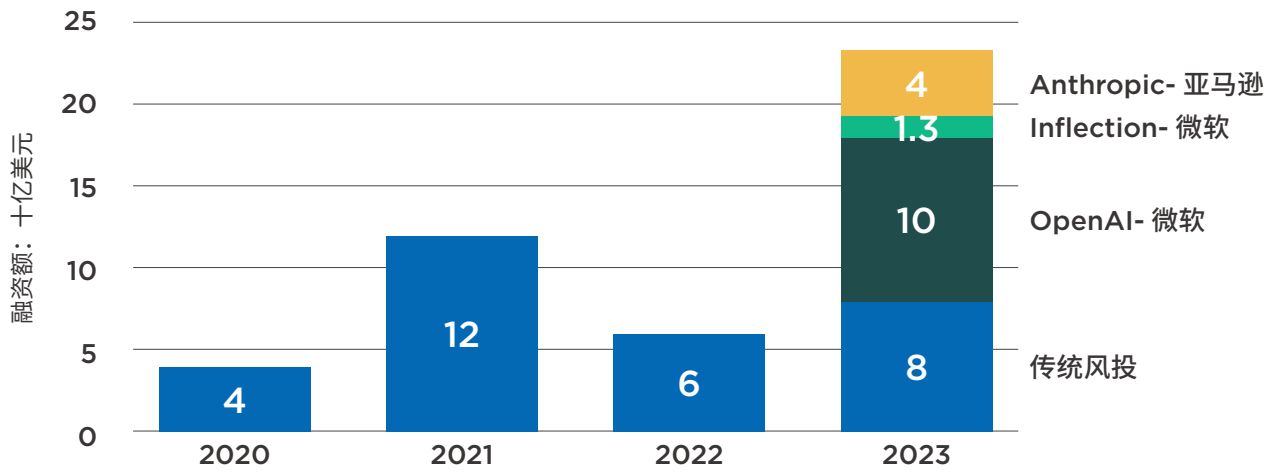
巨头不仅砸下巨资进行前沿技术的和新产品研发，打造先进的企业 IT 系统和技术平台，它们还投资、收购初创企业。2023 年，对生成式 AI 的投资，微软、谷歌、亚马逊、英伟达等几家大型科技巨头对大模型初创企业的投资，金额上远远超过了独立的风险投资机构。

它们投资的战略性也越来越强。巨头投资大模型初创公司，其中相当大的金额就是算力信用的投入。例如微软对 OpenAI 投资 130 亿美元，其中很大一部分是 Azure 云计算；亚马逊以 40 亿美元投资大模型初创公司 Anthropic，其中多数是 AWS 的算力信用。而巨头们自研的 AI 芯片，也将用于这些大模型的训练和推理功能。谷歌则更早建立起这样一个共生链条。硅谷人称“云洗钱”。但目前食物链的顶端仍然是英伟达。

巨头们不仅控制了最强大的闭源大模型，而且控制了最流行的开源大模型，如 Meta 推出的 Llama，微软研发的 Phi 系列的小型开源模型等。Google DeepMind 还能结合最强的强化学习模型进行科学发现。

这些企业还拥有全球化的优势，它们的用户、业务、供应链和数据中心遍布世界各地。美国巨头与中国巨头的国际竞争主要在东南亚，AI 的兴起，让两国巨头在社交、视频、电商、云计算的竞争愈发激烈。

巨头主导生成式 AI 风险融资



来源：Pitchbook，未尽研究
说明：全球范围内，截至 2023 第三季度。

生成式 AI 十大展望

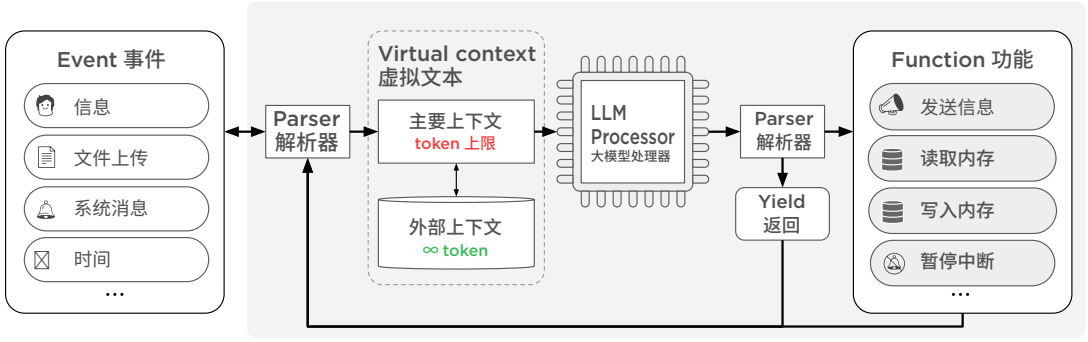
2024 年，优化训练和部署大模型仍然非常重要，大模型的生态加速形成，应用开始在一些领域大规模展开，主要表现在如下十个领域：

1. 智能体作为任务助理进入更多应用场景和业务流程

智能体能有一定的主动性，能帮助完成任务，而不仅仅是问答。在感知环境后，通过其大脑（大型语言模型），调动其他的程序、应用、知识，甚至自己编程，规划和执行更复杂的任务。有了智能体，许多人可以用经验和专业知识，通过自然语言而不限于编程代码去写软件。

2. 操作系统集成下一代大模型，成为下一代操作系统

大型语言模型日益操作系统化，AI 芯片为它设计，PC 和手机的操作系统为它升级，AI 应用成为它的下游，上下文管理类似于操作系统的内存。微软将推出 Windows12 操作系统，在 PC 上与下一代大模型和 Copilot 深度集成。在移动设备上部署的模型，也与 iOS 与安卓操作系统紧密结合，实现 AI 功能和建立 AI 应用商店。



来源：*MemGPT: Towards LLMs as Operating Systems*

3. 生成式 AI 制作的影视剧大量出现，冲击影视行业发生剧变

图像和视频是生成式 AI 迭代最快的领域之一，GPT-4V 等多模态大模型的推出，基于扩散模型的 Dall-E 3、Midjourney 和 Stable Diffusion 的功能不断增强，LCM-LoRA 等技术达到了实时生成图像和视频的效果，对影视、音乐、游戏等内容娱乐行业的影响是颠覆性的。这方面的应用也是巨头目前还染指不多的领域。2024 年将大量出现由生成式 AI 产生的影视剧，创作者、用户以及角色之间将会出现崭新的交互方式。

4. 人形机器人开始量产，自学习与环境互动能力进一步强化

在已有的机器人技术之上，多模态和具身智能的大模型，不断展示出惊艳的效果。大型语言模型的推理和规划能力，与视觉模型结合，可以通过获取周围环境数据，学习人类用手脚完成任务。2024 年人形机器人开始量产，开始在工作场景中进化迭代人类的灵活性。

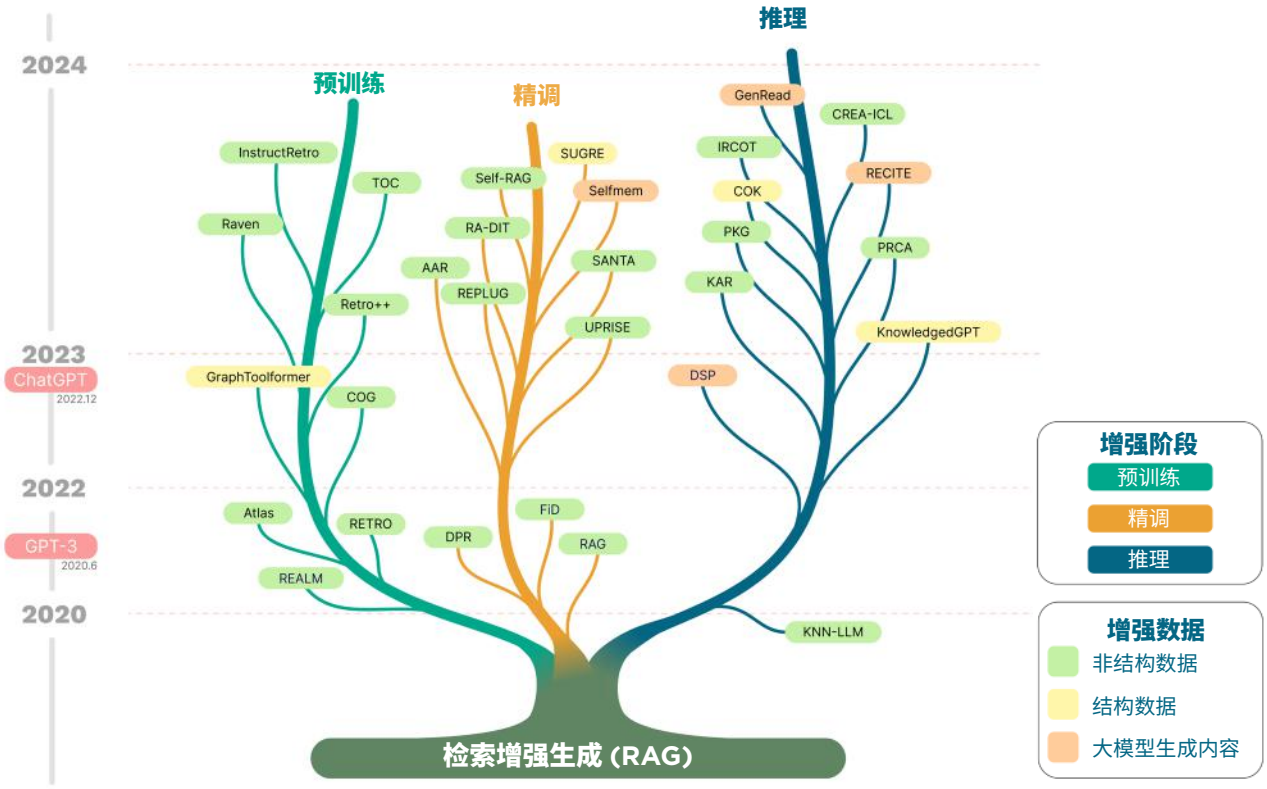
5. 终端设备加载 AI 模型，推动换代升级

小型化的大模型可以加载到笔记本电脑和手机等终端设备上，用户不仅可以更快捷地生成内容，而且可以利用自己本地的数据和知识进行检索生成，建立起定制化的智能体，更快捷地进行推理，也保护了数据安全和个人隐私。AI 设备的主流硬件规格将包括内置 7-10B LLM 模型、40-50 TOPS AI 算力、10-20 token/s 以上推理速度、8-16GB 以上 DRAM 等。

6. 下一代闭源大模型推出，开始出现胜任人类水平的 AGI“火花”，但规模边际效应递减

OpenAI 与微软将推出 GPT-5，谷歌将推出 Gemini Ultra，亚马逊也在训练数万亿参数的大模型。下一代大模型将是多模态的、使用更多合成数据的、混合专家系统的，会消除一些幻觉、增加上下文长度、信息更加准确和及时、基础数学水平有所提升，等等。更多更好的数据、更强的算力、更顺的搜索，依然是产生智能的根本因素。加上 RAG（检索增强生成）补充非参数化的知识，闭源大模型会应用于更多的场景。

检索增强生成技术的时间线



来源：*Retrieval-Augmented Generation for Large Language Models: A Survey*，2023，Yufan Gao et al

7. 数据来源的深度和广度进一步开拓，进一步规范，更多合成数据与自然数据结合用于大模型训练

数据决定了泛化的边界。自然语言数据，以及直接从现实世界事件或对象中收集得到的数据，已经无法满足下一代大模型的训练的胃口。在专业领域和垂直场景，非公开的数据将会发挥更大的价值。大模型训练、自动驾驶、机器人、图像生成、模拟仿真等，都在大量使用合成数据的同时生成新的数据。越来越多的数据标注也由 AI 来完成。但是，只使用合成数据可能会造成数据多样性不足和自循环训练的问题。2024 年将会看到 AI 企业寻求合法获取更多非公开数据，以及使用更多的混合数据。

8. 苹果真正入局，力争复现 AI“iPhone 时刻”

2023 年被称为大模型之年，苹果表面上在作壁上观，但实际上在芯片及软硬件方面的研发一直在加大力度，只做不说。2024 年苹果将把 Vison Pro 推向市场；PC 和手机加载大模型，苹果是其中最重要的玩家；为了建立 AI 应用生态，操作系统封闭的苹果拥抱开源模型。苹果被广泛期待能给消费市场带来更好的 AI 产品体验。

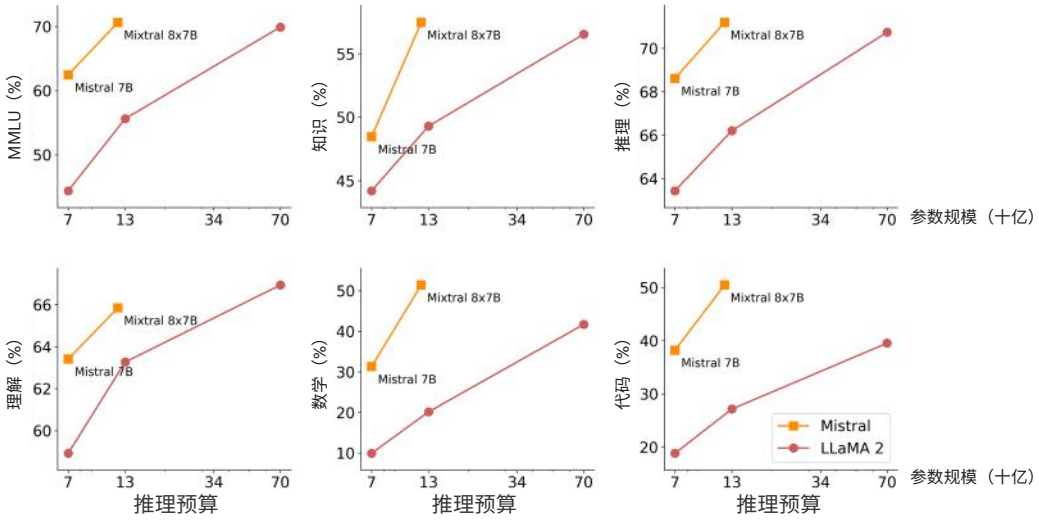
9. 一些开源模型及 AI 应用，因为无法建立起商业模式将面临生存危机

绝大多数初创企业的开源模型，目前还无法在提供推理服务、授权、训练和部署模型方面建立起用户基础；消费类模型 + 应用的初创企业，在激烈的竞争中多数将遭淘汰；纯应用类的初创企业，许多将遭到巨头碾压或者很快在更新的开源技术迅速推广中出局。快速获取用户并且在反馈中建立起数据飞轮的企业将赢得生存。而能结合起应用场景、行业深度和垂直数据来源的企业，将能保护好自己。

