生成式 AI

State of Generative AI

2023

启明创投 × 未尽研究

概要

如果说 2022 年被称为生成式人工智能之年，扩散模型应用取得突破，ChatGPT 出世，一系列开创性的研究论文发表，2023 年则把大模型推向了一个高峰，GPT-4 的发布，标志着生成式人工智能，进入了面朝通用人工智能创新应用的阶段。研究、应用、监管，合力开辟着生成式人工智能的发展之路。

创新应用

生成式人工智能的生态包括了基础设施层、模型层与应用层，创新在每一个层面发起，竞争也在科技巨头、行业龙头和初创公司之间展开。

在整个生态中，受益于以参数规模为代表的大模型不断扩张，算力目前是最稀缺的资源，也处于最容易获利的要津。算力是大模型成本结构中最大的一块，GPU 的性能，决定了这个新兴行业的步调。但是，GPU 性能提升的速度，已经落后于大模型训练和推理需求的增长。

面对这一革命性的技术，不论是主动还是被动，企业都被卷入其中。不管是技术的守成者、创新者还是采纳者，业务模式都将发生变化，进而影响企业的发展。

当前，生成式 AI 尚处于技术发展的早期阶段，基础架构和核心技术并不成熟；科技巨头忙于研发大模型，尚未顾及深度切入具体的应用场景。但巨头何时添加相似的功能（feature）始终是悬在初创企业头上的达摩克利斯之剑，而大模型能力边界的扩张也可能在未来挤占初创企业的发展空间，可以说，这是初创企业的蓝海，但也有航道下的暗礁。

在中国，目前从模型出发的公司受到看好，通用大模型和垂直大模型的创业如火如荼，而自建模型的应用也在努力构建着自己的壁垒；同样，科技巨头正在利用自身算力优势来构建大模型。我们有理由相信，在众多模型层和科技大厂的合力下，模型层的整体能力将进一步完善，在未来为应用层企业提供可靠的支撑。

前沿研究

生成式人工智能领域的一个突出特征，是研究与创新过程的密切结合，许多在企业内部实现，迅速推出用例和产品。这种研究与创业的一体化，初创企业和风险资本起到了重要的作用，而美国科技巨头和主要人工智能企业的研究投入、人才密集度、包括一些底层技术的研究，这些年来已经超过了大学等研究机构。

GPT-4 迸发出通用人工智能的“火花”，需要研究和解决的问题反而更多了，如信心校准、长期记忆、持续学习、个性化、规划和概念跨越、透明度、认知谬误和非理性，等等。而过去半年最重要的研究方向，是破解和理解大模型神秘而又令人兴奋的智能“涌现”。大模型既需要超越对下一个词的预测能力，也需要一个更丰富、更复杂的“慢思考”深层机制，来监督“快思考”预测下一个词的机制。

大模型不仅用来生成文章和图片，而且可以当成智能代理，帮助管理和执行更复杂的任务。开源模型实现了低成本、小型化、专业化的训练，与闭源的基础模型竞争互补，共同推动了生成式人工智能技术的应用，也加快了模型向边缘侧和移动端部署。生成式人工智能大模型日益向多模态发展，具身智能也成为一个重要研究方向，帮助生成式人工智能更好地理解和处理现实世界的复杂性和多样性。大模型更安全、让智能更可信，成为新兴的研究热点。生成式人工智能对于就业和经济的广泛影响，正在吸引经济学、社会学、心理学等不同领域的研究兴趣。但仍然需求实证性的研究。

监管、 安全与人才

生成式人工智能加快了中国、欧盟和美国的监管和立法的进程。欧盟努力在今年底让《人工智能法案》生效，为全球人工智能立法定下基调。中国也预计将于明年提出综合性的人工智能立法。而美国重点在于建立风险控制技术标准。

中国对通用人工智能表现出很大热情与期待。地方政府中北京、上海、深圳是第一梯队，均提出了较具雄心的人工智能科研、创新与产业目标。中国研究人员发布的论文在数量上已经超过了美国，但在金字塔顶端，无论是研究还是创业，美国仍然占据明显的优势 。

科技部要求人工智能企业应该接受科技伦理审查；审查主体应该设立科技伦理（审查）委员会。美国人工智能企业较早开始设立负责任与可信人工智能部门，从去年到今年以来经过一些调整，反映出在生成式人工智能发生变革之际，企业正在寻求用更好的技术和方案，来安全和负责地部署新技术。

十大前瞻

基于上述研究，报告对未来一至三年的大语言模型、多模态模型和商业竞争态势，做出了十点前瞻。

目录

第一章 行业变革

生成式人工智能的生态包括了基础设施层、模型层与应用层。创新在每一个层面发起，竞争也在科技巨头、行业龙头与初创公司之间展开。

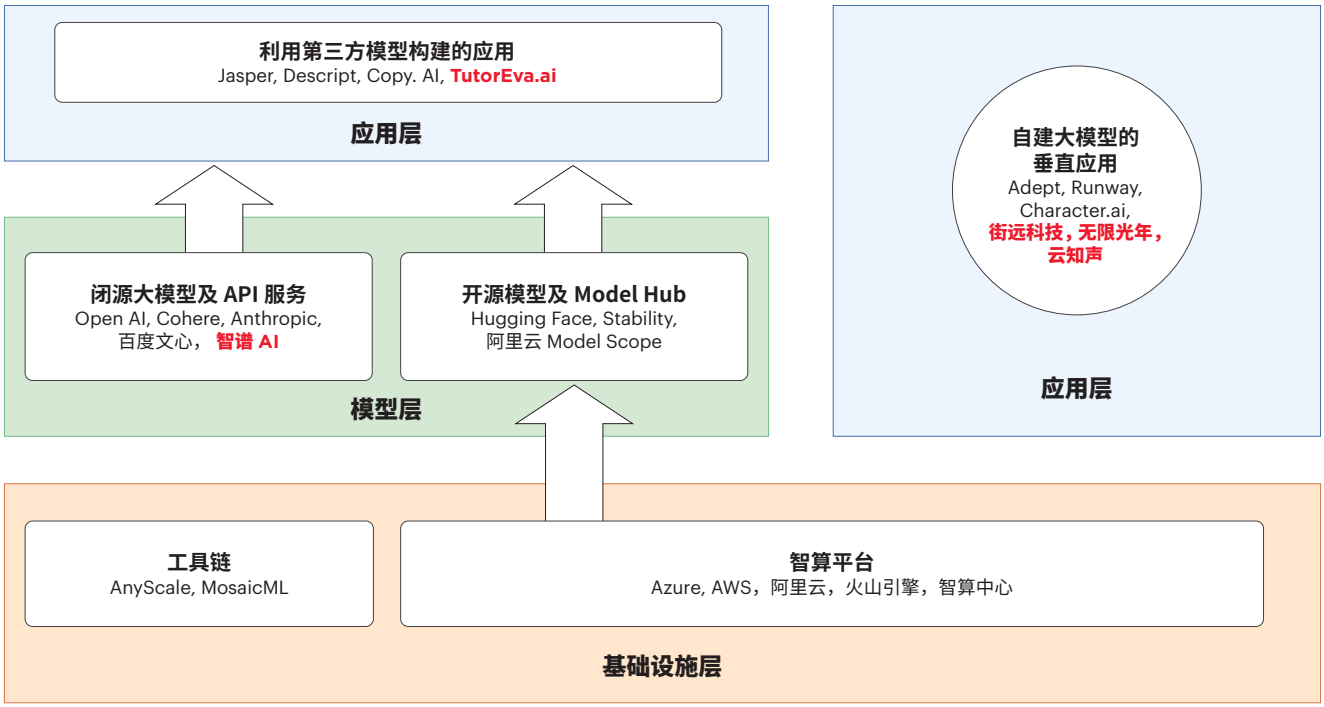
生态架构

四代底层技术的进步，催动了四波人工智能的发展。第一波小规模专家知识，用了 40 年走完；第二波浅层机器学习，用了 20 年走完；第三波深度学习，用了 8-10 年走完，并取得一定的成就。最近这一波 AI 新浪潮，以 2017 年基于 Transformer 的预训练模型为起点，并在 2020 年 GPT-3 大模型发布后突破技术奇点。

AI 1.0 时代，需要针对特定任务，利用相关的数据研发特定模型，任务和模型耦合。AI 2.0 时代，经过大规模数据预训练得到的大模型，带来了极好的效果和泛化能力，可以直接用于下游的各种任务。

AI 2.0 的公司将分为三层：

- **基础设施层：**解决大模型训练 / 推理 / 部署的工具链厂商和提供 GPU 资源的智算中心。智算中心再往下是新一代 AI 芯片或者下一代通用 GPU。
- **模型层：**研发大模型，并对外提供 AI 模型服务或者 API 服务，包括训练（training）和推理（inference）时需要的 GPU 资源。除了这类输出“水电”的底座大模型，也包括提供针对特定行业或场景的垂直模型的公司。
- **应用层：**专注于解决某个特定领域的应用公司，包括自研大模型的应用公司和利用第三方大模型的应用公司。



图中标红的企业为启明创投已布局企业。

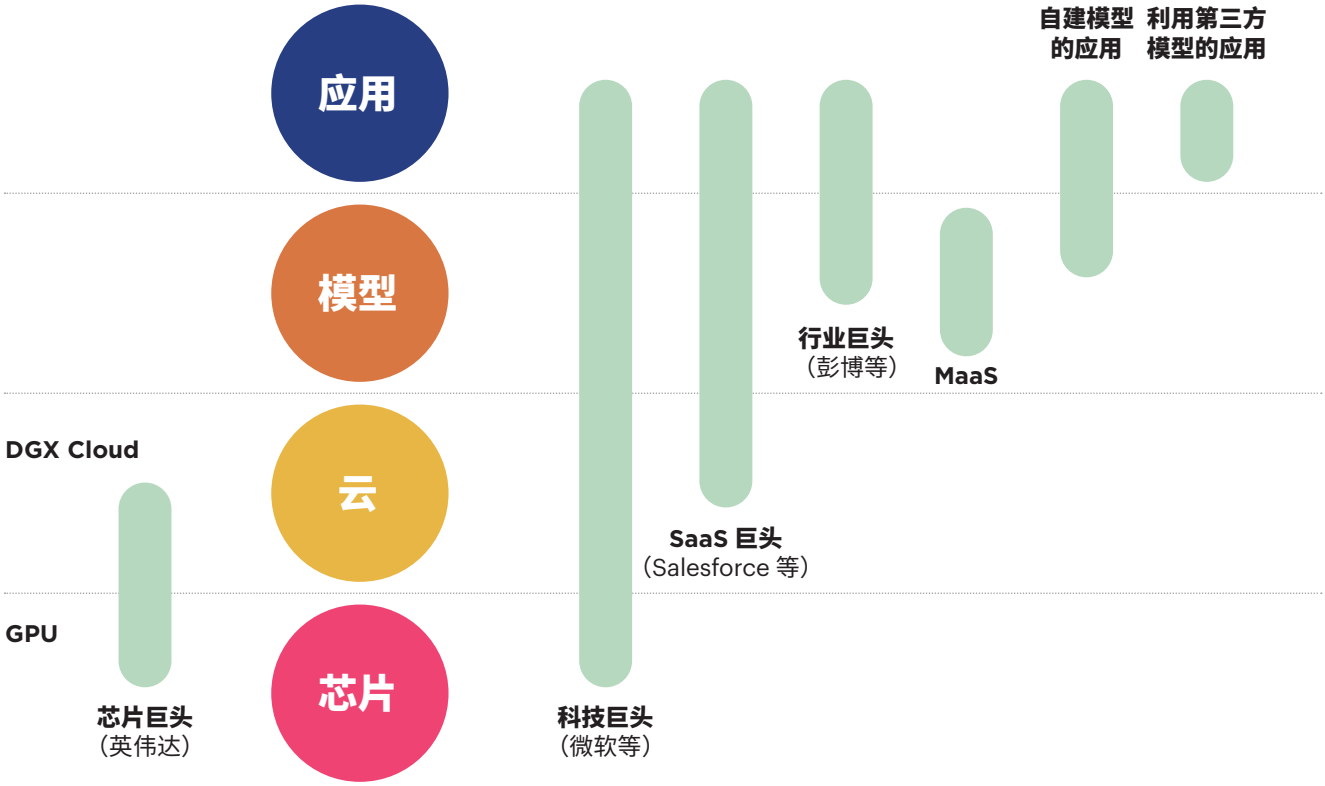
生态位与新物种

在生态系统中，每一个物种都拥有自己的角色和地位，即生态位。处于不同的生态位，则指示了不同物种之间的合作和竞争关系。

AI 2.0 的生态同样如此。新的“模型即服务（MaaS）”企业，以及自建模型、微调模型或调用 API 服务市场具体应用场景的企业，蔚为这个生态中的“新物种”，寻找着属于自己的新市场，同时为竞争做着准备。以下是一些对于这些新物种的观察：

- OpenAI 是“新物种”的代表，率先打造出具备涌现能力的大模型，激活了整个生态系统。这让在 AI 1.0 时代有所成就的企业紧张，但又让更多的创业者与投资者兴奋。
- 生成式 AI 的原生企业，它们遍布基础设施层、模型层和应用层。从提高研发和使用模型效率的工具链企业，到致力于打造下一代模型的大模型公司，再到众多通用或着面向行业的应用公司，这些企业的创新日新月异，为生成式 AI 带来了无限活力。
- 云巨头研发通用大模型，服务于自身业务，也对外开放 API。微软旗下操作系统、生产力工具、企业管理系统、代码平台、安全套件都拥有了副驾驶（Copilot）；百度要把每个产品重做一遍。同时，这些巨头还在开发自己的芯片，谷歌已有了 TPU，微软则是在研发雅典娜（Athena）。
- 芯片厂商也在拓展自己的边界，英伟达针锋相对地推出了 DGX Cloud，它还在强化赋能元宇宙（Omniverse）与大模型工厂（AI Foundations）的云平台。
- SaaS 巨头原本就是基于云的应用，正在从大模型汲取新的动能。未来，绝大多数 SaaS 企业都会是包含生成式 AI 功能的 SaaS 企业。
- 彭博等行业龙头开始防御性地采纳自有大模型技术，也盯着基础模型的机会。

此外，还有闭源与开源的路线，由于 License 的限制，开源模型并不一定可以商用，并且开源模型无法确保在未来一直迭代来匹敌闭源模型的效果。而基于闭源模型，很多企业又会担心未来的迭代可能受制于人。



定价模型：基础设施层

新的应用要有新的基础设施。AI 2.0 的基础设施是以提供智能算力为中心的智算中心。无论是模型还是应用，它都离不开硬件厂商或云服务商。

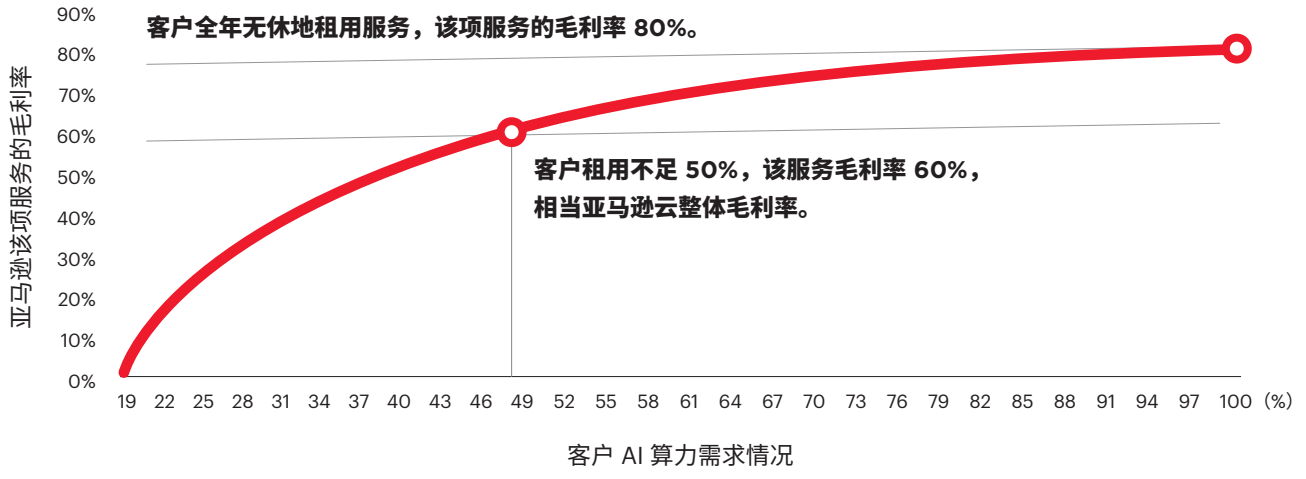
GPU 是训练模型与加速推理的关键算力硬件。大模型还拔高了对数据中心带宽、数据存储的门槛。云服务商会采购各类硬件，辅以冷却系统与运维服务，构建灵活、可扩展的 IaaS 平台，按需为客户提供算力。

传统云巨头获利颇丰。

- **设备定价：**假定亚马逊直接采购英伟达组装好的 DGX A100 平台。它集成了 8 片 A100 GPU，配置了内存、CPU、网络等软硬件组件，初始售价 20 万美元。实际上，亚马逊选择了采购 A100 芯片，自己搭建数据中心，这虽然能够压低一些成本，但仍然使英伟达获利颇丰。
- **年均成本：**亚马逊 AWS 数据中心按五年线性折旧，年均 4 万美元。
- **服务定价：**假定亚马逊 AI 算力出租的收入，全部来自 p4d.24xlarge，它向客户提供 8 片 A100 算力性能的加速服务。（亚马逊目前还规模化提供基于英伟达 V100、自研 Trainium 等硬件的算力服务，此处选取当前最主流的 A100 为测算基准。）如果承诺一年内稳定的用量（Compute Savings Plans），且不提前预付费用，目前它的每小时价格为 24.21 美元（美东俄亥俄的价格）。
- **年均收入：**如果客户一年 365 天一天 24 小时不停的租用算力，年均 21 万美元。
- **该项服务的毛利率：**那么，亚马逊该服务对应的毛利率将是 1-4/21=80.9%

即如果生成式 AI 的生态持续扩展，市场繁荣，客户全年无休地渴求算力，那么亚马逊该项服务的毛利率最高可达 80.9%。如果客户只有 50% 的时间用到了它，那么 8 片 A100 加速服务的年均收入就降到了不足 11 万美元，该项服务的毛利率就只有 1-4/11=63.6%，相当于外界预估的亚马逊云服务的总体毛利率。如果用户只有 20% 的时间用到了它，那么收入只有 4 万美元，该项服务的毛利率为 0。事实上，AI 算力目前是稀缺资源，AWS 正在极大受益。

亚马逊 AI 算力服务的毛利率，随客户需求提升而提升

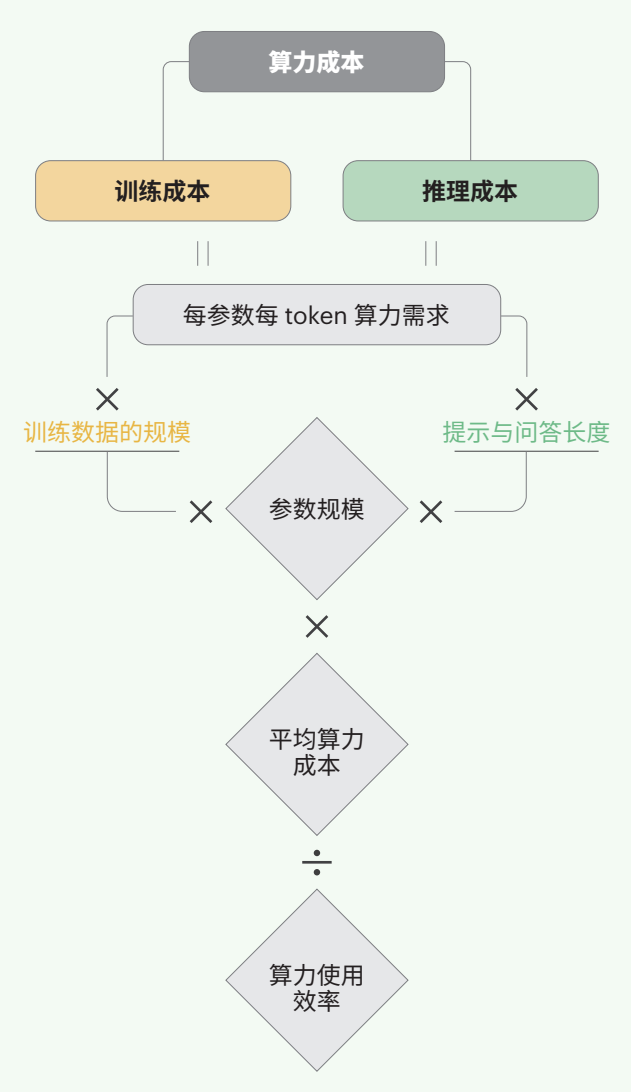


说明：基于硬件 DGX A100 采购折旧价格与亚马逊 p4d.24xlarge 服务预购一年价格，未涉及运维与能耗等各种成本。未考虑不同地区不同时间的市场价格波动。未考虑承诺外用量的额外费用等。亚马逊云服务毛利率估算数据来自 Bear Stearns。假设所有机器都投入生成，仅根据用户的需求导致运转时间有差别，并未考虑有部分机器完全闲置的情况。例如所有机器都 50% 的时间运转，而非 50% 的机器完全闲置。

定价模型：模型层

算力需求是模型层企业成本结构中，占比最显著的一部分。其他还包括数据收集与预处理、MLOps 工具、能源消耗等。算力需求可分为训练与推理两大阶段。一些机构提出了各自的估算方式，它们可以用一个公式来简单概括：

每参数每 token 的算力需求是常数，在训练阶段一般为 6 FLOPs，推理阶段则为 2 FLOPs。其他几项共同导致了不同模型的不同成本，是降低成本的重要方向。



参考论文 Scaling Laws for Neural Language Models，与 Semianalysis 等

- 平均算力成本主要由 GPU 性能等决定，每 FLOP 的价格平均每 2.5 年下降 40%-50%。
- 算力使用效率取决于软硬件优化水平等。据谷歌 PaLM 的论文，在训练阶段，缺乏优化经验或堆叠过多芯片，效率可能低至 20%，目前谷歌与 OpenAI 都能达到 50% 左右。前述机构推测目前推理阶段的效率在 25% 左右。

训练一次类似 GPT-3 的大模型，即 1750 亿参数规模，3000 亿 tokens，需要 $6*1750*10^8*3000*10^8=3.15*10^{23}$ FLOPs 的算力需求。如果只用 1 片 V100，在 FP16 精度的 28TFLOPS 的理论算力下，需要训练 $3.15*10^{23}/28/(1*10^{12})/(365*24*60*60)=357$ 年；要缩短训练时间，就要增加硬件投入，但算力使用效率就会下降。

租用云服务，亚马逊刚推出 8 片 V100 算力的 p3dn.24xlarge 时，预购一年（Savings Plans）每小时 18.3 美元。按 50% 算力使用效率估算，在 2020 年时，训练 GPT-3 的成本约为 357*(18.3/8)*365*24/50%=1430 万美元。

现实会复杂一点。不同云服务商的可用算力资源不同，价格也不同；大模型训练时长与并行多个模型同时训练的行为，也影响着算力使用需求。事实上，OpenAI 采购了 GPU，还得到微软支持，实际单次训练成本会比估算更低；但反过来，实际上训练一次是几乎不可能训练成功的，在大模型构建的过程中，存在着大量的失败和反复，此外为保证模型迭代的更快，需要进行大量的并行训练。即便打造出第一版大模型，后续模型的持续迭代的成本也无法避免。