10. 小模型结合软硬件应用，新物种涌现

2023 年是大模型之年，2024 年也将是“小”模型之年。更多几十亿到上百亿参数的小模型，通过模型架构、算法、训练和精调的创新，以及结合外部检索，性能可以叫板百亿参数大模型，甚至追平 GPT-3.5 （1750 亿参数）。开源模型许多来自中国、欧洲、韩国、甚至中东等地，以更快的速度推广到各行各业。小模型尤其适于下载到设备上，在许多功能上可以替代从云上提供的大模型服务。小模型 + 终端设备是 2024 年的重要看点。

Mistral 7B 和 Mixtral 8X7B，碾压 Llama2



来源：Mixtral，https://mistral.ai/news/mixtral-of-experts
说明：Mixtral 模型参数规模分别为 70 亿与 120 亿，LLama 2 模型参数规模分别为 70 亿，130 亿与 700 亿。

算力“破墙”

“芯片战”在科技巨头之间、芯片巨头之间、中美之间烈度不减。

深度学习的黄金十年，终于产生了黄金般昂贵的 GPU 芯片。

大模型以每年 10 倍的速度扩大参数规模，对算力的需求每两个月翻一倍；GPU 算力每年翻倍；摩尔定律以每 18 个月翻倍的速度已经放缓；内存增长的速度更慢。服务器和数据中心的扩张已经追不上神经网络的加速野心。算力卡住了第四次工业革命的脖子。

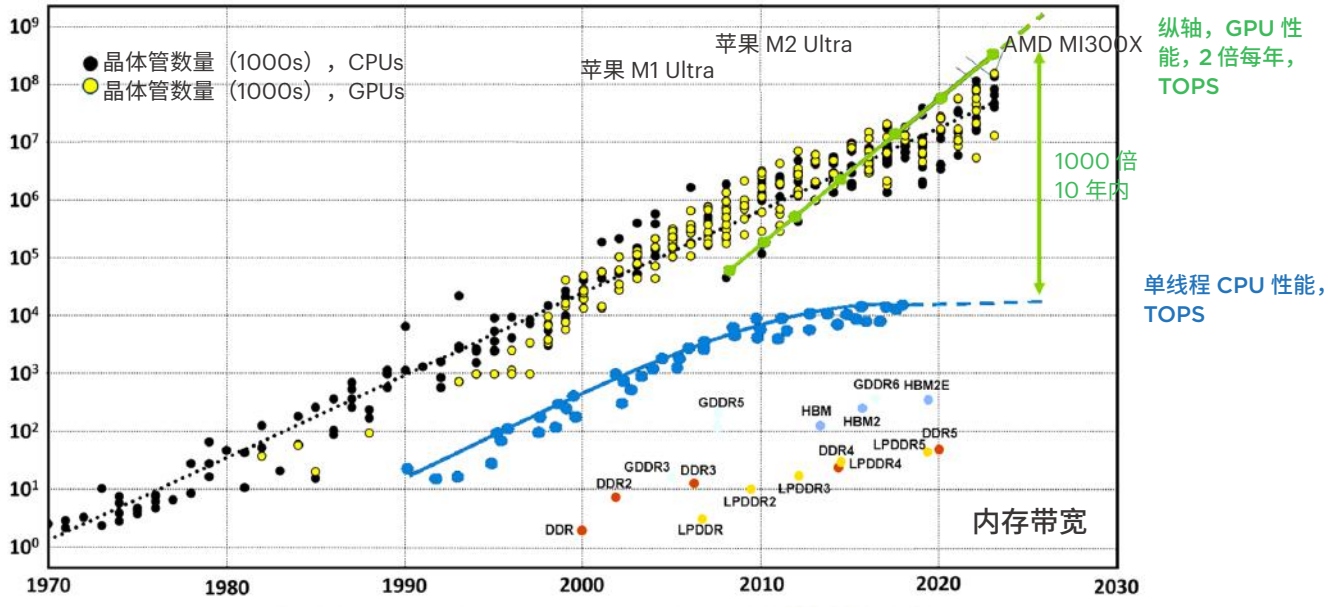
尤其是 ChatGPT 的推出，让更多人看到了通用人工智能的曙光。但也很快带来两个后果，第一是算力非常短缺；第二是短缺的算力使用起来也非常粗放。

在 Transformer 架构中，每预测一个单词，就要在上万亿个词元（token）和上千亿个参数中通过概率计算筛选出最合适、最对齐人类的那个字，这需要在存储和计算之间每秒钟调用 10^12 字节的数据量。

在内存和计算之间高速移动 TB (10^12) 级别的数据量，需要 TB/ 秒级的数据传输带宽，这远远超出了目前的内存能力，被称为“内存墙”。如果处理器没有及时接收到数据，它就会处于空闲状态，影响其效率。有研究发现，GPT-4 在最先进的芯片上运行的效率可跌至 3% 或更低。

为了弥补数据中心处理模型训练和推理的低效率，云服务提供商增加了更多硬件来执行相同的任务。这种方法导致成本急剧上升，电力消耗也成倍增加。

GPU vs. CPU vs. 内存



来源：*Arteris/NVIDIA and "Fast validation of DRAM protocols with Timed Petri Nets,"* M. Jung et. al, MEMSYS 2019

大模型经济，在过去的亢奋的一年中，基本上就是在这么昂贵而又短缺的算力基础上开始建立起来。

这样的结果，就是让英伟达这家 30 年前创办的企业，从一个做游戏显卡的公司，迅速膨胀为一个大模型训练和推理芯片的垄断者。因为其 GPU 的并行计算、张量计算、存储、Nvlink 等的集成，及其 CUDA 软件形成的生态，不仅在算力上遥遥领先，而且在生态上无处不在。AI 公司只有足够的数量的 GPU 卡，才能吸引顶级的 AI 科学家。

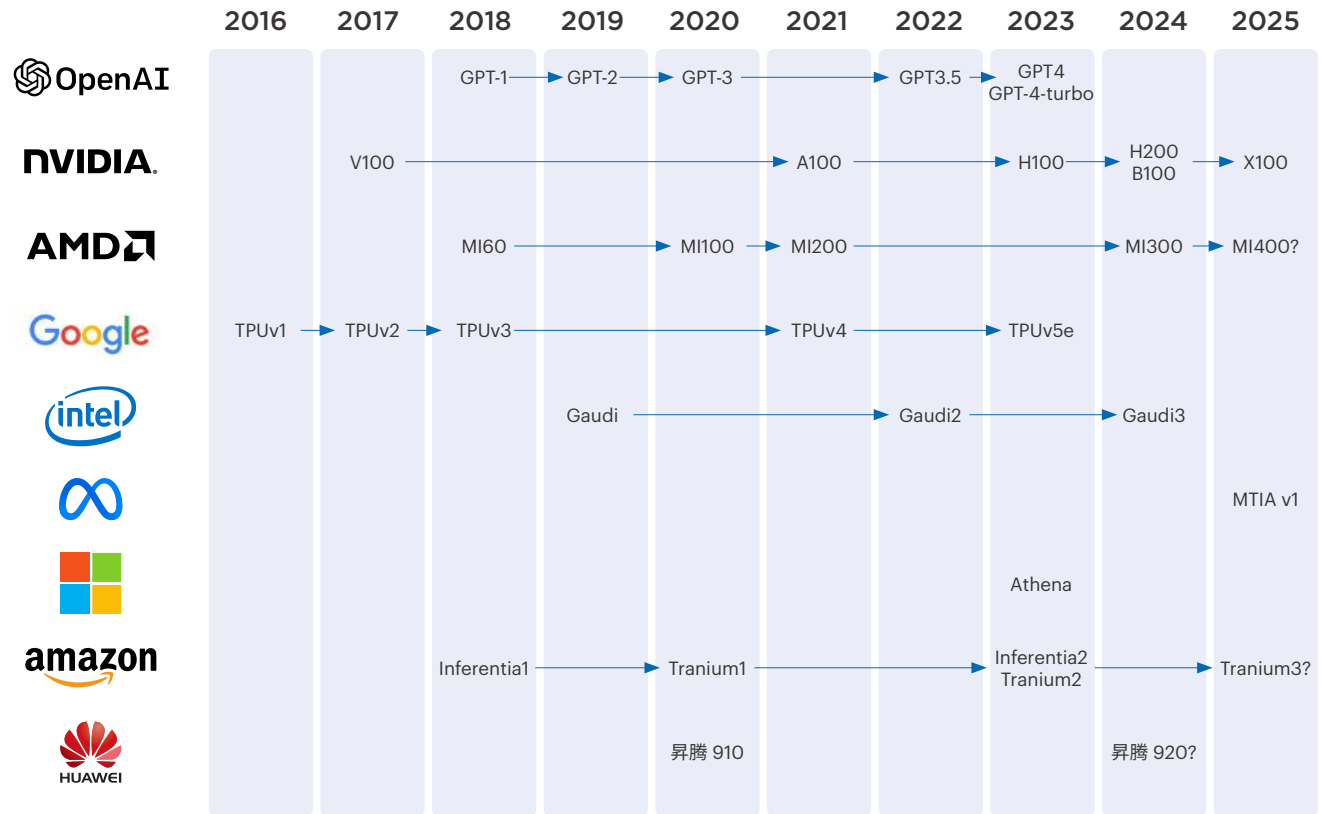
生成式 AI 对于科技巨头的未来如此重要，而算力又如此昂贵，在 2024 年会带来一系列效应。

科技巨头加快推出主要用于推理的自研芯片。因为一旦部署了训练好的模型，支持业务运营的推理成本将会更大。科技巨头首先要考虑的是降低推理成本，并且结合算法，为用户推出差异化的产品与服务。

在数据中心，GPU 相对于 CPU 的主导作用越来越强。芯片架构从异构转向超异构。CPU、GPU、DPU、NPU，以及匹配用户需求的各种算法的芯片，越来越多地集成到单个芯片中（SOC），或者形成相互协处理的系统。科技巨头纷纷进入芯片设计领域，这对于芯片的架构和封装技术带来挑战，也带来创新的机会。

更加复杂的架构，对于不同计算要素之间的数据传输提出了很高的要求，加上存储墙的存在，片上网络（NOC）对人工智能加速起到决定性的作用。

巨头自研 AI 芯片



来源：公开资料，未尽研究
说明：指正式发售时间，而非公开其存在的时间。部分芯片仅官方或传言预估或披露了其量产时间。

尽管黄氏定律还跟不上大模型的规模法则，这种从 PC 时代沿袭下来的软硬件互相加速的节奏，如英特尔 CPU 与 Windows 操作系统的互相借力，正在 AI 时代延续，只是这次换成了英伟达的 GPU 和 OpenAI——大模型正在成为新的操作系统。

生成式 AI，把芯片战提升到了一个新的高度。这不仅是科技巨头之间的竞争，世界上最先进的大模型之间的竞争，芯片企业之间的竞争，而且也是国家之间的竞争。

在过去的一年，美国升级了对中国的芯片管制，小院高墙扩大了地缘遏制的范围，长臂之手伸向了盟国和友岸。美国对中国半导体出口控制，商务部的 BIS 每年十月将会审核。2024 年，中美芯片战是否会扩大升级，仍然值得关注。

美国对中国禁运的高端 AI 芯片

GPU	内存 GB	内存带宽 Tbps	算力 TFLOPS	位宽	TPP 算力 算力 * 位宽	晶粒尺寸 mm^2	性能密度 TPP 算力 / 晶粒尺寸
H100 SXM	80	3.4	1979	8	15832	814	19.4
H20 SXM	96	4	296	8	2368	814	2.9
L40S	48	0.9	733	8	5864	608	9.6
L40	48	0.9	362	8	2896	608	4.8
L20	48	0.9	239	8	1912	608	3.1
L4	24	0.3	242	8	1936	295	6.6
L2	24	0.3	193	8	1544	295	5.2
A100 SXM	40	1.6	312	16	4992	826	6
V100 SXM	16	0.9	125	16	2000	815	2.5
RTX 4090	24	1	661	8	5285	609	8.7
RTX 4080	16	0.7	320	8	2560	379	6.8
AMD MI210	64	1.6	181	16	2896	770	3.8
AMD MI250X	128	3.2	383	16	6128	1540	4
AMD MI300X	192	5.6	2400	8	19200	2381	8.1
Intel Gaudi2	96	2.5	700	8	5600	826	6.8

来源： semianalysis，未尽研究
说明：部分无公开资料，为估算数据。RTX4080 与 RTX4090 等高端消费级 GPU 并未用于构建数据中心。

中国在 2023 年开始出现突破点，主要是 7 纳米制程。通过用于 14 米制程的深紫外光（DUV）技术的两次曝光，华为代工方生造出了 7 纳米的麒麟 9000s 芯片。华为通过魔改 ARMv8.2，形成了自己的鲲鹏 CPU 架构。华为用于数据中心的昇腾 AI 芯片达到了 A100 至 H100 之间的性能。华为已经站到了一个新起点，2024 年值得期待。

Gartner 预计 2024 年，全球半导体收入将增长 17%，其中内存市场将强劲反弹，增长达 66.3%。而世界半导体行业统计（WSTS） 预计， 2024 年将出现强劲反弹，预计增长 13.1% ，这一轮增长主要由存储器推动。2024 年，是中国芯片行业值得期待的一年。

自动驾驶换道大模型

大模型重置了智能驾驶竞争格局，特斯拉领先，中国规模量产电动车加速追赶。

今年，大型语言模型改变了自动驾驶技术路线的竞争格局。它正在教会规模量产的电动汽车，像个五星司机一样开车。特斯拉正处于有利位置，开始探索自动驾驶的世界模型。

Waymo 继续在多个城市运营自动驾驶车队。它向凤凰城、旧金山、洛杉矶和奥斯汀的公众开放。中国加速更快，年初，北京与加州各自披露年度路测报告，百度近五年来的累计路测里程，实现了对 Waymo 的超越；小马智行也在飞驰。

北京速度仍在延伸。很多大中城市都将成为北京。截至今年三季度，百度的萝卜快跑累计服务订单超过了 400 万单，去年同期累计 140 万，按这个速度，明年将超越千万订单。百度的全无人自动驾驶车队，也已驶入北京、武汉、重庆、深圳、上海五城，还将进一步扩容。