尽管如此，理论上，随着硬件性能提升，软件优化程度提高等，大模型的训练成本会随着时间的推移而下降。如果只用 1 片 FP16 精度下理论算力 312TFLOPS 的 A100，来重新训练一次 GPT-3，则需 $3.15*10^{23}/312/(1*10^{12})/(365*24*60*60)$=32 年。亚马逊刚推出 8 片 A100 算力的 p4d.24xlarge 时，预购一年（Savings Plans）每小时 19.22 美元，按 50% 效率估算，目前，GPT-3 的训练成本已降至 32*(19.22/8)*365*24/50%=135 万美元。

去年，英伟达 H100 发布，性能进一步提升，也将带来成本的进一步下降。SXM 版本 H100 的 FP16 精度（FP16 Tensor Core），算力达到了 1979TFLOPS，是 SXM 版本 A100 的 624TFLOPS 的 320%。但据 Lambda 测算，H100 的训练吞吐量（Training Throughput）为 A100 的 160%。

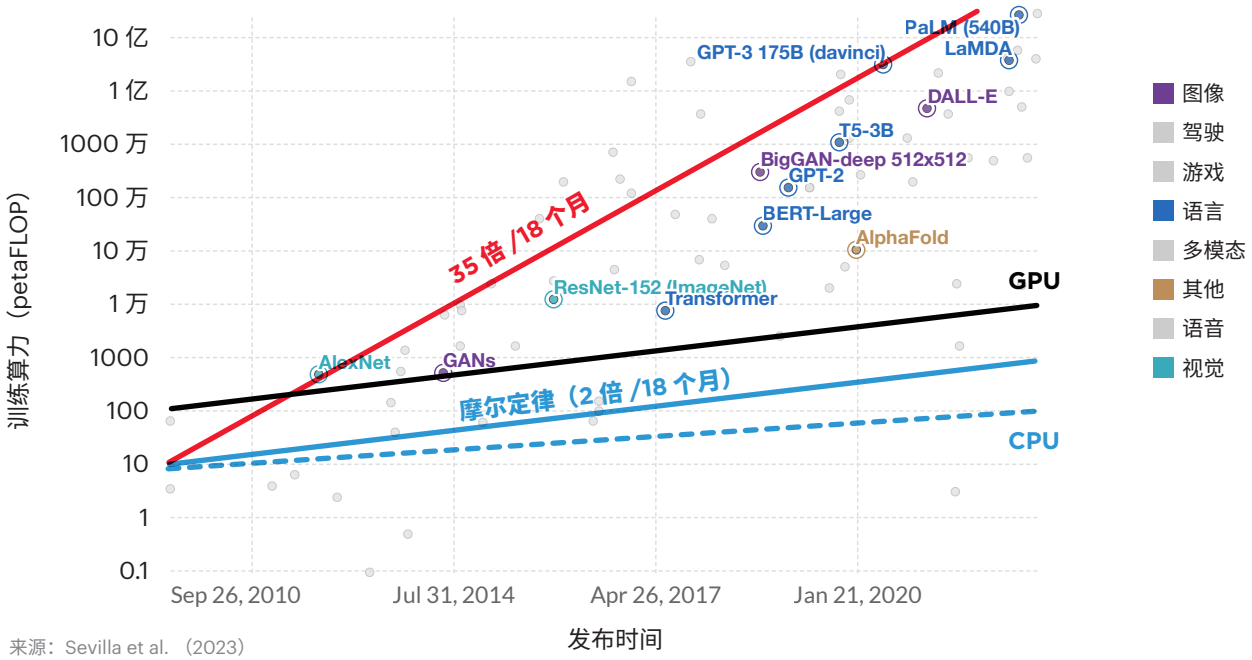
当然，如果大模型参数持续膨胀，训练成本将令市场难以接受。在当前 GPU 以类似摩尔定律的进步速度提升的情况下，大模型参数规模的增长可能会遭遇瓶颈。一方面是算力硬件迭代速度跟不上，另一方面则是因为现实世界生产高质量的训练数据的速度也不够快。这也是为什么 OpenAI 的 CEO 奥特曼认为，“现在已经接近巨型模型时代的尾声”，要寻找其他诸如分布式训练、任务调度优化等方式进一步提高训练效率。

随着 A100 逐步替换为 H100，推理成本也在下降。去年，OpenAI 的 gpt-3.5-turbo (4K context) 的调用价格为 0.02 美元 / 千 tokens。假定 GPT-3.5 的参数规模为 1750 亿，用户调用时，输入 500 tokens 长度的提示词，获得 500 tokens 的内容输出，且这一推理过程完全基于 A100 实现，算力使用效率为 25%，那么单次推理算力需求为 $2*1750*10^{8}*(500+500)=3.5*10^{14}$FLOPs，单次推理成本为 $19.22/8/(312*1*10^{12})/(60*60)*3.5*10^{14}/25\%$ =0.003 美元 / 千 tokens，毛利率约为 1-0.003/0.02=85%。

OpenAI 具有先发优势，为在竞争中赢得更多市场，它的定价策略更为激进。目前，同样的 API，服务价格已较去年下降了 90%，低于 0.002 美元 / 千 tokens。推出更多样的相对高价的 API 服务，以及在算力硬件中提升更高性价比的 H100 的占比，都有助于稳住毛利率。

但这取决于英伟达的产能。目前，亚马逊尚未成规模地对外提供 H100 算力资源，因此无法参考亚马逊上 H100 的定价。即使忽略现实资源有限的情况，采用当前 Lambda 平台上 1.99 美元 / 小时的 1x NVIDIA H100 PCIe（该款芯片单片 FP16 理论精度 1513TFLOPS）服务，OpenAI 该服务的单次推理成本变为 $1.99/(1513*1*10^{12})/(60*60)*3.5*10^{14}/25\%$ =0.00051 美元 / 千 tokens，毛利率约为 1-(0.00051/0.002) =74.5%，已低于去年。

大模型参数规模增长速度超过摩尔定律



来源：Sevilla et al. （2023）

定价模型：应用层

应用层企业的成本结构中，除了软件本身的成本外，就是调用大模型 API 时产生的费用，这部分的成本与活跃用户规模、单个用户日均推理次数，单次推理输入提示词与预置文本的长度，单次推理输出的内容的长度等相关。

这些变量又与应用层企业所在的应用场景相关。有些场景用户量较少，或问答频次较低，但需要更长的提示词或预置文本让大模型更懂自己。有些场景问答则相对简短，但用户与大模型间可能会聊得停不下来。

假设现有三家应用企业，调用 OpenAI 的 gpt-3.5-turbo（4K context）服务，该模型的计费规则为输入 $0.0015 / 1K tokens，输出 $0.002 / 1K tokens，它们对应如下应用场景：

- **查询工具：**企业内部知识查询，偶尔遇到问题，就查询一下。特点是低频 (假设日均 3 次)，短输入（假设单次 50 tokens），中等输出 (假设单次 300 tokens)。当百万 DAU 时，单日成本为 0.2 万美元，千万 DAU 则达到 2.03 万美元。

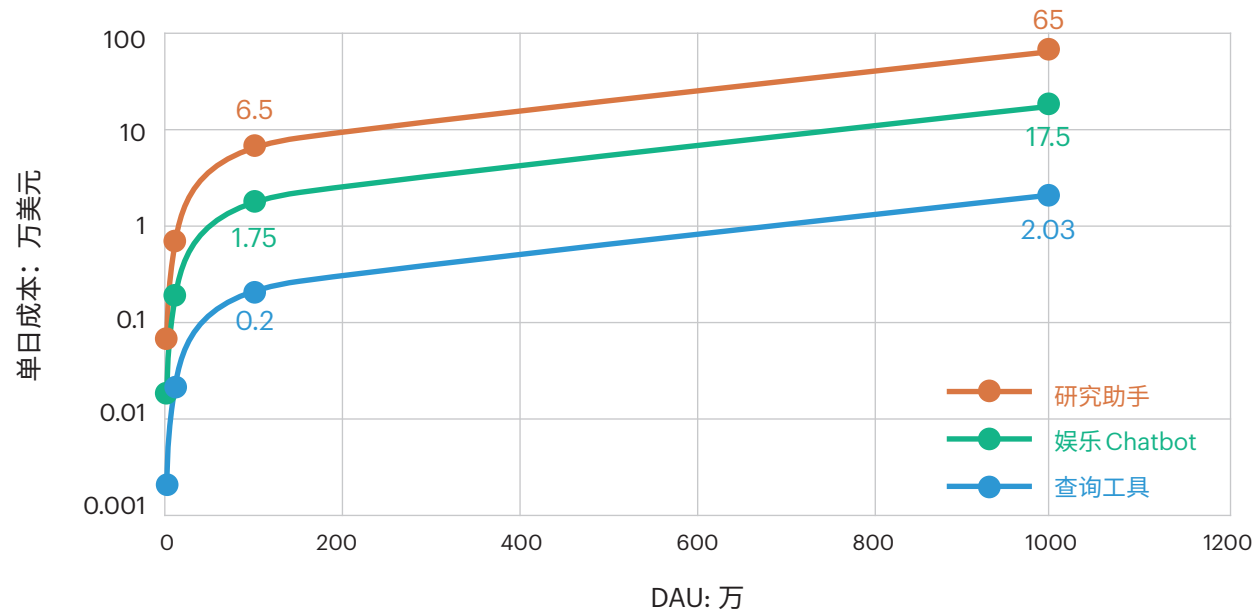
- **研究助手：** 日常工作和研究使用。特点是中频 (假设日均 10 次)，长输入 (假设 3000 tokens)，长输出 (假设单次 1000 tokens)。当百万 DAU 时，单日成本为 6.5 万美元，千万 DAU 则达到 65 万美元。事实上，这类应用达到千万 DAU 非常不易。

- **娱乐 Chatbot：**吃掉了用户大量空闲时间。特点是高频 (假设日均 100 次)，短输入 (假设单次 50 tokens)，短输出 (假设单次 50 tokens)。当百万 DAU 时，单日成本为 1.75 万美元，千万 DAU 则达到 17.5 万美元。事实上，娱乐 Chatbot 往往需要依赖上下文的记忆，如果计入记忆的 token，则单日成本还需增加数倍。

应用企业通过预估每次输入输出需要用到的 token 数量，以及自己想达到的 DAU，即可预估出每天在大模型 API 上的开销。

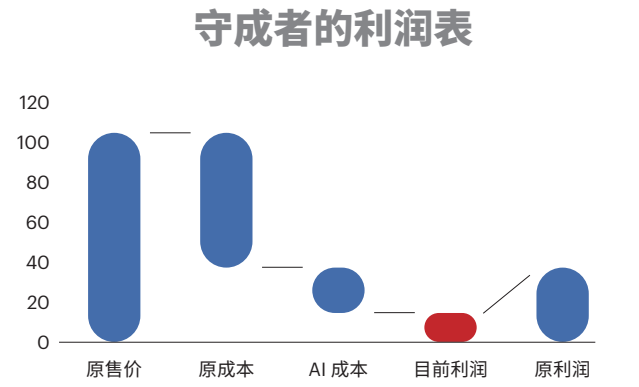
当然，这就是充满混乱与诱惑的早期市场。想要达到百万和千万量级的 DAU 需要企业跑得越快。但由于竞争，应用层企业的利润空间很快就可能收窄，例如 Copy.ai 的定价策略就与 Jasper 针锋相对，用更低的价格夺取市场。即便现在是生成式 AI 的早期市场，在拥有多家创业企业的特定市场中，单纯调用 API 的应用企业更可能无法做到差异化，那么行业整体毛利率的下降将很快到来。

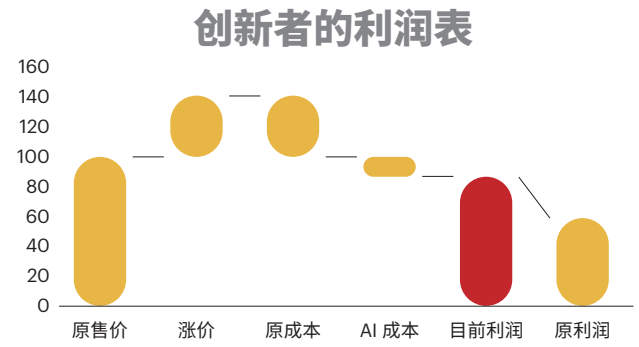
不同应用场景下的推理成本变化

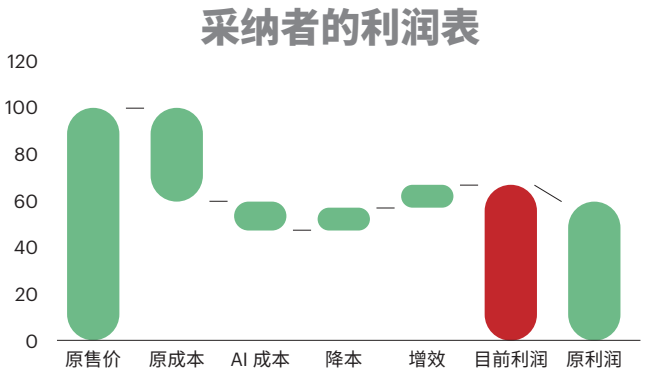


企业运营发生改变

生成式人工智能不仅意味着技术变革，还意味着流程再造。面对 AI 2.0 的冲击，市场诞生了三类玩家：







- **守成者：**这类玩家受到传统业务既得利益的束缚，转型缓慢。企业将受到市场冲击，如果不采纳 AI 2.0，将会因为市场竞争对手提供了差异化的产品，逐步失去市场；如果采纳 AI 2.0，但一时又无法提价，就会因额外 AI 成本的上升，导致利润下滑。谷歌的搜索业务就是如此。它成为市场眼里的守成者。微软 CEO 纳德拉声称，搜索的毛利率将永远下降。

- **创新者：**这类玩家积极拥抱了新的技术，为原有产品提供了新的功能，甚至是新产品新品类，获得了服务溢价与竞争溢价。尽管企业也需要支付额外的 AI 成本，但涨价弥补了这一切，让它的利润较之前有所提升。微软是创新者的代表。最近，微软宣布 Microsoft 365 Copilot 涨价 40%。很多 SaaS 企业最终都会如此。Notion 为它的 AI 每月收费 10 美元，相当于为其最受欢迎的那档订阅服务，提价了 100%。

- **采纳者：**还有部分企业只是在公司内部业务流程中采纳了新技术。这能压缩研发人员、行政人员、销售人员等人力成本。多项研究表明，知识工作的岗位，受本轮生成式人工智能的冲击较大。此外，随着人员缩编，流程优化，沟通中的效率损耗也随之减少。目前，已有券商用生成式 AI 来帮助阅读财务报告，律所用来起草合同文本，营销企业用来撰写文案，软件企业用来编写代码，客服用来回答问题。不同行业甚至同一行业内不同企业，内部的岗位结构不同，受 AI 2.0 暴露影响的情况也不同，企业的利润表变化差异也会较大。

现有研究普遍认为，客户互动、文书撰写、代码编写、资料搜索与收集、数据分析研究等工作内容及工作时间占比较高的金融及咨询行业、客服行业、营销行业、软件行业将是 AI 2.0 的最积极的采纳者。如果其中一半工作可以由 AI 2.0 取代，将节省大量人力成本。

创新者与采纳者之间的界限并非那么清晰。采纳者也可以在重塑自己的业务流程后，对外服务，赢得创新溢价。创新者也可以利用 AI 技术，提升服务溢价的同时，减少自己的研发成本。比如 SaaS 服务商，在交付软件扣除成本后，可以将剩余的毛利润用于运营，即研发、行政、销售营销等。对初创期的 SaaS 服务商而言，研发与营销的成本往往很高；对于成熟期 SaaS 运营商而言，销售成本占比又往往难以下降。

即使是被颠覆者，也还有调整架构的机会。面对冲击，除了模型优化与算力提升外，谷歌正在围绕大模型走向生成式搜索，重新设计自己搜索业务与广告引擎的技术架构与用户体验。

市场格局

AI 1.0 时期，从 2012 年到 2015 年，AlexNet 等技术突破促使 AI 成为创业和投资热潮，融资数量不断上升。但由于产业落地不畅，此后总融资额和新创立的 AI 企业的数量开始下降。

市场在 AI 2.0 时期重新活跃。去年底以来，每周都有新的生成式 AI 产品发布。据 CB Insights，这个领域全球至少诞生了 13 家独角兽企业。

中国没有错过这一轮技术创新。5 月底，科技部旗下研究中心新发布的《中国人工智能大模型地图研究报告》统计，近 5 年来，中国研发的大模型数量排名全球第二，仅次于美国；2023 年中国发布的大模型数量超过了美国。

科技巨头吸引了更多的目光，但初创企业孕育着新的希望。它们或者迭代上一代 AI 技术，或者创造新的产品与服务。

文本与图像是目前相对成熟的两大模态，与美国类似，中国同样有较多的初创企业聚焦于此，但它们正在寻找如何将其融入视频、3D、编码等更多模态。

部分原 AI 1.0 时代的初创企业开始转型，计算机视觉、语音识别、自动驾驶乃至生物医药等企业，迅速结合业务数据与应用场景，将其融入了大模型，试图用更泛化的能力解决更普遍的行业痛点。创新正在外溢至周边领域，元宇宙、数字人等也在通过靠近 AI 2.0 获得新的活力。

法律、金融与营销领域存在更明确的新机会，无论是搜索、翻译、对话、摘要还是创作，都已有初创企业的身影；游戏、家装、服装等行业的概念图设计，也正在交给生成式人工智能。有些则更通用一点，为不同企业各自员工学习公司或行业知识，提供了更便捷的交互界面。

更多的增量正在被发掘出来，现在仍是 AI 2.0 的早期，基础架构和核心技术并不特别成熟；巨头忙于研发大模型，尚未顾及深度切入具体应用场景。这是初创企业的蓝海，也有航道下的暗礁。

市场地图

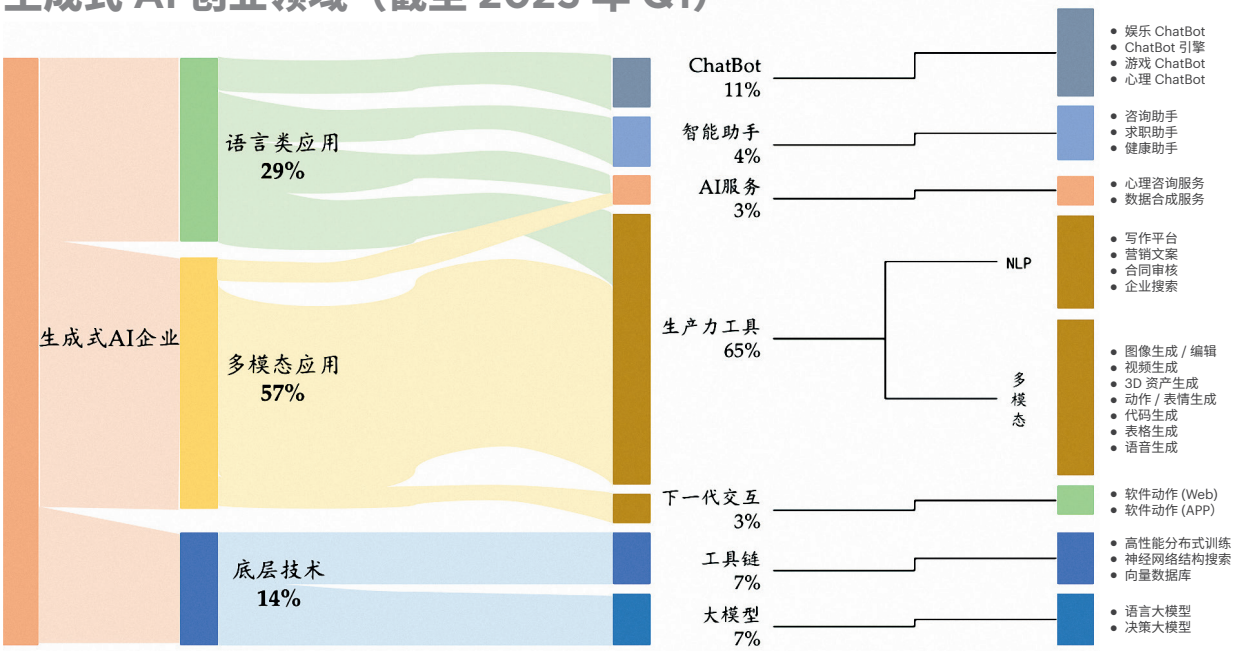
说明：不完全列举。部分企业尚未进行公开宣传，或无 Logo 等宣传资料，暂不予展示。部分企业横跨多个领域，此处仅列入相对典型的一项。

GPT-3 之后的新公司

截止到 2023 年 Q1，根据启明创投投资团队与超过 100 家在 GPT-3 发布后成立的大模型和生成式 AI 相关的中国创业企业的交流，其中，将近 30% 做语言类应用；企业数量最多是多模态应用方向，占比 57%；大模型企业，以及为更好地训练和应用大模型提供支持的工具链企业共占比 14%。

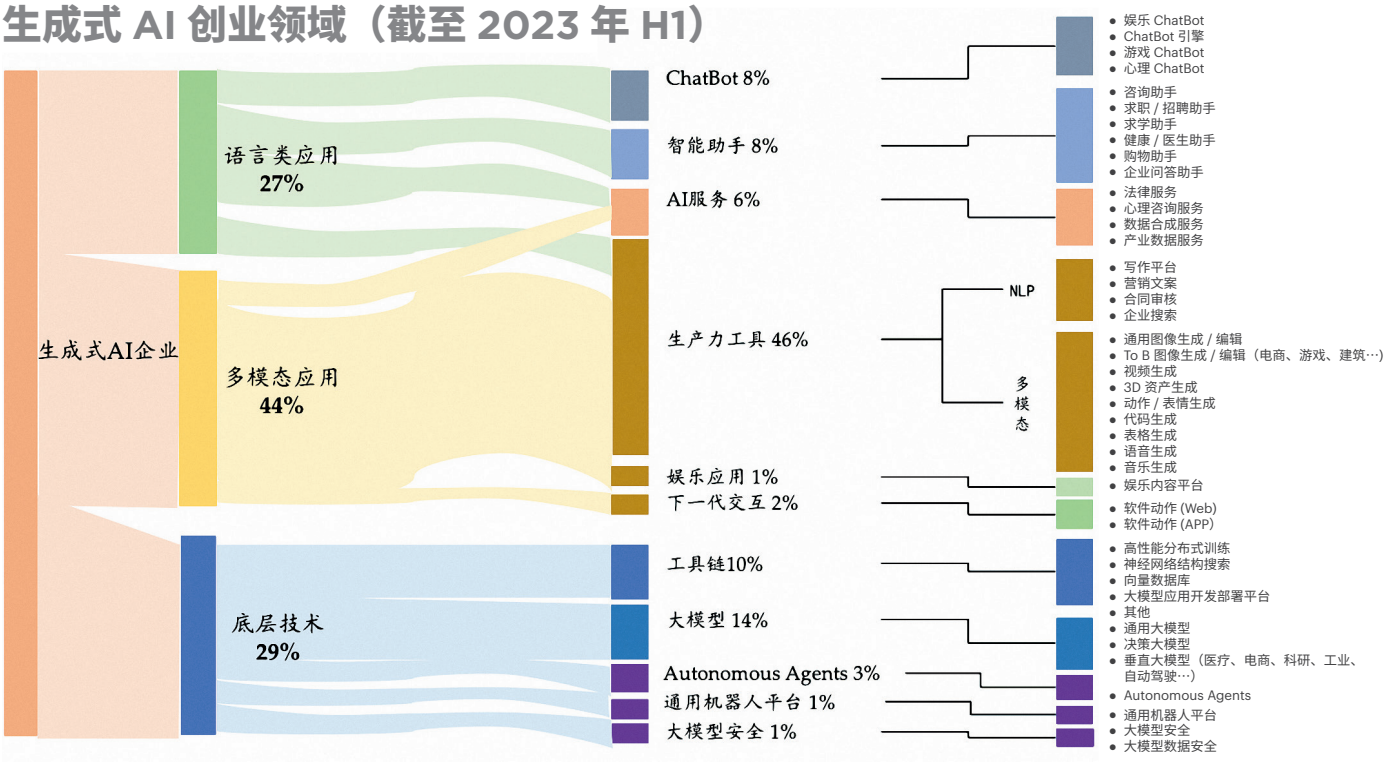
在 100 余家公司的具体方向中，ChatBot 占 11%，而生产力工具占得最多，高达 65%，包括文案写作、图像生成、视频脚本生成、3D 资产生成等。以下是截至 2023 年 Q1 的生成式 AI 市场情况：