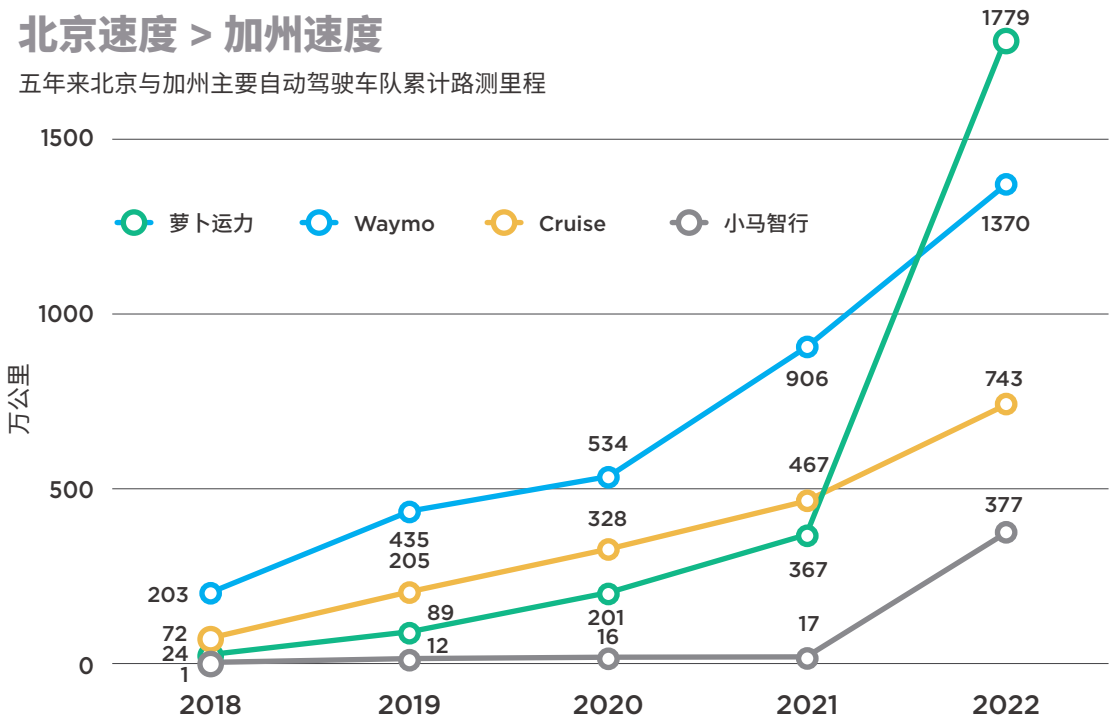
但变局在于大型语言模型以及生成式人工智能，尤其是今年以来，它迭代了包含机器视觉在内的多模态能力。很多学术机构认为 GPT-4V 对自动驾驶影响深远。

这个时代更有利于特斯拉这样的规模量产玩家。特斯拉的 Autopilot（自动辅助驾驶功能）系统，此前依赖基于规则的方法。今年，马斯克的全自动驾驶技术新版本 FSD V12，开始用数十亿帧人类驾驶的视频，来教会自己如何驾驶。

这种端到端（end-to-end）的训练，不需要人类明确编写代码或脚本，它的瓶颈很大程度上不再是代码量，而是视频输入量。神经网络在训练了至少一百万个视频后才能见效。

北京速度 > 加州速度

五年来北京与加州主要自动驾驶车队累计路测里程



来源：DMV，智能车联，未尽研究
说明：萝卜运力（含百度）、小马智行（不含小马智卡）为北京路测数据；Waymo 与 Cruise 为加州路测数据。百度等企业也在加州路测，此处未做统计。为统一单位方便，本报告中，1 英里 =1.6 公里。从 2018 年至 2022 年。

特斯拉是全球电动汽车销售冠军，今年考虑将 FSD 授权给同行；还传出将在中国继续扩建产能，从目前的每年 125 万辆，提高至 175 万辆。特斯拉的超算 Dojo 已经投入生产，相当于明年再新增 30 万片全球稀缺的 A100 芯片，来对付随之而来的海量数据。

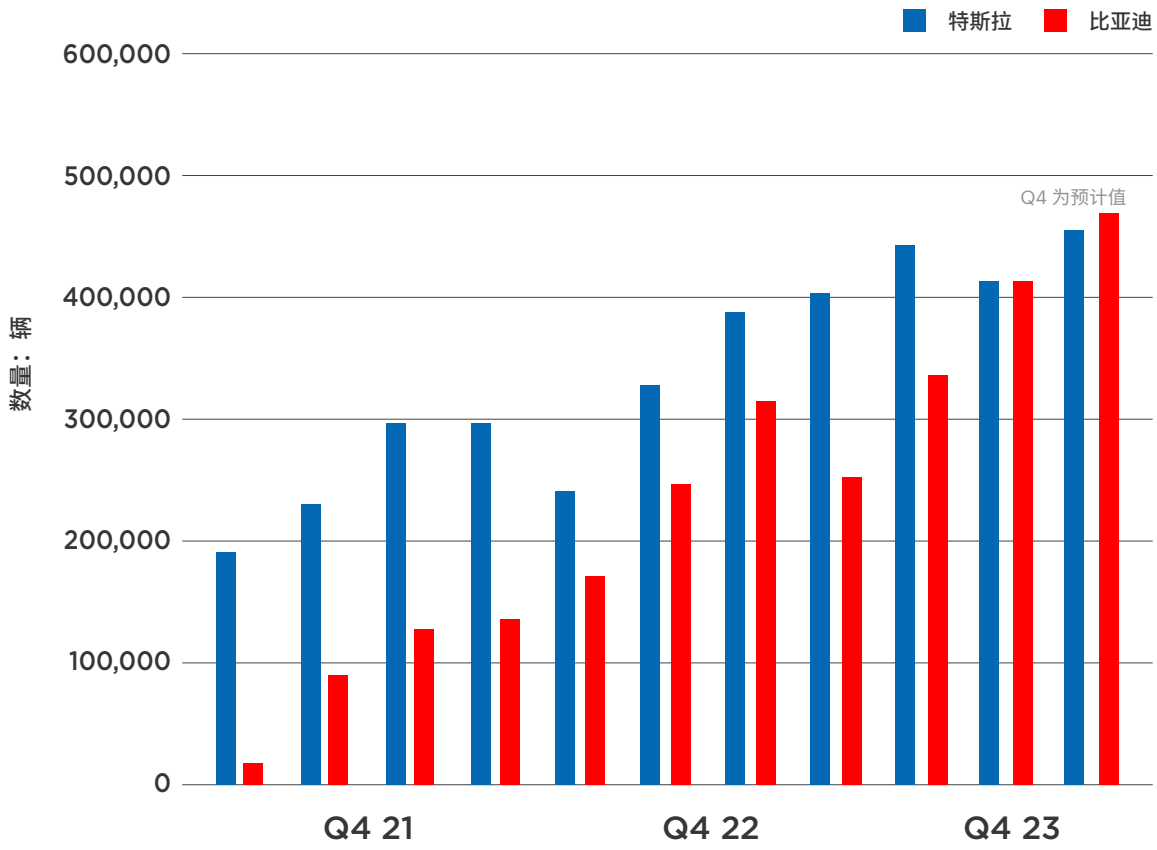
但 Dojo 架构早在几年前就定下，非常适合特斯拉当时独特的算法。如果拿它来跑大型语言模型，内存带宽恐怕不够友好。换道大模型的特斯拉，需要调整它的算力基础设施，或者突破模型底层技术的束缚。马斯克的应急方案是，赶紧抢购了万卡 H100 集群的庞大算力。

马斯克渴望让完全体的 FSD 进入中国。安全监管的障碍正在有序撤去。11 月，工信部等四部门发布了《关于开展智能网联汽车准入和上路通行试点工作的通知》，首次明确了不同情况下的交通事故责任归属，鼓励年底前完成试点集体申报，遴选具备量产条件的智能网联汽车。上海经信委“推动特斯拉自动驾驶在沪布局”或在明年成为现实。

中国汽车行业正在从电动化转向智能化，来自外来者的竞争，将加速中国量产车型拥抱大模型。除了自动驾驶技术企业外，大模型核心玩家的华为、以及“蔚小理”等一众造车新势力，也在尝试融入 Transformer 架构。毫末智行还发布了 DriveGPT。但它们的规模量产与特斯拉相比，尚处于爬坡阶段，在技术的垂直整合能力方面还有距离。

比亚迪的电动汽车销量，最早将在年底实现对特斯拉的超越。它与特斯拉一样，正在垂直整合整条汽车产业链，但尚欠缺一点软实力。比亚迪也在走基于 Transformer 的决策规划大模型，今年大概会有 6 亿公里的数据，标注自动化率超过 95%；内部架构也发生了调整，近期招聘了超过 4000 名软件工程师，“采用人海战术，保持颠覆性迭代能力”。

电动汽车季度销量对比

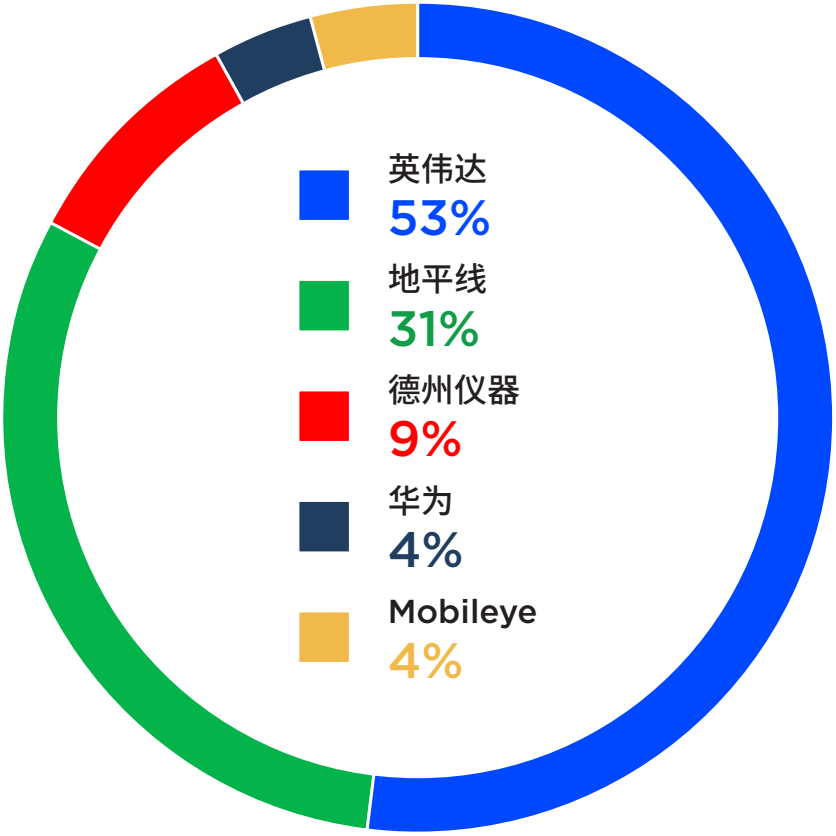


说明：BEV，即动力电池汽车

国内算力也跟得上。英伟达的智能驾驶芯片，占国内前装 NOA（自动辅助导航驾驶）市场超 50%，并未被“高墙”所限；今年，比亚迪进一步扩大了与英伟达的合作，两家企业的共识是未来的汽车是可编程的。地平线占超 30%，年底，它推出了征程 6，专为大参数量 Transformer 设计，明年正式交付；比亚迪是首批量产意向合作车企。华为也拆分旗下智能汽车解决方案业务单元，引入长安汽车合伙。华为已经突破了制造等效 7nm 水平的高算力芯片的封锁。

大模型将重置中国与美国这场从电动化转向智能化的竞争格局。2024 年，在中国 500 多万公里的道路上，将上演国内智能驾驶车企守住领先身位，比拼大模型应用落地的一幕。但无论如何，安全第一。

中国智能驾驶芯片市场



来源：高工智能汽车研究院，未尽研究
说明：截止 2023 年 6 月，仅有上述 5 家计算芯片方案商在 NOA 领域实现量产交付（新车上市销售为准），其余厂商（包括黑芝麻智能、芯驰科技、后摩智能等）目前仍未量产交付。不包括特斯拉自研 FSD 芯片。

参考：A Survey of Large Language Models for Autonomous Driving；GAIA-1: A Generative World Model for Autonomous Driving；Comparative Safety Performance of Autonomous- and Human Drivers；2022 Disengagement Report from California；北京市自动驾驶车辆道路测试报告（2022 年）；关于开展智能网联汽车准入和上路通行试点工作的通知

量产人形机器人

人形机器人完成商业验证，领先企业率先量产，中国加速核心部组件安全可控。

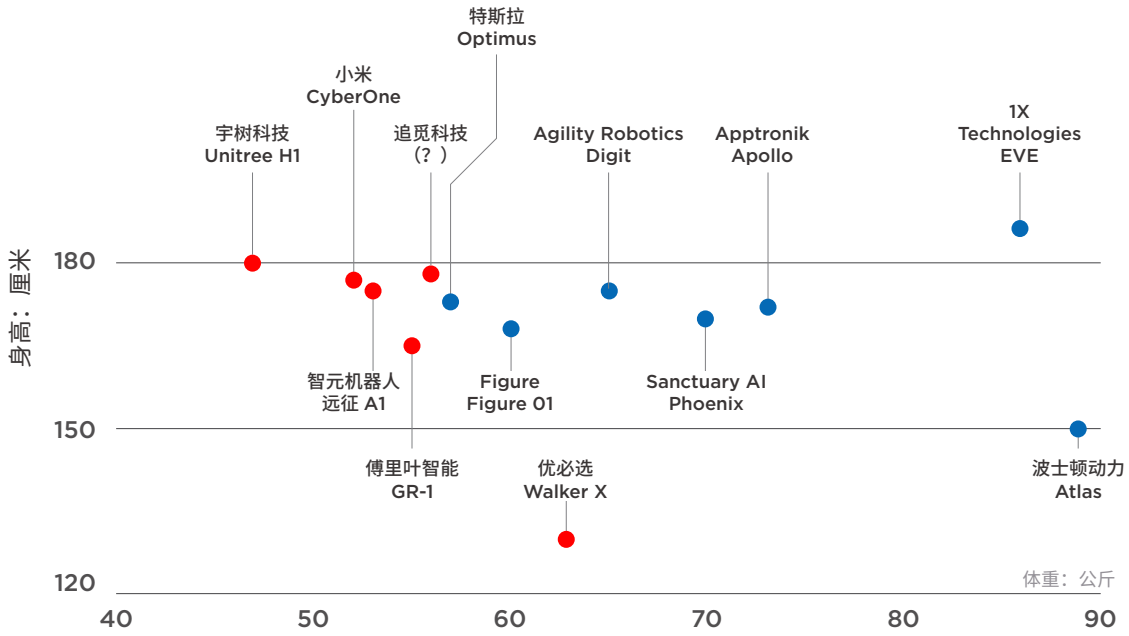
生成式人工智能正在加速人形机器人的生长。技术、市场与政策一起发力，催生人形机器人进入量产时代，比汽车更便宜。人形机器人将是真正制造业强国的标志。