生成式 AI 创业领域（截至 2023 年 Q1）



基于启明创投团队交流过的 100 余家企业的统计。

然而，市场发展是快速的，2023 年的 Q2 又涌现出大量的生成式 AI 创业企业，在 2023 年 H1 结束后，启明创投基于近 200 家生成式 AI 企业的交流，观察到的生成式 AI 市场情况如下图：



基于启明创投团队交流过的近 200 家企业的统计。

根据与这些公司的交流，启明创投发现市场上的创业公司呈现出以下趋势：

- 与 2022 年受到 Stable Diffusion 和 ChatGPT 刺激后快速涌现出的生产力工具方向的创业公司不同，2023 年有更多比例的新公司聚焦在底层技术的创新上，更多大模型公司和 infra 基础设施工具链公司在以技术大拿为主的创始人主导下成立。反映在数据上，具体表现为聚焦在底层技术的创业公司占比从 14% 提升到了 29%，而生产力工具型的应用公司占比则从 65% 下降到 46%。此外，在生产力工具的方向上，不同于此前仅微调 Stable Diffusion 等开源模型的创业公司，最新涌现的创业公司往往由更高级别的 AI 人才领导。
- 大模型创业公司开始分化，在通用大模型创业公司方兴未艾的同时，许多面向特定行业的垂直大模型公司开始出现，主要聚焦在医疗、电商、科研、工业、自动驾驶和机器人等方向。
- 具备行业属性的智能助手方向的创业企业开始增加，如求职、招聘、求学、法律、健康、购物、企业知识问答等方向的个人助手和员工助手方向的创业公司持续涌现，这代表着在经过一段时间对 ChatGPT、Stable Diffusion 的熟悉后，具备更强行业知识和资源的行业老炮型创始人逐渐进入生成式 AI 领域。

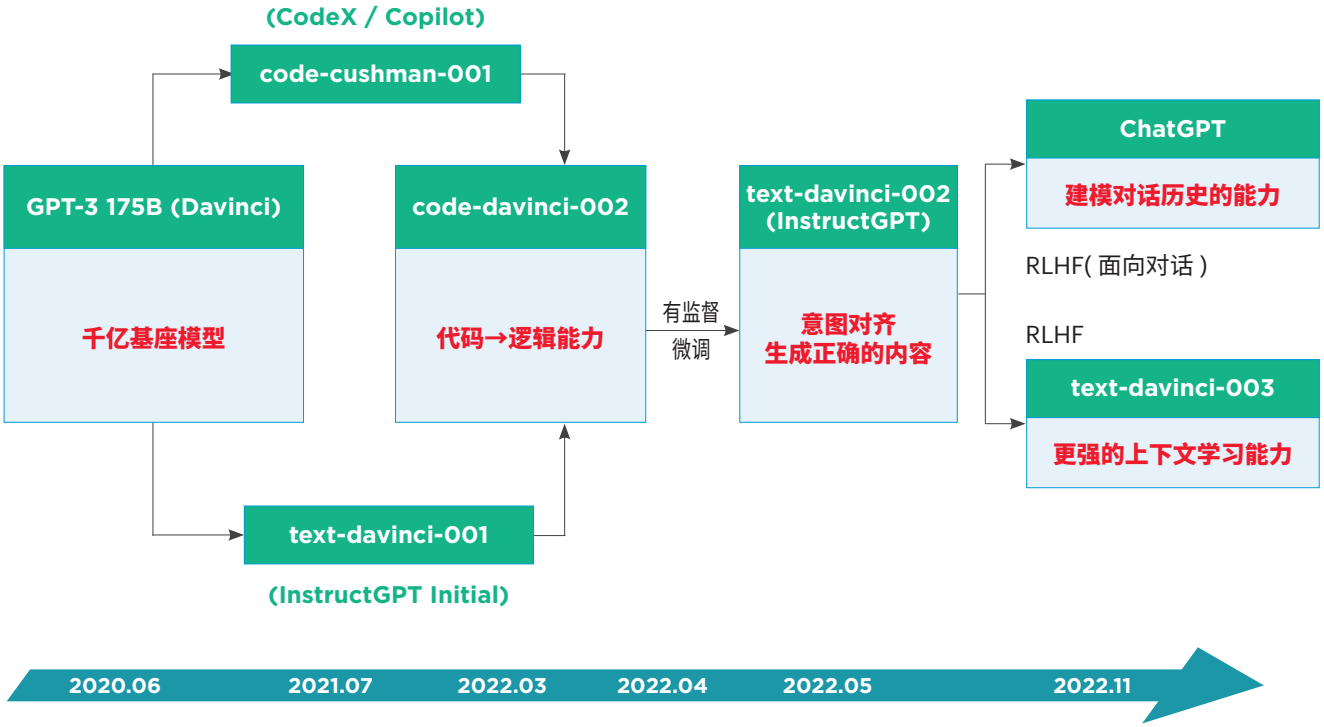
大模型公司

通用大模型

OpenAI 是模型层公司的代表，2020 年发布的 1750 亿参数的 GPT-3 曾一度是 AI 历史上最大的机器学习模型，相比于 15 亿参数量的 GPT-2，GPT-3 参数量提高约 117 倍，预训练的数据量也从 50 GB 提高到 570 GB。2023 年 3 月，OpenAI 发布的 GPT-4 则再次扩展了深度学习的边界，结合多模态能力达到了里程碑式的效果，并在各种专业和学术基准上表现出可以与人类媲美的水平。 可以说，GPT-3 打响了大模型竞争的第一枪，而 ChatGPT 和 GPT-4 的出现进一步加速了大模型主导权的竞争，是否拥有一个大语言模型底座对于大模型企业后续进一步优化出更好的模型至关重要。

ChatGPT 是 OpenAI GPT-3.5 优化后的模型和产品化体现，其背后的技术从 2018 年的 GPT-1（2018）开始，经过 GPT-2（2019），GPT-3（2020）逐渐达到里程碑式的突破，此后 2 年内 GPT-3 又经过两次重要迭代，引入基于人类的反馈强化学习系统（RLHF）后形成 ChatGPT。从 ChatGPT 的发展可以看出，对于模型层公司来说，技术的演进极为重要，公司需要极强的技术掌舵人和融资能力来保障研发投入的稳定性。

此外，通过对海外市场的观察，我们发现当前大模型竞赛中，由高级别 AI 人才主导的创业公司更加领先，例如 OpenAI, Anthropic 和 Cohere 等公司皆是如此。同样，类似 Adept, Inflection 和 Character.ai 等公司以极快速度实现了极高的估值，也表明顶级的 AI 人才正在通过研发大模型来构建有壁垒的应用，以此参与到生成式 AI 领域的竞赛中，而市场也更青睐这些顶级 AI 人才创立的公司。

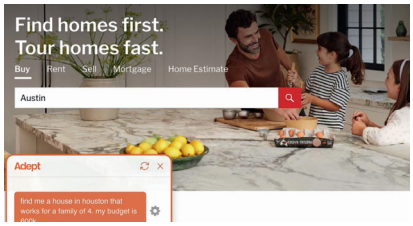


Adept.ai

融资历史 & 核心团队

融资日期	轮次	融资额	投资机构
2023 年 3 月	B	$350M	General Catalyst, Spark Capital, etc.
2022 年 4 月	A	$65M	Greylock, Addition, etc.

- CEO, David Luan, OpenAI 工程副总裁
- CTO, Niki Parmar, Google Brain 科学家
- 首席科学家 , Ashish Vaswani, Google Brain 科学家

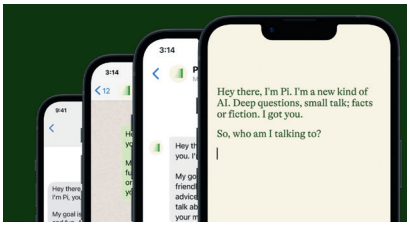


Inflection.ai

融资历史 & 核心团队

融资日期	轮次	融资额	投资机构
2023 年 6 月	B	$1.3B	Microsoft, Nvidia, etc.
2022 年 5 月	A	$225M	General Catalyst

- CEO, Mustafa Suleyman, DeepMind 联合创始人
- 联合创始人 , Reid Hoffman, LinkedIn 联合创始人
- 首席科学家 , Simonyan, DeepMind 首席 (Principal) 科学家

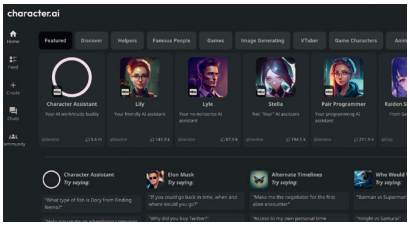


Character.ai

融资历史 & 核心团队

融资日期	轮次	融资额	投资机构
2023 年 3 月	A	$150M	Andreessen Horowitz

- CEO, Noam Shazee, Google 首 席 (Principal) 软 件 工 程 师，Transformer 作者之一
- 总裁 , Daniel Adiwardana, Google 资深软件工程师



同样，目前中国市场普遍看好从模型出发的公司，当前大模型公司具备以下三个特点：

- **投入大：**底层模型的构建需要超重资源投入，包括大量算力、数据和人才；
- **工程强：**由于大模型具备更强的泛化能力和提供方的商业追求，大模型发布时就提供各类用法的样例；
- **营销强：**受到 OpenAI 高调营销（如高管频繁接受各种访谈）的带动，国内大模型公司召开发布会已经成为常态。

在通用大模型百舸争流的今天，国内绝大多数的大模型团队在 2023 年之后成立，在同时起步并角逐大模型皇冠的路上，团队至关重要。正如 GPT-4 报告中披露的，研发出 GPT-4 至少需要六个方向的研究团队（Pretraining, Long context, Vision, Reinforcement Learning & Alignment, Evaluation & Analysis, and Deployment），国内大模型创业团队需要有极强的算法、工程和数据能力：

- 将市面上存在的算法用艺术的形式组合起来，成为最终模型的某个环节；
- GPT-4 未公开算法，企业需要创造性地提出自研算法才能研发出达到或超过 GPT-4 效果的通用大模型；
- 基础模型的研发需要极强的分布式训练等工程能力的支持，团队需要确保对计算资源的高效利用，并建设高质量数据集以保证模型的效果。

当然，巨头不会懈怠，如何与科技巨头竞争和合作，始终是贯穿初创企业成长的难题。国内科技巨头几乎每周都会宣布大模型的研发进展与行业合作动态，它们横跨了云基础设施与大模型，而且在它们那里模型层与应用层的界限相对模糊。百度宣称要把所有的产品都重做一遍，而坐拥最多用户的腾讯决定先聚焦产业。但竞争的关键，还是提供效果最优的模型，辅之以足够可靠的产品与服务。

垂直大模型

垂直大模型企业往往不会充当模型提供商，较多采用“自建大模型的垂直应用”的模式。 除了创业公司以外，有兴趣研发垂直大模型的组织主要还有互联网公司、AI 1.0 企业和行业龙头等。

对于自研垂直模型的企业，行业数据尤为重要，拥有高质量的行业数据和私有数据，是针对特定行业优化大模型表现的关键。以彭博自研的 BloomBergGPT 为代表，金融行业数据超过了公开数据，占比达到 51%。因此，最终模型效果在很多金融任务上有出色的表现。

目前构建面向垂直行业的模型有以下三种方式：

- 在已经完成训练的通用大模型基础上，结合大量自有的行业数据进行微调（fine-tuning），在此之前是否对通用大模型进行蒸馏、后续是否外挂知识库则视情况而定。
- 通过改变数据的分布，结合更多特定行业的数据进行预训练，直接打造行业大模型。
- 通过自定义一种专属语言，并用（文本，专属语言）这样的 pair 对大模型进行 fine-tuning，并将生成的专属语言输入到自研的 AI 模型中，完成【用户输入 – 大模型 – 专属语言输出 – 自有 AI 模型 – 业务结果输出】的全过程。

应用层公司

模型层公司的分量虽重，应用层公司的数量仍是最多的。这是创新最活跃的地方。绝大多数应用层公司的创业者不需要从头训练大模型，只需要直接利用底座模型的能力，叠加对于场景和行业的深刻理解，就可以支持一家应用公司的发展。

根据AI能力来源及其占比，这些应用公司大致可以分为三类：

- 调用外部大模型的 API 为主的模式。这类团队本身通常不会有很强的预训练模型开发能力，更多是具备应用层的能力。他们往往是年轻创业者，或是来自垂直产业的老兵，搭配几位 IT 专家，基于 API 或开源模型去开发应用，至多做一些微调与修改。

- 结合了 AI 1.0 模型能力的模式。他们仍以调用 API 或使用开源模型为主，但又涉及大模型技术以外的 AI 算法。这类团队内部培养了一些深度学习算法的工程师，才能更好地实现既定效果。

- 自研 AI 2.0 模型能力的模式。这就是“模型 + 应用”的垂直大模型模式。这类团队通常需要高度熟练的机器学习科学家、大量相关的数据、训练基础设施和计算能力。团队领袖往往是 AI 行业的顶尖人才，有过成功的大模型预训练经验。当然，这些公司也不会介意借鉴一些开源模型加快研发速度。

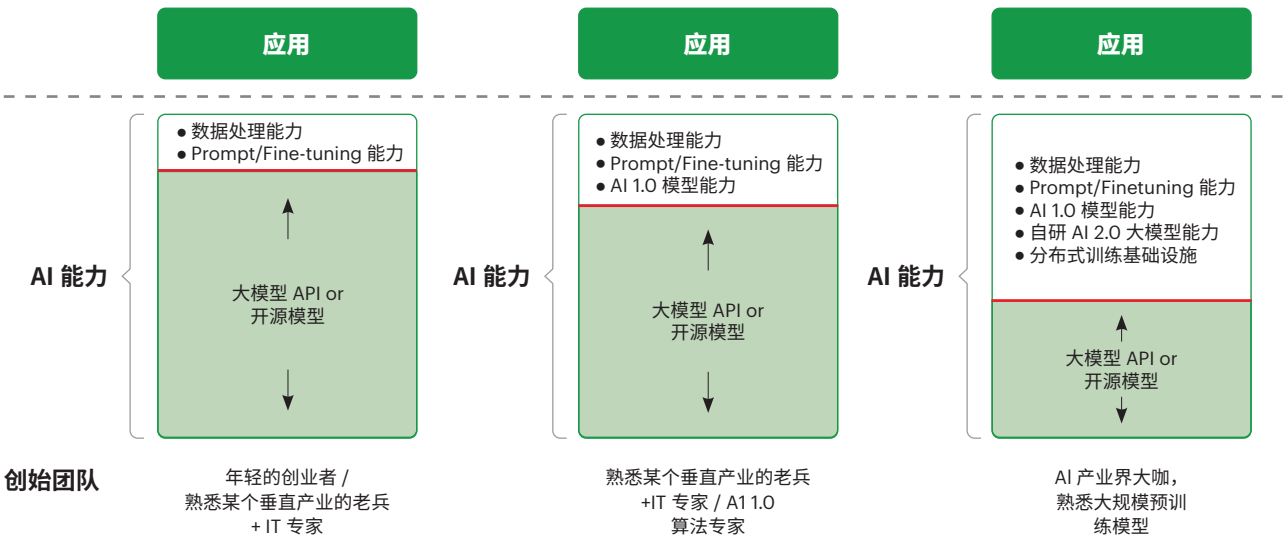
三类模式并没有孰优孰劣之分。不同应用场景，不同发展阶段，需要合理采用不同的模式。随着调用 API 为主的初创企业逐步发展，团队变得更为成熟，会很自然地将提升自身 AI 能力提上日程。

如果没有自研 AI 2.0 模型的能力，想要成功，就要快速推出产品并占领市场，并持续领先地迭代出更合适客户的产品。它的产品或服务，成为工作流程中的一环，或建立新的用户社区，是能否持续快速规模化的关键。但长期来看，它的竞争壁垒仍是传统软件的规模效应、切换成本，而技术壁垒较低，最底层技术很难实现差异化。

调用外部大模型的 API 为主的模式，它们还面临被原厂大模型迭代到下一个版本后吃掉市场的威胁。而结合了 AI 1.0 模型能力的模式，也将面临大量同质化产品的竞争，即便公司在早期发现了蓝海市场，在实现产品与市场的匹配（PMF）后，也可能引起竞争对手快速跟进，并且容易受到科技大厂的竞争。

自研 AI 2.0 模型，想要成功，就要持续拿到大量融资，在实现对早期大众的占领前，始终保持自研模型效果不低于第三方模型，同时需要兼顾好产品打磨、业务发展、销售和营销等。它们面临较少的同行业竞争对手，但面临大模型边界扩展的威胁。它的竞争壁垒在于，如何扩大自己的技术领先优势与资本投资热情。

创业公司使用 GPT-3/GPT-4 等第三方大模型或开源模型的三种形态：



语言类应用公司

在全球范围内，基于自然语言处理的应用，在 transformer 应用中的占比 40%。在国内，根据启明创投交流过的公司统计，语言类应用占了近三年成立的生成式 AI 企业的 27%，此外，多模态类应用中还有占比近 1/3 离不开自然语言处理。

语言类公司，按功能来分，可以分为翻译、对话、摘要、生成、推理等，可以用于构筑智能对话、智能助手、智能服务与生产力工具；按应用场景来分，这些公司出现在社交、咨询、招聘、健康、心理、金融、法律与营销等领域。

语言类公司面临强大的竞争对手。微软的生产力套件挤压了原生应用的市场空间；排名靠前的全球 SaaS 巨头纷纷推出自己的语言类 AI 应用，部分还是自研的大模型。

不同应用场景决定了语言类公司不同的竞争策略，有些场景需要快速跑出流量占得先机，比如招聘和社交，流量爆发后形成的规模效应将是这些场景下创业公司的核心壁垒；而另外一些场景，则需要深耕行业，例如金融、法律、心理和营销，具备信心的创始人还需要有极强的行业认知，在攻下一个个行业客户 / 用户后，切换成本将为他们建立显著的优势。

简单的文本处理和套壳的 Chatbot，将很快成为红海。创业公司要从取悦早期用户的兴奋中冷静下来，构筑更高的壁垒。自然语言正在成为一种新的交互界面，连接用户与世界。问题不断出现，知识不断更新，绝大多数用户并不仅仅满足获取可能与事实不符的娱乐性对话，如何把知识嵌入到大模型的需求非常迫切，当前相关技术和产品尚供给不足，仍处于蓝海状态。

文本模态的应用企业，要警惕被自己杀死。它需要对齐来满足公众价值和国家的监管要求。越来越多的生成式人工智能，正在制造良莠不齐的文本内容，它们会成为训练数据，也会成为搜索来源。谁先解决这个问题，谁拥有更广阔的发展空间。同样，如何帮助 Facebook 和 Twitter 这样的社交网络或用户社区防止泛滥的 AI 数字人和 AI 回复，也将带来不小的创业机会。

语言类应用的初创企业

说明：不完全列举。部分企业尚未进行公开宣传，或无 Logo 等宣传资料，暂不予展示。部分企业横跨多个领域，此处选择相对核心的业务重复展示。

多模态应用公司

多模态方向上的技术创新与应用场景，也为中国的生成式 AI 应用公司提供了巨大机会。在中国庞大的互联网、消费市场、实体经济中，蕴藏着丰富的多模态数据。同样，抖音、快手、微视等短视频或直播应用也占据了用户大量的时间。

最早的一批多模态应用经过将近一年的发展，已经有公司的总注册用户量突破百万甚至两百万，并初步实现了早期的商业化收入。但如何进一步扩大用户量，或深入到游戏、电商等特定行业实现大规模收入的路径尚不清晰。另外，更强技术背景的创始人正在进入这个赛道，准备研发更强劲的模型来解决可控性等问题。未来，如何从创意工具走向可控性极强的生产力工具，将是多模态应用公司需要回答的关键问题。

多模态应用的初创企业

说明：不完全列举。部分企业尚未进行公开宣传，或无 Logo 等宣传资料，暂不予展示。部分企业横跨多个领域，此处选择相对核心的业务重复展示。

与 Text-to-Image 企业已经拥有了相对不错的生成效果，而在争夺可控性的制高点不同，Text-to-Video 和 Text-to-3D 企业则在比拼生成内容的效果。视频和 3D 生成领域尚未出现如图像领域的 Stable Diffusion 一样风靡一时的模型，因此，这些方向上的公司进入商业化阶段的条件并不充分，需要通过模型层面的创新（无论是自研还是使用第三方模型），来生成符合用户预期的内容。

数字人企业重新焕发生机，在 AI2.0 到来之前，它们拥有很强的 CG（计算机图形学）能力，但对话能力却显得薄弱，很多情况下是没有灵魂的皮囊。大模型的出现补足了数字人企业的短板，让通用的 to C 数字人可以和用户进行更丰富和深入的交互，基于内容提供更强的情感链接；并让 to B 数字人从原来的“客户宣传需求驱动”和“电商平台合规驱动”，真正走向“效果驱动”。然而，大模型也激化了数字人企业的竞争，原本独特的 NLP 能力如今不再新鲜，通过接入大模型 API，每个应用公司都轻易具备。

当前，多模态的应用正在超越虚拟世界，向具身智能领域进军，从而直接与现实世界进行互动。例如，机器人需要在虚拟环境中模拟和仿真各种操作、理解用户的需求、感知周围物理世界的环境并规划要实现的动作。

总之，当前多模态方向上的创业公司尚处于发展的早期阶段，商业的想象力让这条赛道充满前景，但技术的不成熟又让这个方向充满了挑战。这个方向的创业公司同样面临着生成式 AI 公司无法回避的问题——即用户被生成的内容所吸引，与传统的 CRM 和 ERP 等软件不同，生成式 AI 的用户并没有表现出足够的黏性和切换成本。用户跟着优质的内容走，而谁能够提供优质的内容，就可以在提高渗透率的同时，把竞争对手的用户吸引过来。在技术尚未成熟的今天，谁能够提供更优质的模型，往往意味着能够提供更优质的内容，吸引更多的用户。可以说，致力于在多模态方向上打造出爆款应用的创业公司，必须具备极强的模型研发能力和创新能力。简单来说，颠覆式的 AI 应用的核心驱动力来自于底层模型的创新，两者无法解耦，一定时间内，模型的作用将大于产品设计的作用。

第二章 前沿研究

语言大模型需要超越预测下一个词的“快思考”能力，也需要一个思维更丰富、推理更严密的“慢思考”深层机制。

生成式人工智能领域的一个突出特征，是研究与创新过程的密切结合，许多在企业内部实现，迅速推出用例和产品。这种研究与创业的一体化，初创企业和风险资本起到了重要的作用，而美国科技巨头和主要人工智能企业的研究投入与人才密度，包括一些底层技术的研究，这些年来已经超过了大学等研究机构。