所有的生物，都是通过身体体验这个物理世界，逐步产生智能的。人形机器人要成为真正的智能体，也要经过这一步。今年，谷歌先后发布了能理解视觉语言的 PaLM-E 等多模态大模型，让机器人将视觉转化为行动的 RT-2；微软则发布了“ ChatGPT for Robotics ”，允许人类用熟悉的自然语言对机器人下达指令。

人形机器人本质是通用机器人：它可以适应多种环境，执行不同的任务；稍加学习，还能做得越来越好。具身智能泛化了人形机器人的能力。通用机器人还意味着在人类社会即插即用。不像上一代机器人，为了让它更好地工作，人类还得花一大笔钱，为它修建标准化的场地，制造专门的工具，甚至还要让人类离得远远的。

生而为“人”

今年以来亮过相的人形机器人的身高与体重



来源：公开资料，未尽研究
说明：仅列举今年以来公开亮相且有相关数据的人形机器人。不含高校研发的暂无商业化计划的人形机器人。不包括四足机器人、轮腿机器人或机架机器人。部分公司迭代了多款人形机器人，仅列举最新型号。

这就是为什么现有的人形机器人，身材与造型基本与人类相仿。这不仅让它们看上去更亲切，更让它与人类社会无缝交互。而且，这也能更好地形成人类与机器人之间的直接映射，让训练与反馈更具体。

市场正在探索，拥有了泛化场景的感知、理解与决策的“大脑”的最小可行（MVP）的人形机器人是什么样子。一条路线更侧重稳健有力的双足，一条路线更侧重灵巧精密的双手。两者都需要负责运动控制的“小脑”与刚柔耦合的“肢体”。中国希望到了 2025 年核心部组件安全可控，到了 2027 年供应链体系安全可控。

二十多年来，人形机器人技术专利的申请量逐步上升，重心逐步从下肢结构与步态控制，转向手臂结构及其运动控制，且有进一步增加的趋势。今年很多“具身智能”的演示，几乎都是机器臂完成的。“人”就是这样的生物，大约 65% 的工种需要移动，其中 20% 需要两条腿来完成；高达 98.7% 的工作，需要灵巧双手精细操作。

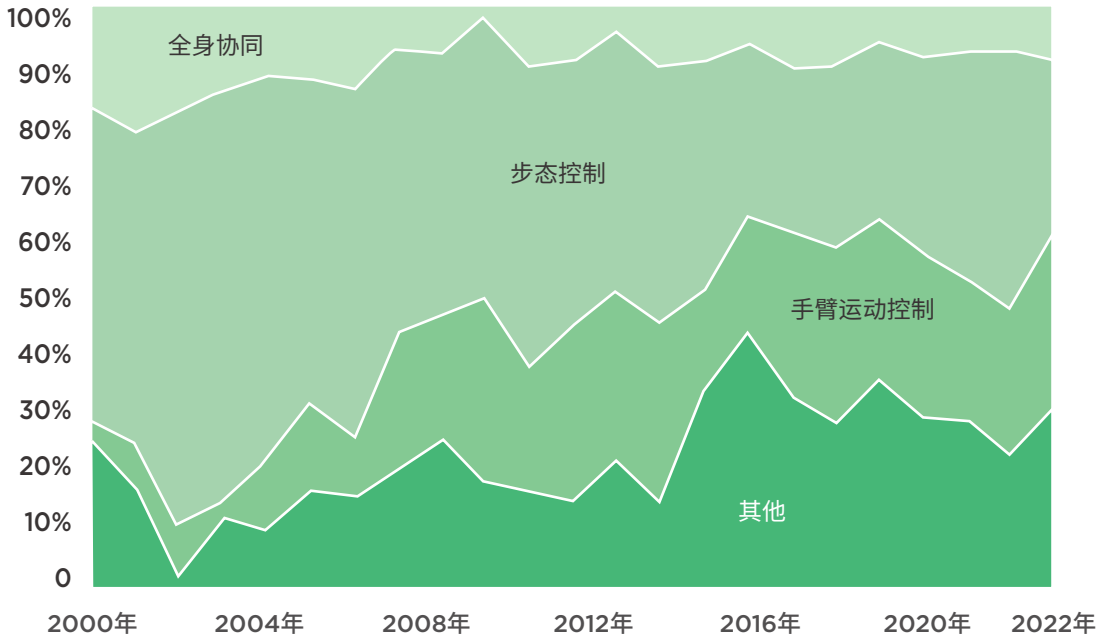
市场尝试拥抱人形机器人。它不知疲倦，没有人口危机，长期来看，单位时间成本更低。今年，亚马逊在自己的物流仓库试用了人形机器人 Digit。Digit 计划于 2024 年量产。第一条生产线是位于俄勒冈州的 RoboFab，年最大产能 1 万台。

Digit 原型的成本高达 25 万美元。特斯拉希望 Optimus 成本降至 2 万美元。Figure 创始人认为没有理由做不到：一个人形机器人，大约 1000 个零件，重量 70 公斤；一辆电动汽车 1 万个零件，重量 2000 公斤左右。秘诀在于规模量产。根据经验曲线，每当量产翻倍，成本有望下降至少 15%。

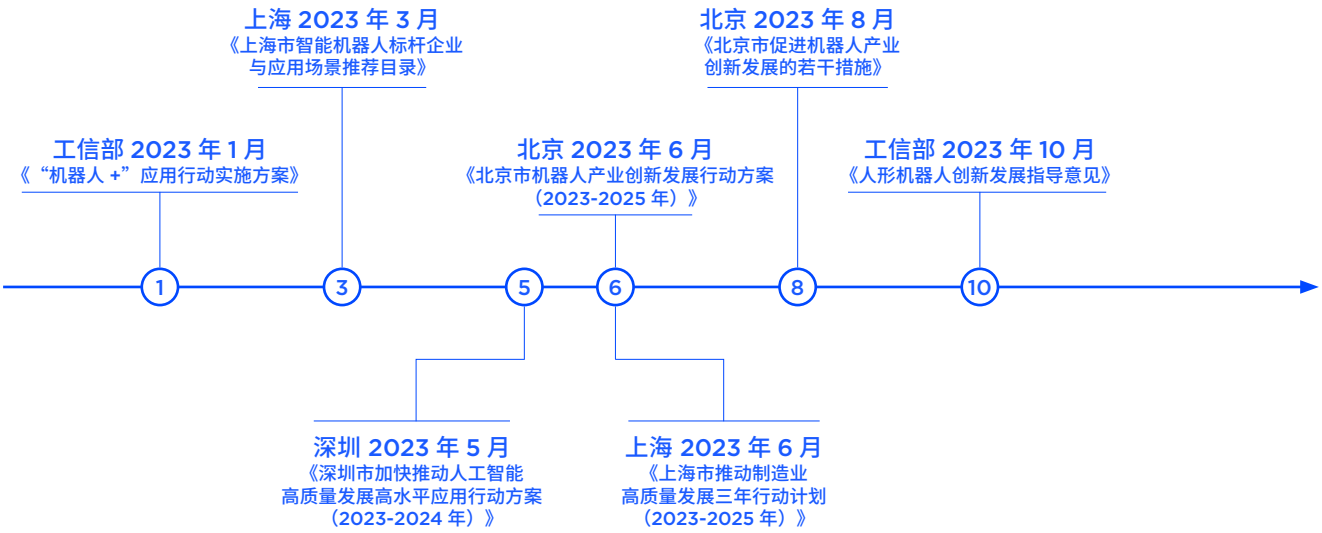
人形机器人技术与动力电池技术、自动驾驶技术等日益融合，核心供应链高度复用。这也是为什么汽车企业热衷人形机器人。除了特斯拉，Ashimo 来自日本本田，波士顿动力被韩国现代收编；小鹏发布 PX5，比亚迪投资智元，小米则同时宣布了汽车与 CyberOne。

只有少数几个国家，具备量产商用人形机器人的条件。中国是世界上最大的机器人市场，国际机器人联合会（IFR）称，供应链企业不断在中国增加产能。此外，应用市场的繁荣，为训练“具身智能”提供了更充沛的高质量数据。

人形机器人驱动控制技术专利占比趋势



2023 年中国密集出台鼓励政策



今年以来，中国各级政府正在引导创业者与投资者抓住机遇。《人形机器人创新发展指导意见》是全球第一部由政府出台的顶层设计文件。京津冀地区早已行动起来。北京设立 100 亿元规模的机器人产业基金；河北省 20 亿元的机器人产业基金成立；唐山成立 50 亿人民币的机器人产业基金。长三角与珠三角地区也有相应政策法规，那里产业集群密集。

设计和商业化下一代人形机器人的竞赛正在进行中。它在人类社会的渗透曲线，将与电动汽车相似。特斯拉先后在 2008 年与 2012 年开始交付 RoadSter 与 Model S。2024 年，将是人形机器人的“RoadSter 时刻”，卖的不多，但完成了商业化验证，为 2027 年的“Model S 时刻”蓄力。这一次，它可能会首先发生在中国。

参考：人形机器人技术专利分析报告；人形机器人创新发展指导意见；WR Industrial Robots 2023；WR Service Robots 2023；RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

从元宇宙到“新现实”

更多科技巨头卷入空间计算平台竞争，苹果硬件优先与 Meta 用户优先初次对决。

巨头们很少再提元宇宙。苹果与 Meta 似乎已经心照不宣地达成共识，空间计算才是最近的入口。更自然的交互方式，很快将与多模态的生成式人工智能合流。

今年，苹果与 Meta 先后展示了 Vision Pro 与 Quest 3。它们兼具 AR（增强现实）或 VR（虚拟现实）的功能，具备直通显示（VST）技术，通过摄像头与传感器，让人“透过”高分辨率的屏幕，看到身边的物理世界；它交互更加自然，能捕捉手势或目光的细微移动。

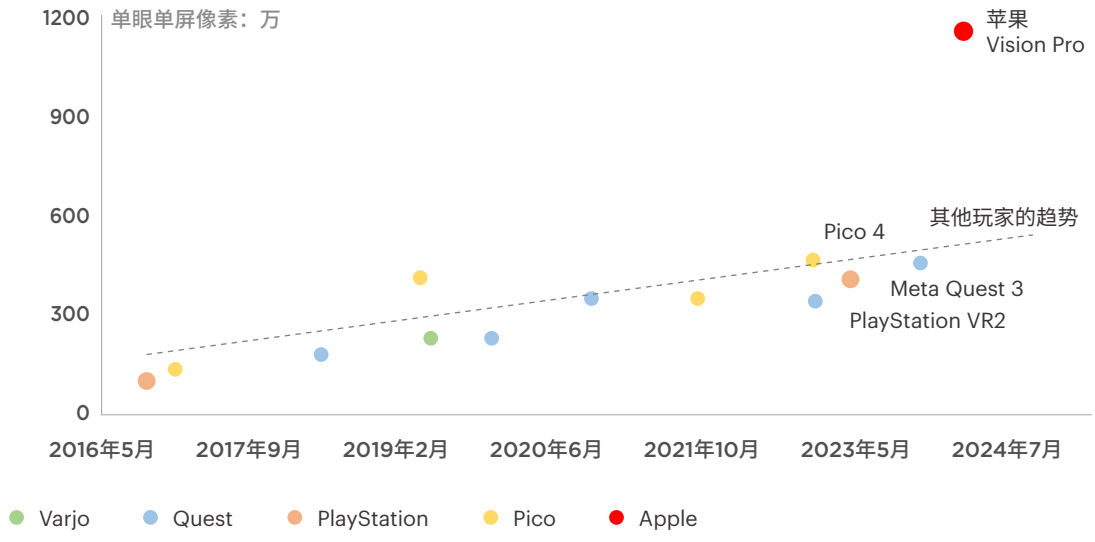
数字世界正在融合物理世界，构成“新现实世界”。先在显示屏上，实现对现实世界的增强与交互，是巨头们追逐智能眼镜的退而求其次。尽管大方向正在空间计算收敛，但是苹果与 Meta 选择了不同的道路。

硬件创新是苹果的重点。苹果决定从合适的设备开始，然后回到合适的价格。首发的 Vision Pro 定价高达 3,500 美元，拥有 12 个摄像头、多种传感器，成熟的 M2 芯片，全新的 R1 芯片，试图将从移动计算时代走向空间计算时代。只有这样的配置，才能驱动苹果定制的那两块 Micro OLED 屏幕。它们的分辨率远超消费级设备的行业水平。这决定着它能成为生产力工具，还是停留在游戏与娱乐工具。

像素密度的提升，带来了更高的功耗和发热。苹果采用了注视点渲染（Foveated rendering）技术，跟踪用户的眼睛，确定他们的注意力集中在何处，在此基础上决定使用多少算力，多大范围内运行全分辨率。

Meta 押注用户规模与网络效应。最大的底气是每天活跃在它的社交网络平台的 30 多亿用户。Meta 试图从合适的价格开始，努力开发合适的设备，将其打造成新的注意力消费平台。

VR 越来越清晰



来源：公开资料，未尽研究
说明：选定品牌的选定 VR 机型。目前，与苹果类似分辨率的主要是 Varjo 年底最新推出的第四代设备，但它主要面向工业场景。这里统计的主要是消费电子产品。

作为 Meta 最成功的尝试，300 美元的 Quest 2 累计销售了超过 2000 万台。今年发布的 Quest 3，售价 500 美元仅为 Vision Pro 的 1/7，硬件成本只有苹果的 25%。但 Meta 几乎仍是亏本在卖硬件。

从现有的布局来看，苹果与 Meta 将在 2024 年展开错位的竞争。它们都有胜算，最主要的对手还是自己。今年，它们都没来得及直接参与大模型公司业务的竞争。

市场期待 2024 年人工智能的潜力在端侧爆发。年底，苹果发布了机器学习框架 MLX，简化研究人员在硬件平台设计和部署模型的过程。明年，苹果还将发布 iOS 18 系统，生成式人工智能有机会跑在端侧。最新发布的 M3 芯片，将为此提供算力。Vision PRO 是其他硬件的扩展。iPhone 15 Pro 已经可以为 Vision Pro 拍摄空间视频，期待更多第三方的多模态的空间视频生成技术。