致敬 2022

负责谷歌研究的副总裁 Jeff Dean 在总结 2022 年时说：“自然对话显然是人们与计算机交互的一种重要且新兴的方式。” 实际上，2022 年不仅发生了这种人机交互革命，也充满了令人兴奋的 AI 论文。我们与全球科技大厂和顶级研究机构的 AI 领袖进行访谈和交流，请他们推荐 2022 年杰出论文，加上我们的最终评议，筛选出这 10 篇论文，其中的每一篇，都被行业专家认为会影响人工智能技术的发展方向。排名不分先后。

1 生成城市 3D

Block-NeRF: Scalable Large Scene Neural View Synthesis
Matthew Tancik 等， UC Berkeley, Waymo，Google Research

用 280 多万张图像训练了一个 Block-NeRF 的网格，渲染了旧金山的整个街区 。此前 Mega-NeRF 也刚开源。Block-NeRF 是一种神经辐射场的变体，可以表征大规模环境。该研究表明，当扩展 NeRF 以渲染跨越多个街区的城市场景时，将场景分解为多个单独训练的 NeRF 至关重要。重建大规模环境在自动驾驶、航空测量等领域具有广泛应用前景。

2 ConvNeXt: 卷积神经网络的“复兴”

A ConvNet for the 2020s
Zhuang Liu 等，Facebook AI Research， UC Berkeley 等

这是一篇在 2022 年被引述次数最多的论文之一。卷积神经网络在 Transformer 诞生前称霸了整个计算机视觉领域，而作者的这篇论文就是让 ConvNet 重新在视觉领域大放异彩。

3 跟随人类指令

Training language models to follow instructions with human feedback
Long Ouyang 等，OpenAI 团队

研发团队影响力的真正考验当然是技术如何在产品中出现，OpenAI 紧随其 2022 年 3 月论文《训练语言模型以遵循人类反馈的指令》后，于 2022 年 11 月底发布了 ChatGPT，震惊了世界。如此快速的产品采用是罕见的。

4 最优计算

Training Compute-Optimal Large Language Models
Jordan Hoffmann 等， DeepMind

表现最好的模型不是按参数衡量最大的模型，而是一个较小的但在更多的数据上训练过的模型。

5 隐含扩散模型

High-Resolution Image Synthesis with Latent Diffusion Models
Robin Rombach 等，慕尼黑大学、海德堡大学和 Runway 团队

MidJourney, Dall-E 和 Imagen 等模型所创造的精美的图片都有一个重要的共同点，它们都依赖于扩散模型。研究人员开发了一种新的图像合成方法，称为隐含扩散模型 （latent diffusion models ），可以在一系列任务中获得最先进的结果。

6 缩放自回归

Scaling Autoregressive Models for Content-Rich Text-to-Image Generation
Jiahui Yu 等，谷歌研究团队

谷歌介绍了一种自回归文本到图像生成模型 Parti（Pathways Autoregressive Text-to-Image model），可实现高保真照片级图像输出，并支持涉及复杂构图和丰富知识内容的合成。谷歌还推了另一个文本到图像生成模型 Imagen。Parti 和 Imagen 分别是自回归模型和扩散模型，两者不同但互补，代表了谷歌的不同探索方向。

7 涌现

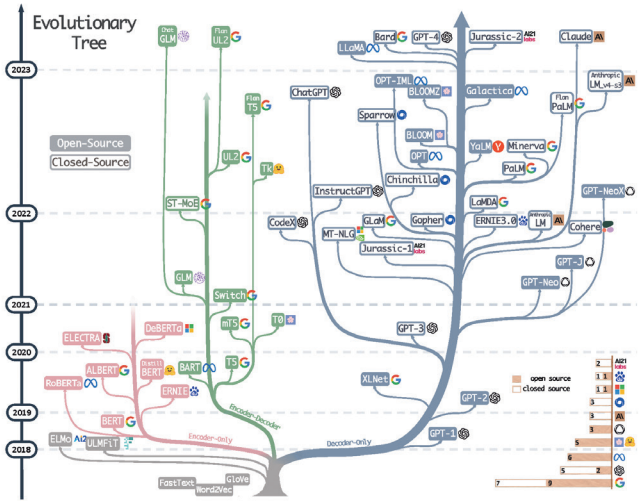
Emergent Abilities of Large Language Models
Jason Wei 等，谷歌、斯坦福、DeepMind 等

事实证明，扩展语言模型可以显著提高各种下游任务的性能和样本效率。本文讨论了一种不可预测的现象，“我们将其称为大型语言模型的涌现能力。” 如果一种能力不存在于较小的模型中但存在于较大的模型中，则认为该能力是涌现的。因此，不能简单地通过推断较小模型的性能来预测新兴能力。这种现象的存在提出了一个问题：额外的扩展是否有可能进一步扩大语言模型的能力范围。当这篇文章发表时，OpenAI 已经完成了 GPT-4 的预训练。

8 通用智能体

A Generalist Agent
Scott Reed 等， DeepMind

这个被称为 Gato 的多模态模型，可以完成 600 多种任务，被称为通用智能体，引起了小小的轰动。科技圈内一度以为这就是未来通用人工智能（AGI）的方向。 Gato 同时学习多个不同的任务，这意味着它可以在这些任务之间切换，而不必在学习另一项技能之前忘记一项技能。这是一个小小的进步，但意义重大。



自然语言大模型的进化， 2017，Transformer 诞生；
2020，GPT-3 发布，进入大模型时代；
ChatGPT 开启了通用人工智能（AGI）时代

9 语音识别，99 种语言

Robust Speech Recognition via Large-Scale Weak Supervision
Alec Radford，Ilya Sutskever 等，OpenAI

OpenAI 开源的语音识别系统：Whisper，支持 99 种语言的语音识别（包括汉语），其中英语的识别能力几乎接近人类。OpenAI 联合创始人 & 首席科学家 Ilya Sutskever 说：“终于有一个可靠的语音识别系统能听懂我的口音了。”

10 思维链

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
Jason Wei 等，谷歌研究，大脑团队

“模型能力可以通过扩大参数规模来提升，但是研究人员却另辟蹊径，想了一个更好的方法，即把一个复杂问题分解为多步推理的简单问题，让模型能够明白并学习人类到底是怎么一步一步推导到这个答案的，这个就叫作思维链。”

“大型语言模型经过海量预训练之后，已经见过了很多推理方式，我们只需要一步一步引导，就可以让它按照自己想要的方式去推理。思维链方式进一步解放了模型潜力，让本来模型不会解的一个个复杂问题分解成很多简单问题，然后通过逐一解决简单问题，最终使复杂问题迎刃而解。”（邱锡鹏）

大模型的“慢思考”

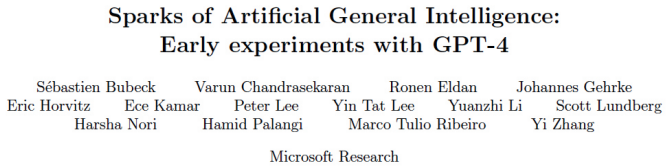
在生成式 AI 的各种基础模型中，GPT-4 至今仍代表了最高的水准，人工智能研究者们还在忙于发表测试的论文与报告，试图理解涌现出来的智能。

微软的测试论文指出：GPT-4 展示出比以前的 AI 模型更具普适性的智能。我们讨论了这些模型不断提升的能力和其所带来的影响。我们展示，除了掌握语言之外，GPT-4 能够在数学、编码、视觉、医学、法律、心理学等领域解决新颖且困难的任务，而无需任何特殊提示。此外，在所有这些任务中，GPT-4 的表现与人类水平的表现非常接近，往往远远超过了 ChatGPT 等先前的模型。鉴于 GPT-4 的广度和深度，我们认为它可以合理地被视为人工通用智能（AGI）系统的早期（尽管仍不完整）版本。[1]

该研究提出了，未来大模型需要解决的一些问题，也构成了研究的方向：信心校准、长期记忆、持续学习、个性化、规划和概念跨越、透明度、认知谬误和非理性、对输入的敏感性挑战，等等。而过去半年最重要的研究方向，是理解和破解大模型神秘而又令人兴奋的智能“涌现”。大模型既需要超越对下一个词的预测能力，也需要一个更丰富、更复杂的“慢思考”深层机制，来监督“快思考”预测下一个词的机制。

预训练几乎可以产生所有大模型的知识，只需要有限的指令调整数据，就可以指导模型产生高质量的输出。[2] 而调动大模型的智能，发现其泛化能力的新领域，可以用更有效率的方式，如用直接偏好优化（DPO）的算法，训练和微调的过程大为简化。[3]

可以说大模型的成功，在于捕捉词汇之间的大量统计相关性，但实验表明，大模型在发现因果关系的表现方面，有时甚至不及随机猜测。[4] 克服这些局限，还是需要继续引导大模型正确的思考方法，或者借助外部的资源。一种新的语言模型推理框架，“思想之树”(ToT)，在流行的“思想链”方法的基础上进一步发挥，允许大模型通过考虑多个不同的推理路径和自我评估选择来进行深思熟虑的决策，以决定下一步的


Sparks of Artificial General Intelligence:
Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

行动方向，以及在必要时进行预见或回溯以做出全局选择。[5] 此外，还有大模型可以自己编写 API 调用，这些生成和执行代码的能力，可以减轻幻觉问题，增加了输出的可靠性和适用性。但也可能带来一些控制大模型方面的风险。[6]

还有研究人员提出了基于 Transformer 训练出来的推理模块，可以在大模型上即插即用，改善其推理能力。[7]

大型语言模型理解人类常识推理，还往往取决于其“情商”，即理解人类的信念、目标和心理状态，这被称为心智理论（ToM）任务。适当的提示可以提升大模型的心智推理能力（甚至共情能力），对上下文的依赖非常重要。[8]

此外，研究员发现了节省计算资源的训练方法，有的能提升 2 倍的效率。[9]

最后，是训练大模型的数据问题：由人类原生的数据，将来可能会越来越稀缺；高质量的自然语言数据，最快有可能到 2026 年就被大语言模型耗尽。[10]

一项针对数据众包的研究，发现其中 30%-40% 来自承包者使用大模型获取的数据。这就产生了大模型喂自己数据的问题，就像一条蛇，它咬住了自己的尾巴，要把自己整个吞下。[11]

但随着大模型在人们生活和工作中作用日益重要，合成数据在大模型训练中的数据来源占比越来越大。如用扩散模型的合成数据，可以提升 ImageNet 中分类的准确度。[12]

1. Sparks of Artificial General Intelligence: Early experiments with GPT-4 2. LIMA: Less Is More for Alignment 3. Direct preference optimization: your language model is secretly a reward model 4. Can Large Language Models Infer Causation from Correlation? 5. Tree of Thoughts: Deliberate Problem Solving with Large Language Models 6. Gorilla : large languge models connected with massive APIs 7. Tart: A plug-and-play Transformer module for task-agnostic reasonin 8. Boosting Theory-of-Mind Performance in Large Language Models via Prompting 9. Sophia: a Scalable Stochastic Second-Order Optimizer for Language Model Pre-training 10. Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning 11. Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks 12. Synthetic Data from Diffusion Models Improves ImageNet Classification

开源

Meta 于 2 月份推出了一个开源大模型 LLaMA，这个模型原本只对从事大模型研究社区开放，但很快在社交网站上泄露，迅速流行开来。已经被公认超过了 GPT-3。在此基础之上微调，一个开源模型的“羊驼家族”迅速繁衍。与此同时，一份谷歌内部文件泄露，称面对正在兴起的开源大模型，闭源大模型并没有任何门槛。这样，开源大模型能否达到闭源大模型的水平，如何实现大模型技术的民主化，更垂直、更小型、更个人化的模型，以及各种测试与研究，形成了一波研究热点。

开源模型四年来进步不大?