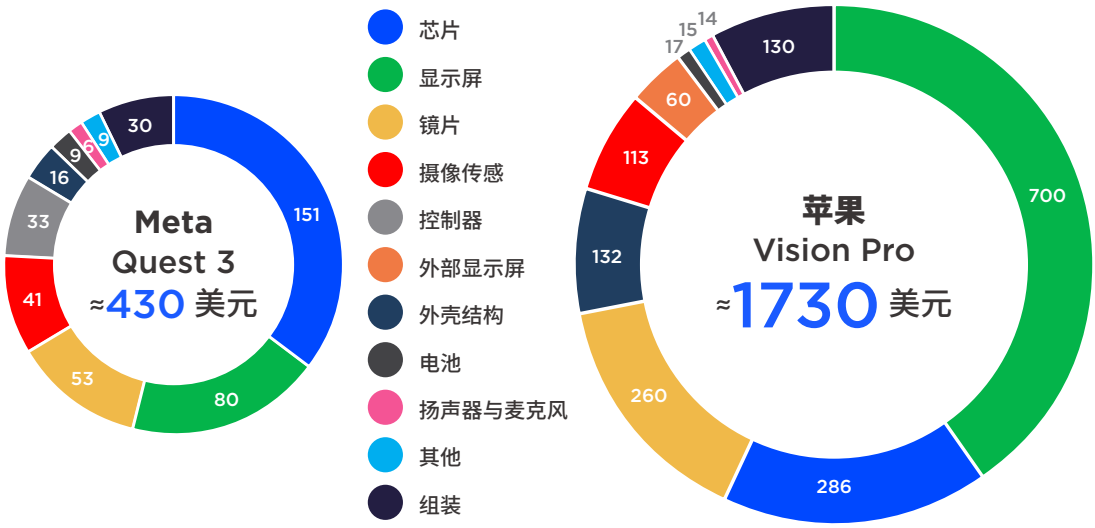
市场也期待生成式人工智能加持社交。Meta 的未来是“具身互联网”，人与人的社交关系，将进化为包含数字化身与虚拟角色的合成社交网络。生成式人工智能驱动的聊天机器人，将化身为地平线世界的 NPC，用户可以跨平台与之交谈。明年，Meta 将允许用户将游戏里的数字奖杯，“放置”在物理世界的书架上；甚至还可以将 Facebook 视频等其他应用挂在“墙上”播放。这是社交平台间的相互扩展。

2024 年将变得非常热闹。苹果早已动手开发下一代的 Vision Pro，Meta 还计划明年推出低配版的 Quest 3。谷歌、高通与三星结成了联盟，将阻击苹果独占高端市场。字节跳动旗下 PICO 仍在推进硬件升级，市场传言华为也将更新它的产品线。

尽管竞争激烈，但无论是谁，都能从对方的进步中获益。量产带来的供应链生态的完善，用户规模扩大；应用生态的繁荣，催生自然交互的原生应用，以及面向“智能家居”和“物联网”的交互。

市场正处于其短暂历史的关键时刻。明年会是“空间计算元年”，产能不足问题逐步缓解，应用生态摸透开发平台。2024 年，新品竞相上市，行业出货量增长会达到约 1200 万台，依然在为爆发铺垫。

下一代入口的成本对比



来源：Wellsenn Xr，未尽研究
说明：没有特殊说明，显示屏指设备内部向用户展示画面的屏幕，苹果的 Vision Pro 还有一块显示屏位于设备外侧前方，主要用于向周围人群展示用户的眼睛与表情等。镜片指光学相关的镜片与 IPD，芯片还包括冷却系统，其他包括包装与充电器等。苹果的 Vision Pro 没有控制器。

参考：苹果 MR 交互方式与内容开发研究报告；Taking the Hyperbole Out of the Metaverse；IDC：Worldwide Quarterly Augmented and Virtual Reality Headset Tracker

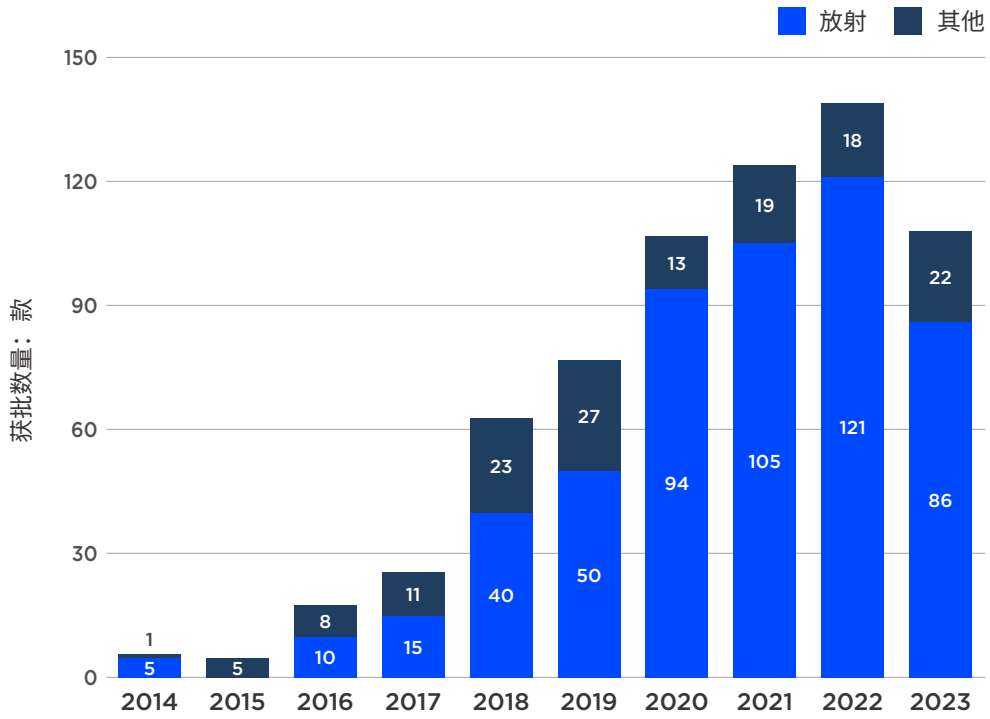
医疗智能体

大模型的“医生助理”，先从文书工作做起，专家模型逐步集成于医疗智能体。

医院永远人满为患。“鲍莫尔病”是医疗健康行业的顽疾。多年来，科技巨头数度高调进入这个行业，希望让服务变得像药品那样，可以规模化复制，提升可及性，降低成本。大模型、生成式人工智能以及智能体，是最近也是最有希望的一次。

上一轮人工智能热潮并非一无所获。近十年来，美国 FDA 批准了 500 多款支持人工智能或机器学习的医疗设备，截至今年 7 月，已经接近去年全年水平。其中，放射学诊断占了 75%。在今年的北美放射学会（RSNA）会议上，一半的讨论涉及人工智能。上月，阿里巴巴联合多家医院通过“平扫 CT+AI”，在 2 万多真实病例的回顾性试验中，发现了 31 例临床漏诊的早期胰腺癌病例。Nature 称基于医疗影像 AI 的癌症筛查，即将进入黄金时代。

近十年 FDA 批准的 AI/ML 设备

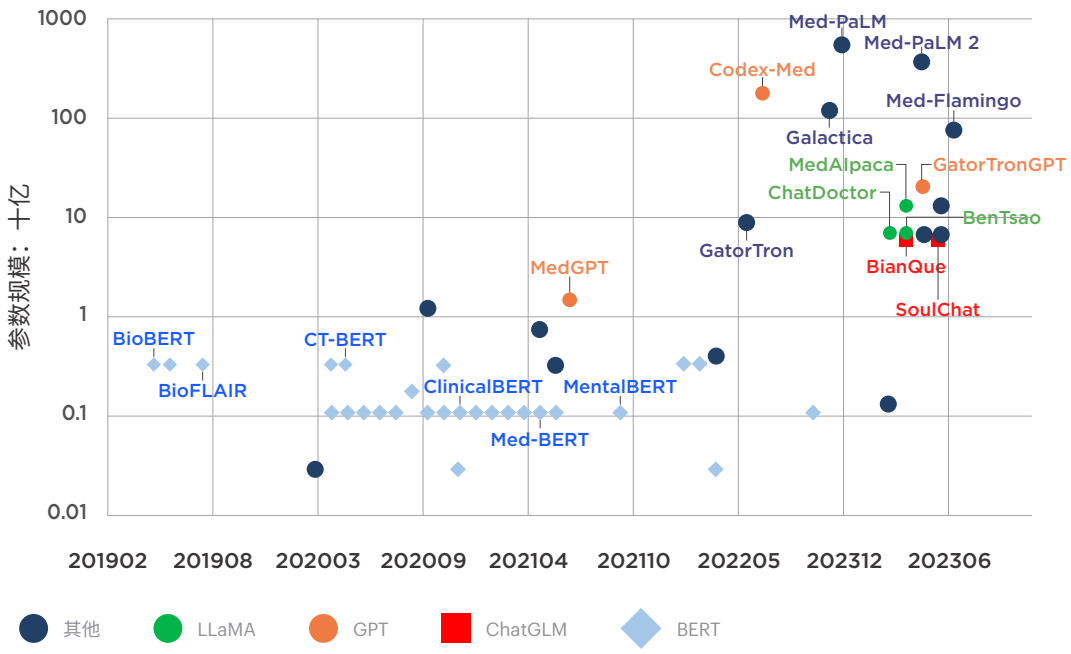


来源：FDA，未尽研究
说明：支持 AI/ML 的医疗设备。2023 年截至 7 月。

问题是不够泛化。已经获批的人工智能算法，往往倾向于专注于特定任务，而不是全面分析图像各种可能，或考虑到患者病史。原先的解决方案之一，是添加更多的人工智能工具。但这意味着算法过载。

医疗大模型的涌现，提供了新的解决方案。医疗保健数据，本质上由文本、图像和时间序列数据组成的，甚至可以把专业医生视为这些数据的“标注员”。今年以来，随着强大的预训练大模型尤其是开源模型相继问世，医疗大模型家族化演进，迭代升级加速。中国也是重要的参与者，扁鹊、本草大模型等相继问世。

医疗大模型“涌现”



来源：公开资料，未尽研究
说明：ALBERT 与 RoBERTa 等归入 BERT。不完全举例。部分未公开参数规模或基座的医疗大模型未予展示。

科技巨头已经构造了可以理解多种数据模态的全科医疗人工智能（GMAI），可以根据交互对象的不同，输出或专业或通俗的解释。谷歌的 Med-PaLM 2 是其典型，医学考试的表现基本接近“专家”，准确率达到 85%。关键不在于它能在何时取代多少专家医生，而在于它能惠及多少缺少顶级医疗资源的患者。谷歌搜索每天都会有 10 亿个健康相关的搜索，它们有需求获得更好的服务。

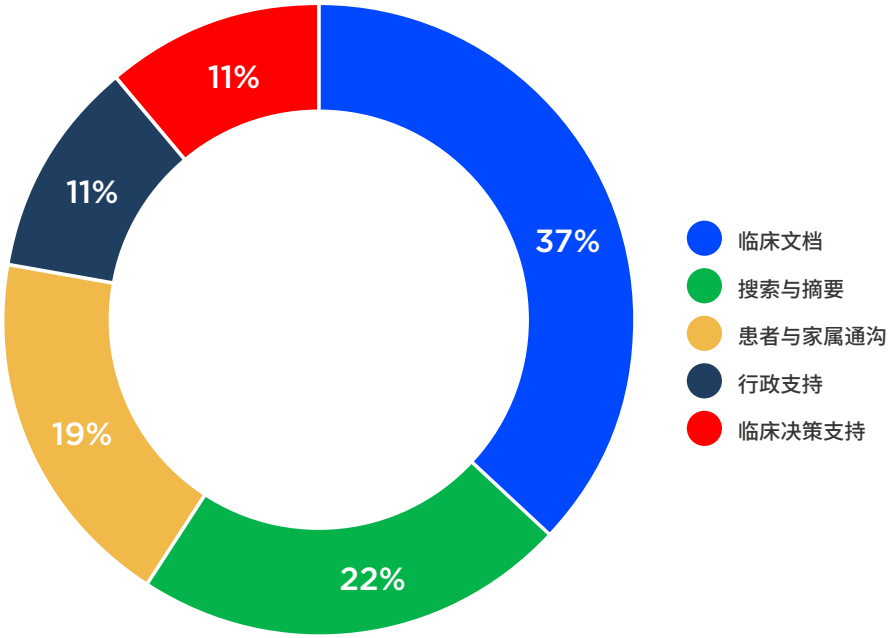
特定任务的专家模型也仍有其用武之地。分诊导航、辅助诊断、临床文档、预后追踪等这些相对较小的模型，甚至各类人工智能支持的细分科室的诊断工具，都可以集成到一个全面的人工智能平台中，智能体（Agent）就是它的中央调度中心，理解意图，分拆任务，调用模型，输出结果。最终，医生负责审核并给出最终方案。

短期内，生成式人工智能用于医疗服务，监管阻力最小，确定性最高的应用场景，是扮演医生的“文书助理”。医生希望人工智能带走他们工作中最无聊和最乏味的部分。

美国数十家综合医疗卫生系统正在试点的生成式人工智能应用，几乎主要面向临床医生提供服务，作为他们的助手，在接诊过程中，捕获与患者的对话，记录符合规范的电子病历；搜索病人的既往病史与检查结果、临床指南手册、临床试验机会等；对即将接手的护士给出注意事项摘要；向患者解释术语，叮嘱按时按量服药；还可以生成转诊、出院文件等。

它们的顶级开发者包括微软与谷歌等科技巨头，以及 Epic 这样的医疗软件巨头。今年，微软宣布将 GPT-4 集成到 Nuance 全新的 DAX 平台，减少了 50% 的临床文档记录时间，并与 Epic 合作，将生成式 AI 工具集成到后者的电子健康记录系统中。

美国医院中的生成式 AI



来源：STAT，未尽研究
说明：美国大型卫生系统公开披露的处于试点阶段的生成式人工智能服务。不完全列举，因为许多卫生系统选择不在风险更高的早期阶段就分享试点情况。

科技巨头已经展开全面竞争。亚马逊推出了自动生成病历的 HealthScribe，谷歌则与梅奥诊所等测试类似的生成式人工智能工具。百度的灵医大模型，以及腾讯医疗大模型也围绕病历等场景展开。

还有很多技术需要突破。谁掌握了提示技术，谁获得更好的答案，这在医疗服务领域不可接受；要更好地辅助诊断，大模型要更多地输出针对性的“追问”，而不是一味扮演“回答”者的角色。

2024 年，生成式人工智能扮演的“医生助理”，会先从文书工作与调度工作做起，逐步积累数据与经验，赢得医生与患者的信任；同时，继续与人类专家合作，探索更前沿的诊断与治疗技术。

主要参考文献: *A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics；Towards Generalist Biomedical AI；Generative AI Tracker: A guide to the health systems and companies driving adoption；人工智能大模型赋能医疗健康产业白皮书；中国医学生培养与学生发展调查报告*

“通用”基因编辑

更“通用”的基因编辑工具，更“泛化”的适应症，陆续进入临床，人工智能加速临床前研究。

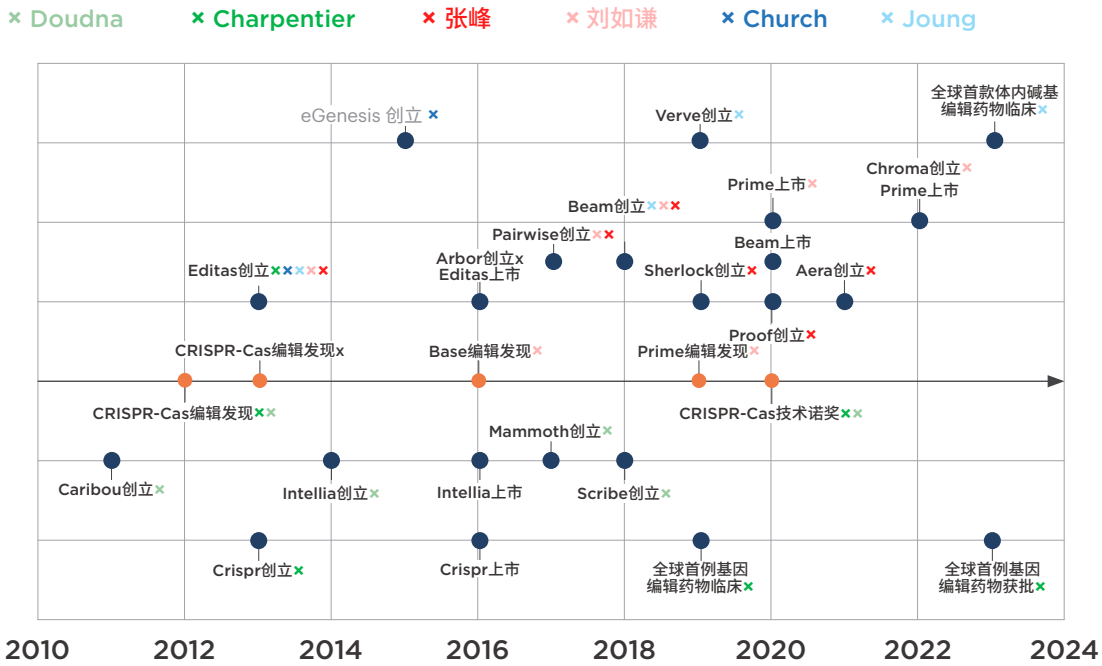
年底，全球首款 CRISPR 基因编辑药物先后在英国与美国获批，预示着药物研发“可编辑”的时代，已经正式到来。它是技术、资本与监管合力创新的典范。

第二场革命也早已酝酿，它的获批开启了下一个十年。它重塑着人们认识疾病与认识药物的方式，它治疗疾病的原因，而不是症状。只要它能在基因组的某个位置编辑，那么就有理由相信它的成功可以复制。何况测序硬件、基因组数据、人工智能算法的创新，仍在迅速扩展基因编辑的工具库。

2012 年，科学家 Emmanuelle Charpentier 和 Jennifer A. Doudna 发现，CRISPR 系统可以精确定位并剪切任何物种的任何基因。2020 年，两人因此获得了诺贝尔奖。张锋证明了它在实验室中的惊人能力。三位科学家各自迅速成立公司，改进技术，寻找应用场景。资本也积极响应。十年来，张锋旗下公司至少达到了 7 家，Doudna 达到了 4 家，尝试用它来满足药物递送、临床诊断、治疗的需求，甚至农作物改良。

基因编辑十年竞速

事件背后大佬



来源：公开资料，未尽研究
说明：基因编辑公司仅列举以上 6 位创始人参与创立的，公司其他联合创始人未列出。Doudna 在创立了 Editas 后离开。

CRISPR Therapeutics 赢得了竞速的第一程。与它同时获批的，还有蓝鸟生物（Bluebird Bio）基于慢病毒载体的基因疗法，也针对同一患者群体，但接到了 FDA 提示风险的黑框警告，获批当日，股价暴跌 40%。医药市场的逻辑，已经彻底改变。单次给药，终生治愈，意味着一旦获批，就会完全占领市场，除非后来者效果更好，或者风险更低。

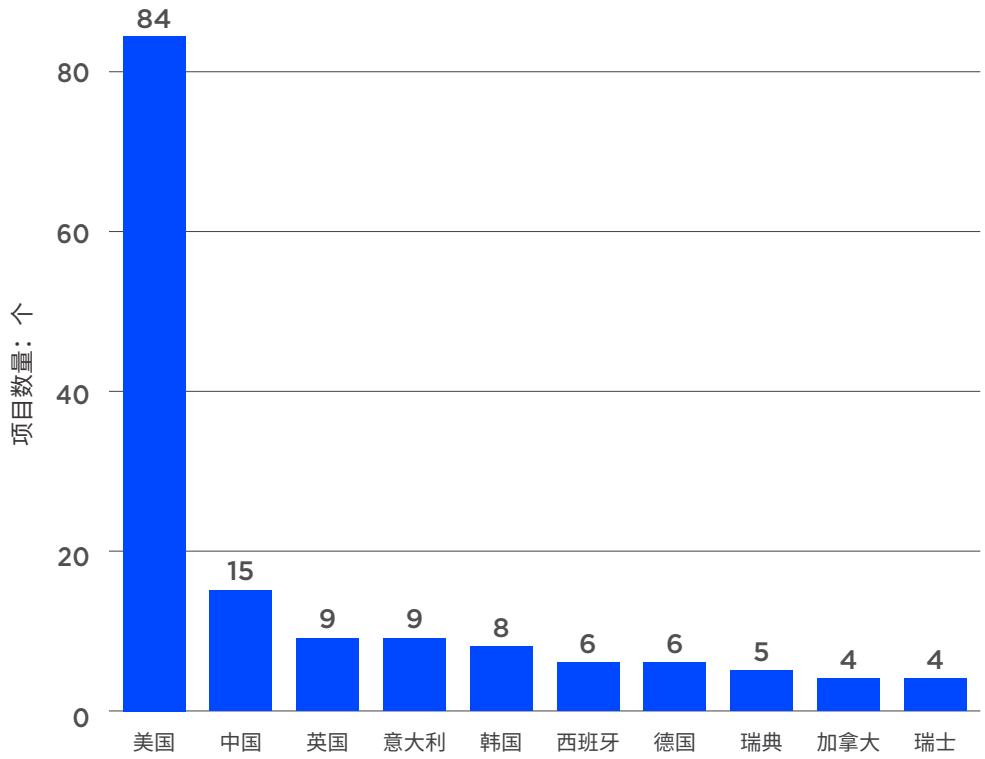
第一款获批的基因编辑药物，迈出了重要但不完美的第一步。标准的 CRISPR-Cas9 方法，实质上是对靶向基因的破坏，相当于将“编辑”功能限定在“删除”上。正如只有很少的情况下，才能通过划掉某几个单词，来纠正整个文本的错误，治疗大多数遗传性疾病，仍然需要更广泛的“编辑”功能，增加或替换某些单词。刘如谦先后发明了了基于 CRISPR 的碱基编辑（Base Editor）与先导编辑（Prime Editor），可以精准插入或替换单个或多个单词。

另一种缺陷是药物的可及性。这款新药非常昂贵，高达 220 万美元。它需要从患者体内获取细胞，然后在患者体外，用 CRISPR 工具编辑纠正它，再注入患者体内。为了给新细胞腾出位置，患者往往还需要先进行化疗，破坏骨髓。整个流程环节复杂，难度巨大。

降低成本的一种思路，是在更“通用”的细胞上，进行基因编辑，以便它们可以用于治疗许多不同的患者。但在体外生长和维持细胞安全稳定，成本高昂。更直接的思路是在患者体内编辑，难度在于 CRISPR 药物必须被递送到这些目标细胞。用于 mRNA 疫苗的 LNP 递送技术在这里也有用武之地。科学家还在不断尝试新的递送工具。

目前，全球近百项涉及 CRISPR 疾病治疗应用的重要研究正在进行或已经结束。美国占了绝大多数，中国居于第二。很多研究已经进入临床。据药明康德统计，中国的博雅辑因、邦耀生物、瑞风生物，同样瞄准地中海贫血；本导基因与中因科技主要聚焦于眼部疾病。

CRISPR 研究项目进展追踪

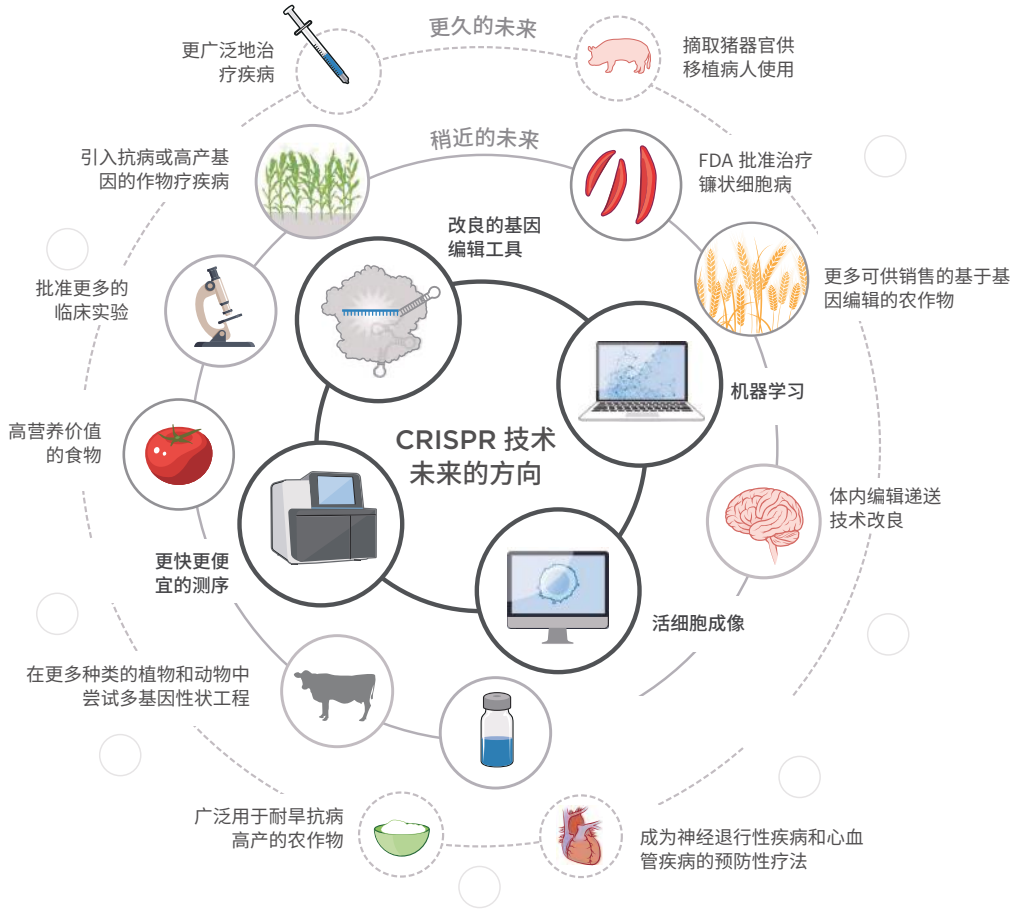


来源：STAT CRISPR TRACKR，未尽研究
说明：包括已经进入临床试验阶段的项目，也包括医疗机构、学术机构针对具体疾病的临床前研究，且论文发表于学术期刊、学术会议与公司公告，但不包括理论研究或文献综述。受披露等影响，并不代表世界上所有正在进行的研究。部分研究多国合作，重复统计。

大多数基因编辑临床试验，都聚焦于罕见病或遗传病，它们多由基因问题导致。但这样的患者往往分散、有限。收入捉襟见肘的企业，有时候不得不因外界风吹草动调整研发节奏。Church 尝试将基因编辑动物的器官安全地移植到人类身上。2024 年，异种移植的尝试还会继续。

今年年初，诺奖得主 Doudna 畅想，在十年后，CRISPR 编辑会适用于所有人，会先在农业世界中体验它。而稍近未来的“FDA 批准治疗镰状细胞病”已经实现。

AI+CRISPR：下一个十年



来源：Science. 2023 Jan 20

CRISPR 编辑正在进入 2.0 时代。这将主要由前述碱基编辑、先导编辑与表观基因组编辑（Epigenome Editing）引领。机器学习也在发力，张锋研究了新算法，搜索海量基因组数据，发现了 188 种新型 CRISPR 系统，丰富了工具库。此前，研究人员确定了 6 种 CRISPR 系统，CRISPR-Cas9 是最常用的一种。

《自然》杂志选出了 2024 年最值得关注临床试验。其中，Verve Therapeutics 首例进入临床的体内的碱基编辑药物，有望取代日常服用的降胆固醇药物。Prime Medicine 计划在明年寻求针对慢性肉芽肿病这一致死性遗传疾病的临床试验。表观基因编辑的 Tune Therapeutics，刚展示了乐观的临床前实验数据。

2024 年将是 CRISPR 编辑 2.0 时代的开端，大佬们的竞速将鼓舞更多创新者涌入，无论它们成功，还是短暂失利，都在推动行业“编辑”出更精准、更安全、更便宜的药物。

主要参考：*CRISPR technology: A decade of genome editing is only the beginning*；*CRISPR 2.0: a new wave of gene editors heads for clinical trial*；*CRISPR TRACKR: Follow the latest developments in genome editing*

中美风投再分岔

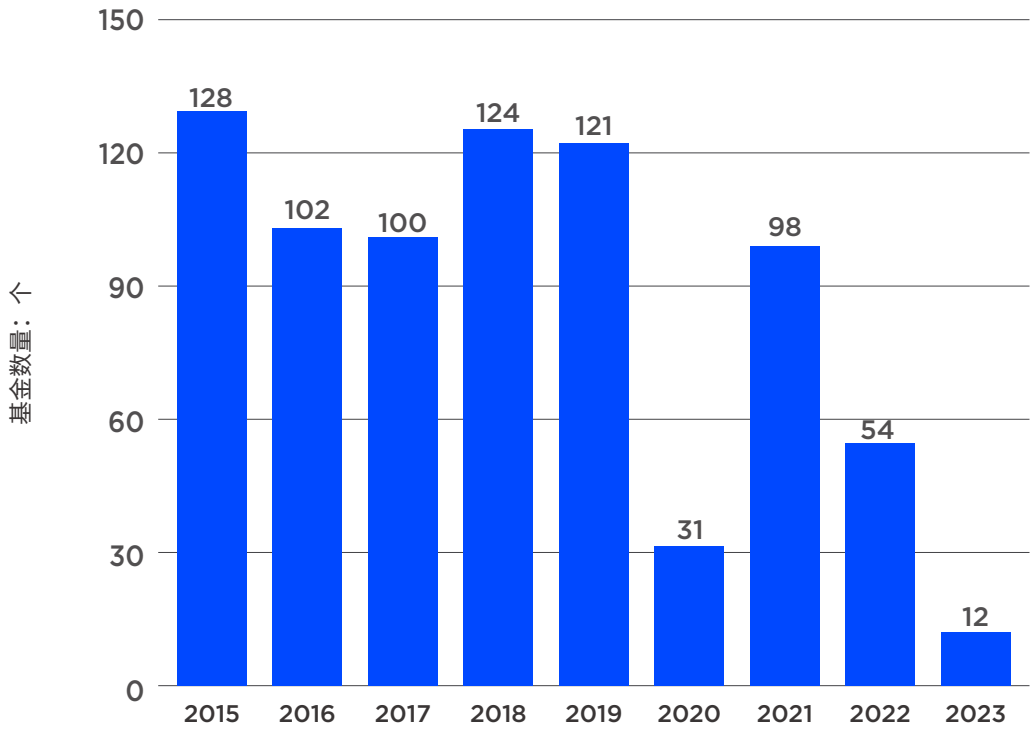
最寒冷的冬天过去，中国与美国创投生态各自主导力量继续分化和强化，中国风险资本开拓多元化资金补缺。

2023 年，全球风投市场整体入冬，募资与投资金额双双下降。作为全球最大最活跃的两大市场，中国与美国的创新生态已然迥异，不仅表现在投资方向上，还表现在资金来源上。

形似的是，大玩家在塑造两国的创新生态中，扮演了越来越重要的角色。在中国，城市主导下风投产业综合体崛起，创新“南下西进”；在美国，科技巨头围绕自身产业布局，风险投资交易逆势上扬。

中国本土风投机构（VC）越来越难获得美国有限合伙人（LP）的资金承诺（commitment count）。今年已经结束募集的中国风投基金，出资人来自美国的，较近十年来的峰值下降了约 90%。中国本土的初创企业也越来越难从美国风险投资基金那里获得融资。今年 8 月，美国总统拜登签署行政令，打着国家安全的名义，实施“反向 CFIUS”，限制美国企业投资“被关注国家”的半导体与微电子、人工智能与量子信息技术。

中国 VC 难以获得美国 LP 承诺

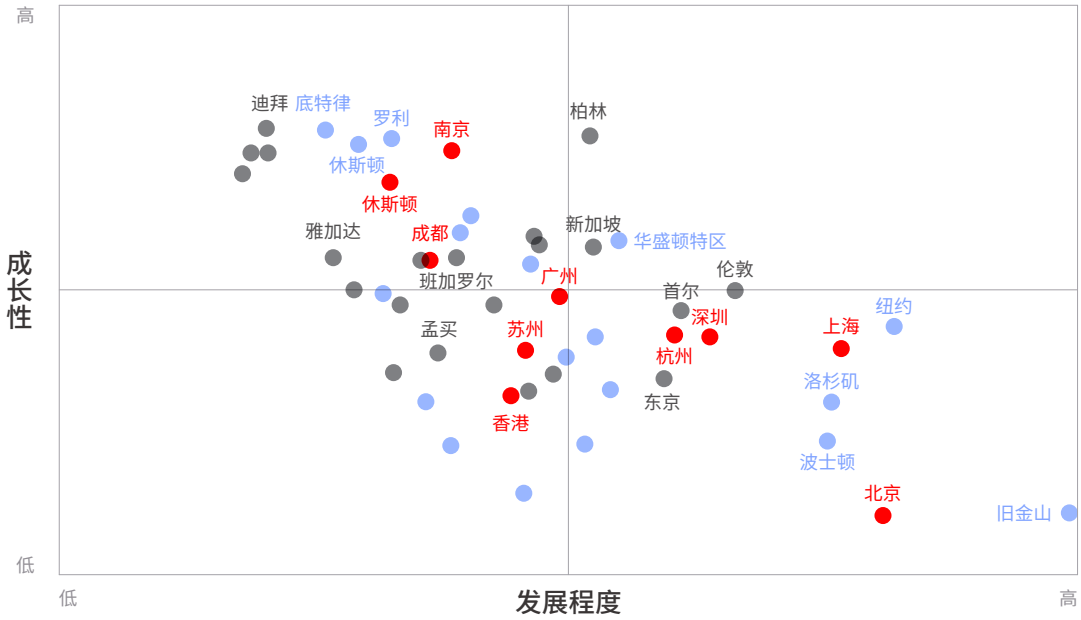


来源：PitchBook，未尽研究
说明：总部位于中国大陆与中国香港的风险投资机构，已经关闭、清算或完全投资的基金数量。截至 2023 年 11 月 15 日。

红杉资本已将其美国和欧洲、中国、印度部门，拆分为三个独立的实体。最近，美中战略竞争特别委员会要求其提供十多年来在华投资限制领域的详细信息。纪源资本、金沙江创投、高通创投和华登国际等多家知名跨境投资的风投机构也有类似境遇。

中国本土风投市场已经是人民币基金的天下，财政资金与产业资本主导市场。政府引导基金目标规模从 2017 年的 9.5 万亿元人民币提升至 2023 年年中的近 13 万亿。多地相继印发创投基金与产业基金发展管理办法，官方“以投促引”正在进一步下沉，中西部省份新成立基金数量逆势增长。南京、长沙与成都，成为中国风投生态成长性最高的城市。长期主义的社保基金也亲自下场，作为单一有限合伙人，在北京、上海出资设立专项创投基金。

全球风投生态 Top50



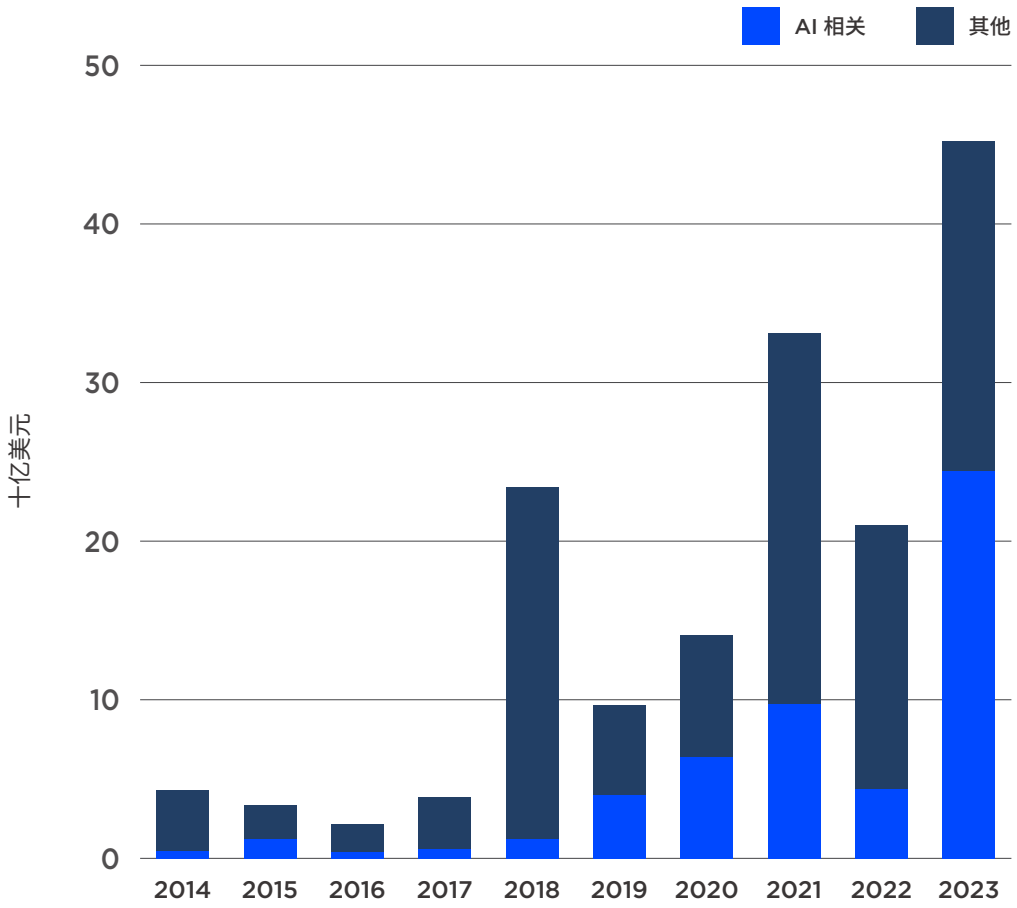
来源：Pitchbook，未尽研究
说明：2017 年至 2023 年。机构募资、企业融资、上市退出金额等是规模（Size）指标，以及独角兽企业数量，非传统风投交易占比等成熟度（Maturity）指标，反映了当地风投生态的发展程度。这些指标在不同时期内的增长势头，代表了当地风投生态的成长性。

中国的投资者与创业者也在奔赴中东，寻求募资来源多元化，蔚来汽车与小马智行是其典型。但沙特等中东主权财富基金，自身转型叙事色彩浓厚，与之磨合尚需时日。在第 28 届联合国气候变化大会上，阿联酋宣布要成立 300 亿美元新气候基金，把这笔钱花在能源转型、工业脱碳以及气候、技术等领域上。

美国的初创企业并不好过。今年至少有 3200 家有融资记录的初创公司倒闭，投向它们的 272 亿美元风险资本打了水漂。软银和老虎环球等曾在热钱汹涌时蓬勃发展的非传统风投机构，已经从最活跃玩家的排名中消失了。

来自“七巨头”的风险投资交易逆势上扬，2023 年至今已经超过了 400 亿美元，为历年最高，其中，超过 240 亿美元投向人工智能相关领域。微软、谷歌、亚马逊与英伟达是最活跃的企业巨头。

七巨头企业风投逆势上涨



来源：PitchBook，未尽研究
说明：苹果、特斯拉、英伟达、微软、谷歌、亚马逊与 Meta 及其旗下企业风险投资机构所投资的初创企业的风险融资总额。AI 相关既包括生成式人工智能，也包括以往的专家模型等方向的投资。

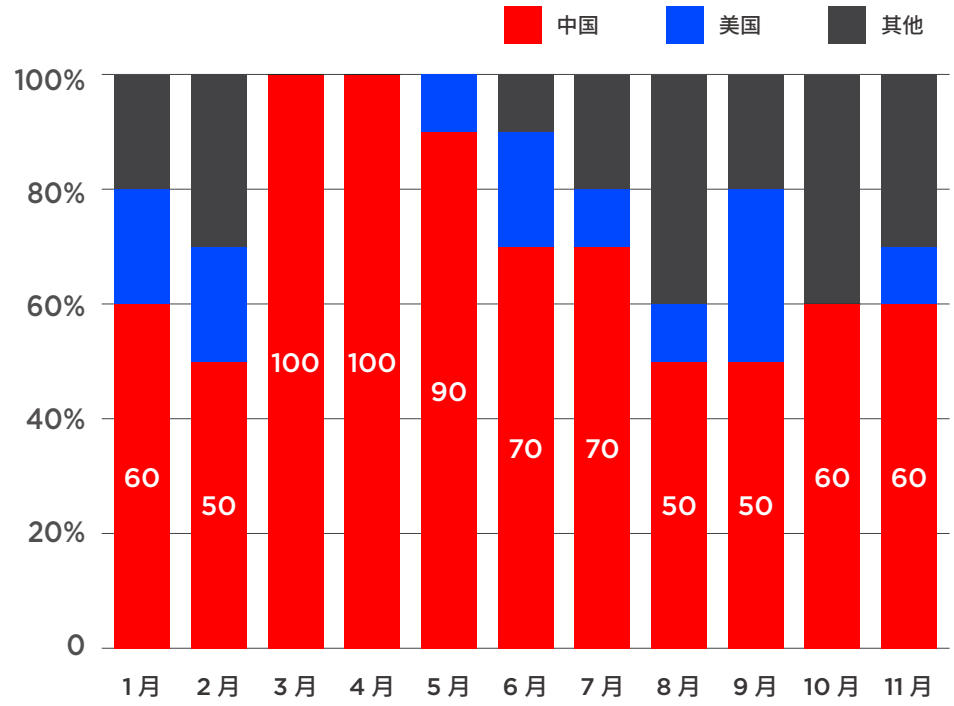
这笔钱既盘活了初创企业，又反哺了产业巨头。今年春天，英伟达加入了万亿美元市值俱乐部。二季度，该公司在早期创投市场出手了 8 次，三季度达到了 11 次，多数是人工智能初创企业，它们也是英伟达的客户，抢购芯片相当于“返投”了英伟达。微软公司至少投资了 8 家人工智能初创公司，它与 OpenAI 的 100 亿美元交易是今年规模最大的一轮。谷歌公司在人工智能领域达成了 16 笔交易，它旗下的风投机构 GV 达成了另外 13 笔交易。

无论是城市主导，还是产业巨头主导，都将推动资金、人才与技术的流动与配置。关键在于防范行政僵化或市场垄断带来的创新效率下降，要让传统的市场化机构发挥积极作用。

到目前为止，中国人民币基金的硬科技投资，成果颇丰。2023 年，全球每个月上市募资规模最大的初创企业，绝大多数来自中国，其中大部分都是半导体或新能源等近年来中国投资最热的硬科技领域。

2023 年全球初创企业 IPO 退出

当月前十大 IPO：企业总部所在地 / 所有



来源：PitchBook，未尽研究
说明：企业总部所在国家，而非企业上市地点。上市企业为风险投资或产业资本支持的初创企业。当月前十大 IPO 以当月上市企业募资规模排序。

在人民币基金较少涉足的消费等领域，美元基金也仍有发挥空间。中国的消费互联网创业者正在接管全球的购物车。今年，字节跳动收入超过腾讯，优势在于全球扩张，它的竞争对手已是 Meta；Temu 与 Shein 合计美国用户，已经逼近亚马逊，超越就在明年。

中国民营企业新一波出海大潮正在泛起，尤其是创业者出海。中国在电商、新能源、机器人、游戏、网络教育、AI 应用、生成式内容等方面具有优势，如果能建立起游刃平行市场的业务架构，不愁吸引不到美元投资。

美元资本呼吁对多元化投资的有限松绑。美国的“反向 CFIUS”监管，交给了与中国往来密切的耶伦，执行细则已经进入起草阶段，或在明年公布；美国国家风险投资协会（NVCA）警告说，过于严格的禁令，将让美国在全球竞争中处于劣势。

2024 年，人工智能仍将是全球创新的主题，中国与美国会是竞争最激烈的两个市场。风险资金将优先满足这种“紧张平衡”，大玩家主导的创新生态仍将扮演重要角色。但在非排它竞争领域，多元化来源的资金将继续在市场渗透，最寒冷的时候即将过去。

参考：US Venture Capital Outlook、Global Markets Snapshot、Global VC Ecosystem Rankings、大中华区风险投资报告、2023 年前三季度中国股权投资市场报告

推理的碳足迹

大模型训练和推理的能源成本，以及碳排放带来的环境成本，促进效率更高的模型训练和部署。

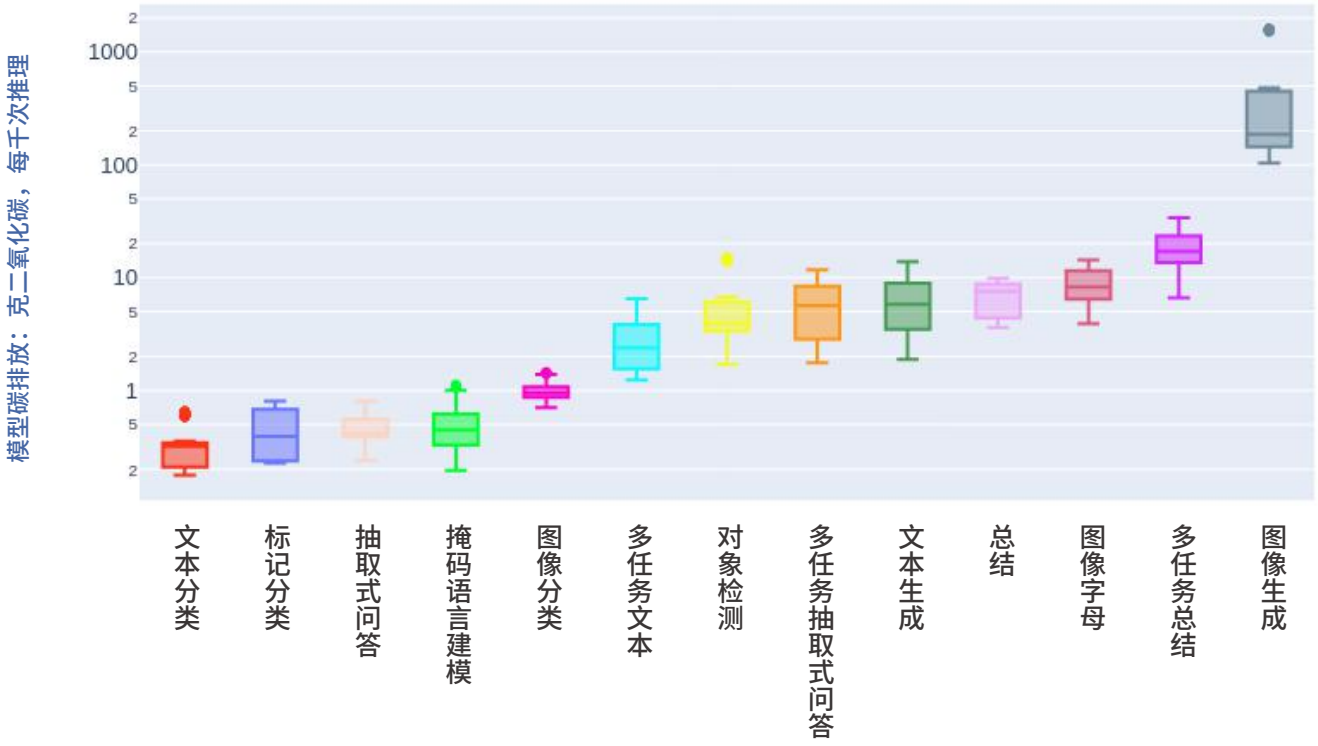
2024 年，我们可能会看到大模型参数数量的收益递减。训练一个千亿参数级的模型，在一些性能上可能比一个万亿级参数的模型产生的结果相差不大，但所需的计算能力能更有效地部署。而且在一些行业和使用场景下，百亿或者十亿参数级的模型，计算效率会更好。

单一的庞大模型是笨重且昂贵的，而一个专家混合体可能会几乎同样有效，它由更小、更具体的模型集合组成，可能还包括多模态模型。

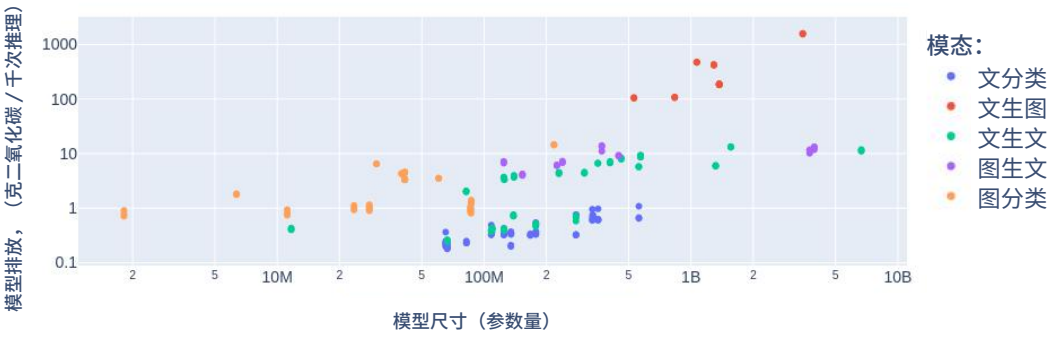
大模型在推理阶段的能耗，生成类任务比分类任务产生更多排放，多任务比单任务更多，生成图像比生成语言更多，更精确的推理、更泛化的任务也产生更多排放。

在自然语言处理和计算机视觉中都很常见在 5 种不同的模态中，选择十个机器学习任务进行分析，可以大致分析出大模型在推理阶段所产生的二氧化碳排放：

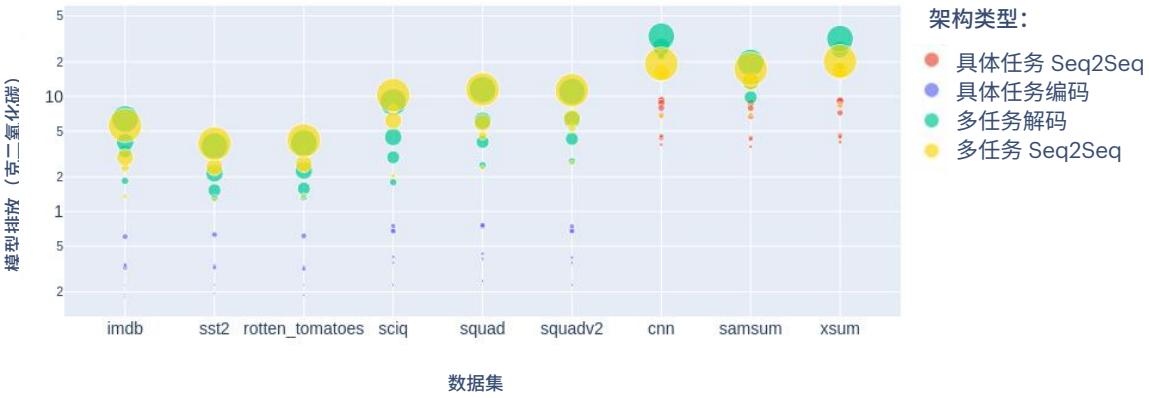
大模型不同任务的碳排放



来源：Hugging Face, 未尽研究
说明：在研究中检查的任务，和它们产生的碳排放量。执行千次。纵轴是对数标度。



多任务的模型比具体任务的模型产生更多的排放。



注：坐标进行了对数处理

值得注意的是，当模型变得更大，以追求智能的涌现、能力的泛化，不仅碳排放总量增加了，而且碳排放强度也增加了。似乎更多的智能来自更多的能源。

不同的推理任务，平均消耗的能源（及其统计标准方差）差别很大。每千次图片生成所带来的能耗强度，是千次数文本分类的近 1500 倍。

同一个模型家族，尺寸越大，不仅能耗及排放的总量越大，而且强度越大。如 700 亿参数模型的能耗强度，是 5.6 亿能耗强度的近两倍。

从 GPT-3 之后，模型的规模越来越大，日益多任务和多模态，更多面向用户的应用开发出来，而且推理的实时性和精确越来越强，所有这些，都意味着在通用人工智能的道路上走得越远，消耗的能源越多，产生的碳排放越多。这些环境成本应该考虑在内。

虽然更大的模型涌现出更多的智能，但是在 GPT-4 之后，继续扩大模型，在训练和推理的阶段，由于能耗的指数型增长，大模型的边际效益递减，综合成本收益是否划算，是需要考虑的一个问题。

为了解决碳排放问题，科技巨头已经成为全球最大的绿电采购方。苹果要求整个供应链 2030 年实现碳中和；微软制定了实现负碳排放的目标——2030 年碳中和，2050 年把历史上的碳排放欠债一并偿还；谷歌要在 2030 年实现 24/7 零碳电力。

目前“小”模型正在大量出现，它们结合具体的使用场景进行训练和部署，在具体功能上并不输闭源大模型，比庞大的模型计算效率更高；而更多模型部署到边缘侧和设备终端，会让推理更有效率。2024 年，应该是“小”模型的一年。

参考文献：*Power Hungry Processing: Watts Driving the Cost of AI Deployment?*Alexandra Sasha，Yacine Jernite , Emma Strubell，A systematic review of Green AI，Roberto Verdecchia, June Sallou, Luís Cruz

结语

巨头准备继续通吃。

由屠龙少年成长为巨龙，它们中间历史最长的苹果，已经近 50 年。时间最短的如 Meta，已经近 20 年。移动互联时代，十年左右时间，就可以从初创企业长出一家科技巨头。在生成式 AI 的技术浪潮中，会出现颠覆性的初创企业，以更快的速度成长为新的科技巨头吗？

目前被人最看好的是 OpenAI，2015 年成立，8 年之后估值已经在 900 亿美元左右。它拥有独特的企业架构，可盈利公司已经跻身估值最高的非上市企业之列，但非盈利公司拥有盈利公司。从 OpenAI 的“董事会政变”事件中可以看出，微软目前通过技术使用、云计算、投资等方式，一段时期以来左右着 OpenAI 的可盈利部分。OpenAI 的新董事会组建仍然没有完成，其新一轮融资寻求估值为 900 亿美元，新的股东结构和董事会组成，也会对 OpenAI 未来的发展产生影响。OpenAI 最初设立为一家非盈利公司，其目的就是为了不成为另一家硅谷的 Big Tech。它在利用资本，但最终不受资本控制，而是由一个捍卫 AGI for Humanity 使命的非盈利董事会行使“监护权”。值得注意的是，OpenAI 的竞争对手 Anthropic 也设立了带有社会影响力色彩的治理结构。马斯克创办的 xAI 也注册为一家赢利性共益企业（for-profit benefit corporation）。

生成式人工智能时代，能否出现在技术、产品和创新都能与科技巨头抗衡，而其社会影响力又大于科技巨头的企业，这是非常值得期待的。这需要在新的技术条件下的企业治理结构的创新，在社会影响力与股东价值之间取得平衡。从目前来看，硅谷的科技巨头和风险资本，也乐意投资这一类颇具技术颠覆性的企业。

人们开始担心已经赢得 IT 和互联网竞争的这几家科技巨头，最终将赢得这场通用人工智能的竞争，而初创公司的成长空间已经非常有限。

初创企业背后的科技巨头

来源：CB insights，未尽研究。
说明：截至 2023 年 11 月 27 日。包括来自科技巨头旗下 M12 与 GV 等企业风投机构的投资。亚马逊对 OpenAI 的投资指 AWS 参与了 OpenAI 的原始融资

颠覆性创新往往发生在初创公司。大公司并不缺乏好的创意，好的论文和专利，但一些研究表明，在技术采纳和应用方面，与初创公司相比，却效率较低。谷歌在与 OpenAI 的竞争中，充分体现出这一点。谷歌是全球高质量 AI 论文产出最多的地方，包括 Transformer 论文，但最终用 Transformer 做出最好模型的，却是 OpenAI。

初创企业的生态富有活力，会在多个点上快速创新。大企业的 R&D 团队与产品团队之间的割裂是其致命弱点。而优秀的 AI 企业，研发与产品团队总是一体的，所以能否做出产品，很快就会得出结果。而大企业这一过程比较迟缓，过于担心失败，或者在推向市场时，顾虑较多。

但科技巨头把手中掌握的最先进的基础大模型，与其本来就已经主导市场的应用结合起来，会轻易碾压做同类应用的初创 AI 企业，如在 SaaS 软件领域。行业巨头在生成式 AI 方面的投入转型也非常快，开源更让许多创新变得没有必要。因此，初创企业的生态位，很多会来自开源模型小型化过程，大模型的基础能力与垂直领域结合的部位，行业深度中蕴藏的数据资源，以及 AI 与硬件结合的产品与供应链能力。

如果 AI 最终让资源和技术向巨头们更加集中，它们将面对监管当局与社会的拷问，控制在少数巨头手里的超级人工智能，对经济增长与就业的好处在哪里?

科技巨头们坚持认为，它们提供了创业和创新的平台，如互联网平台和云计算平台等，还有大量的开源工具，降低了创业门槛，几个人就可以创办一家企业，目前已经开始出现十个人的团队就能创办一家独角兽企业。科技巨头也是风险资本的一个重要来源，它们对初创企业的收购，是风险资本和创始人团队可以退出的机会，其作用已经相当于上市 IPO。而退出往往带来丰厚的回报，这些资金中的大多数，又重新回到风险资本市场上去。科技巨头的资本力量，在创新生态中扮演日益重要的角色。

微软 CEO 纳德拉认为，这一轮 AI 带来的革命，不同于移动。移动带来消费的繁荣，而 AI 是创意者和建造者（builder ）的效率神器，它将显著提升劳动生产率。英伟达的 CEO 黄仁勋认为，企业提高生产力，就会雇佣更多的人，把企业做得更大，或者进入更多的领域。这样企业就会发展下去。但人们也会想起他的另一句话：买得越多，省得越多。

DeepMind 创始人哈萨比斯认为，通用人工智能正在引发一场科学的范式革命，可以用来解决科学难题，改变科学发现的方式。OpenAI 创始人奥特曼相信通用人工智能将会实现生产力革命的“奇点”，未来的问题并不是社会财富的匮乏，而是如何分配已经极大丰裕了的社会财富。技术理想主义者则想到用技术好的一面对付技术不好的一面。OpenAI 首席科学家和联合创始人苏茨克沃正在研究超级对齐的技术，内置于下一代超级智能之中。

在不同的体制内，人工智能可能释放生产力和破坏力的可能性、可控性、可控方式，将会有不同的表现。李飞飞所说的硅谷的 bro culture 及其所推动的技术加速主义，以及巨头可能掌握超级人工智能，已经引发越来越多的焦虑。而大国在 AI 领域的竞争，让这种加速无法放缓。国内监管、国际合作、以及中国与美国之间建立起人工智能的对话机制，成为目前试图让 AI 风险可控的初级框架。

美国游戏公司 Epic 在一起反垄断官司中初步胜诉了谷歌；英伟达对 GPU 供应的控制，包括英伟达在供应链等环节采取的一些排挤竞争对手的做法，正在引发一些国家的关注。

技术从本质上来说是加速发展的。颠覆性的技术，只有一种增长方式，即指数型增长。但这种不断加速的自动驾驶，仍然需要减速装置。

巨头会继续通吃吗?