大模型层出不穷，开源的模型更是令人目不暇接，但是这些模型的水平如何？需要严谨科学的测试。阿里的达摩院和新加坡国立大学的研究团队，用 2019 年的 T5 开源大模型与当下比较流行的开源模型进行测试比较，结果显示：写作能力有提升，但在解决问题和对齐方面还有差距。四年了，开源的模型似乎并没有明显的进步，目前的开源社区已经展开了疯狂的模型开发，但也要建立起对其表现评价的标准。[1]

模仿不是开源的出路

能否通过模仿大模型，让较弱的开源模型获得闭源大模型应用（如 ChatGPT) 的能力？ 研究者对一系列模仿 ChatGPT 的语言模型进行了微调，使用不同的基础模型大小（1.5B 至 13B）、数据源和模仿数据量（0.3M 至 150M tokens）。然后使用众包评估者和经典的自然语言处理基准对这些模型进行评估。

最初，模仿模型的输出质量有些惊艳——它们在遵循指令方面表现出色，输出与 ChatGPT 相媲美。然而，在进行更有针对性的自动评估时发现，在没有大量模仿数据支持的任务中，模仿模型在缩小基础模型与 ChatGPT 之间的差距方面几乎没有任何作用。模仿者只擅长模仿 ChatGPT 的风格，但无法模仿其真实性。

总体而言，模型模仿是一个虚假的承诺：开源模型与闭源模型之间存在着相当大的能力差距，当前的方法只能通过更大量的模仿数据或使用更强大的基础模型来弥合这一差距。因此，改进开源模型的最有效策略是开发更好的基础模型，而不是采取模仿闭源大模型的捷径。[2]

要模仿，就模仿推理

小模型利用大模型生成的输出，来对自己进行解释调整，这种模仿学习，看起来能让增强小模型事半功倍。但也要看情况。如果小模型只是获得大模型浅层输出的有限模仿信号、规模较小且同质化的训练数据，以及缺乏严格的评估导致高估能力，小模型往往只学习模仿大模型的风格而不是推理过程。为了解决这些挑战，微软团队开发了 Orca，一个拥有 130 亿参数的模型，学习模仿 GPT-4 的推理过程。

这样，小模型获得了丰富的信号，包括解释痕迹、逐步思考过程和其他复杂指令，同时借助 ChatGPT 的指教，还利用了大规模和多样化的模仿数据进行谨慎的采样和选择。结果在一些测评和专业考试中，Orca 胜过了最好的开源模型、达到了 ChatGPT 的水平，接近了 GPT-4 的水平。[3]

华盛顿大学博士生 Tim Dettmers 带领的团队，提出了一种高效的微调方法 QLORA，足够降低内存使用量，能在单个 48GB 的 GPU 上微调一个有 650 亿参数的模型。[4]

当音乐不再是“天籁”

Meta 在 GitHub 上以开源方式发布了 AI 音乐生成模型 MusicGen 的代码，该 AI 模型利用 Transformer 架构，可以根据文本和旋律提示创作音乐。与 Riffusion、Mousai、MusicLM 和 Noise2Music 等其他音乐模型相比，MusicGen 在音乐与文本之间的匹配度以及作曲的可信度等客观和主观指标上表现得更加优异。[5]

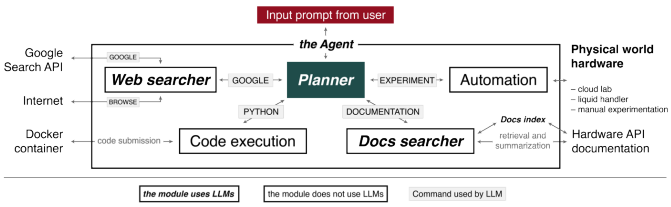
1. INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models 2. The False Promise of Imitating Proprietary LLMs 3. Orca: Progressive Learning from Complex Explanation Traces of GPT-4 4. Qlora: Efficient Finetuning of Quantized LLMs 5. Simple and Controllable Music Generation

智能代理

使用大型语言模型作为核心控制器构建代理是一个很酷的新兴概念。除了下述论文之外，另外有几个概念证明演示，如 AutoGPT，GPT-Engineer 和 BabyAGI，都是鼓舞人心的例子。大模型的潜力超越了生成优秀的复制品、故事、论文和程序；它可以被构架为一个强大的通用问题解决器。

科学研究的智能助理

来自卡内基梅隆大学的研究人员提出了一个 Intelligent Agent（以下简称 Agent）系统，结合了多个大型语言模型，用于自主设计、规划和执行科学实验。[1]



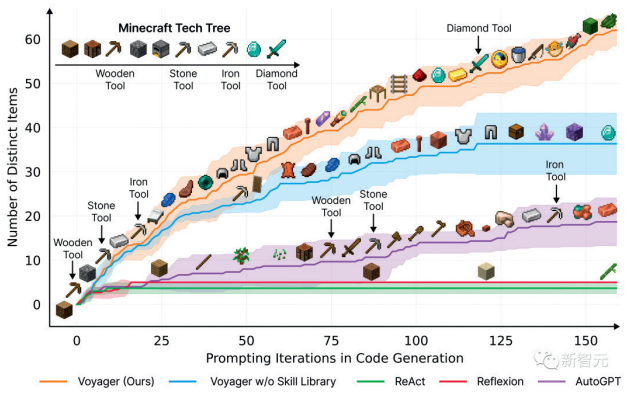
系统架构概述。代理由多个模块组成，这些模块交换消息。其中一些模块可以访问 API、互联网和 Python 解释器。

模拟人生，模拟社会

智能代理除了帮助人类完成较复杂的任务之外，代理之间也可能产生互动。生成代理（generative agents）是一种模拟逼真人类行为的计算软件代理。为了实现生成代理，需要一种架构，将大型语言模型扩展到使用自然语言存储代理经历的完整记录，随着时间的推移，将这些记忆综合成更高层次的反思，并动态地检索它们以规划行为。斯坦福和谷歌的研究团队实例化了生成代理，在一个受《模拟人生》启发的交互式沙盒环境里，用户可以使用自然语言与 25 个代理居民进行交互。在评估中，这些生成代理产生了可信的个体行为和群体行为：例如，仅从用户指定的一个概念开始，即一个代理想要举办情人节派对，代理在接下来的两天内自主地传播派对的邀请，结识新朋友，相互约会参加派对，并协调好在正确时间一起出现在派对上。实验证明了代理架构的观察、规划和反思组件对于代理行为的逼真性至关重要。通过将大型语言模型与计算机交互代理融合在一起，这项工作引入了一种架构和交互模式，实现了逼真的人类行为模拟。[2]

游戏中的生命体：活到老，学到老

Voyager 是第一个由大语言模型驱动、可以终身学习的具身智能体。英伟达团队在之前关于代理在 Minecraft 中玩游戏的研究基础上进行了改进。他们利用 GPT-4 为代理开发了一个课程和构建工具库的方法。这极大地加快了学习速度，并带来了更高质量的解决方案。它可以利用 GPT-4 不停地探索世界，开发越来越复杂的技能，并始终能在没有人工干预的情况下进行新的发现：发现新物品、解锁 Minecraft 技术树、穿越多样化地形，以及将其学习到的技能库应用于新生成世界中的未知任务方面，Voyager 表现出了优越的性能。[3]



VOYAGER 通过自我驱动的探索不断发现新的 Minecraft 物品和技能，显著超过了基线。X 轴表示提示迭代的次数。

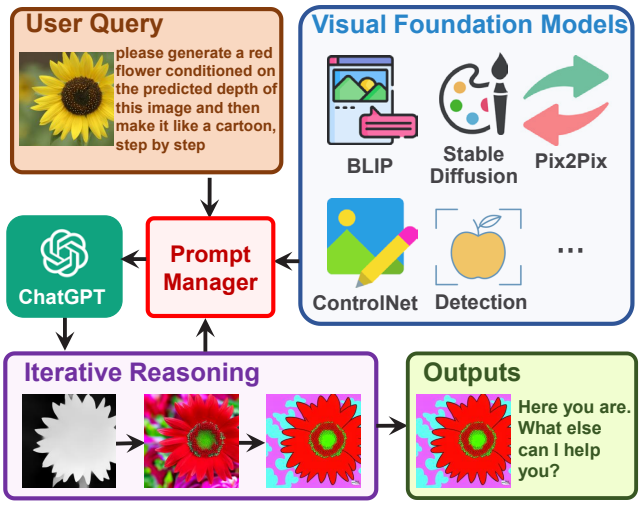
1. Emergent Autonomous Scientific Research Capabilities of Large Language Models 2. Generative Agents: Interactive Simulacra of Human Behavior
3. Voyager: An Open-Ended Embodied Agent with Large Language Models

多模态

多模态指的是机器学习模型可以处理和理解多种类型的数据，如文本、图像、音频和视频等。在现实世界中，信息是通过多种模态传递的，因此一个能处理和理解这些不同类型数据的模型，将更能理解和处理实际问题。多模态能力也是提升 AI 与人类交互能力的关键。如何有效地整合和处理不同类型的数据，以及如何在不同的模态之间转换和翻译等，这些都是当前和未来研究的重要课题。

聊天对话框，一个新的图形界面？

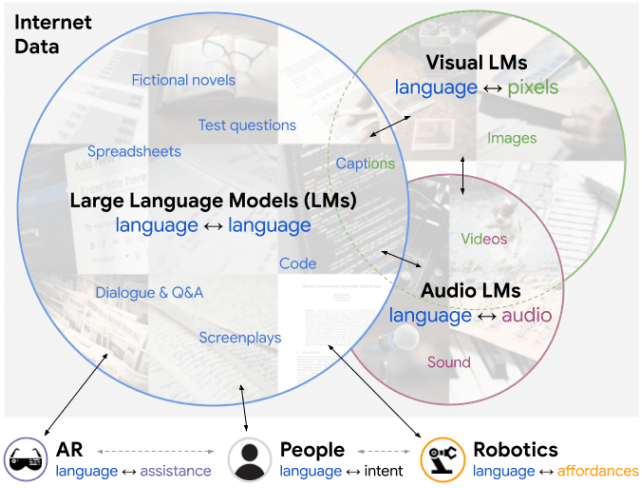
由于 ChatGPT 是通过语言进行训练的，它目前还无法处理或生成来自视觉世界的图像。与此同时，虽然诸如 Visual Transformer 或 Stable Diffusion 等视觉基础模型展示了极佳的视觉理解和生成能力，但它们只是在特定任务上的专家，需要一轮固定输入和输出。为此，微软团队构建了一个名为 Visual ChatGPT 的系统，集成了不同的视觉基础模型，使用户能够通过 1) 发送和接收不仅是语言，还有图像 2) 提供复杂的视觉问题或需要多个 AI 模型多步协作的视觉编辑指令 3) 提供反馈并要求纠正结果。研究团队设计了一系列提示，将视觉模型信息注入 ChatGPT，考虑到需要多个输入 / 输出和需要视觉反馈的模型。实验显示，Visual ChatGPT 在视觉基础模型的帮助下，为研究 ChatGPT 的视觉角色开启了大门。[1]



Visual ChatGPT 的架构

寻找多模态之间的共同语言

不同模态的模型在不同的领域存储不同形式的常识知识，谷歌团队展示出这种多样性是互补的，可以通过苏格拉底模型（SMs）来利用：一个模块化框架，可以通过多模态提示（即零样本）来组合多个预训练模型，以便彼此交换信息并捕获新的多模态能力，无需进行微调。在最小的工程改动下，SMs 不仅能与最先进的零样本图像标注和视频到文本检索竞争，而且还能启用新的应用，例如 (i) 回答关于以自我为中心的视频的不拘形式的问题，(ii) 通过接口与外部 API 和数据库 (例如，网络搜索) 进行多模态辅助对话与人交流 (例如，烹饪食谱），以及 (iii) 机器人的感知和计划。[2]



预训练的大型基础模型，在不同领域训练，学习了互补的常识，语言扮演了中介表示，让这些模型能够彼此交流，为了完成多模态任务而生成联合预测，而不需要微调。可以添加新的应用，如增强现实（AR）、人类反馈、机器人等，参与到多模态讨论中。

大一统：从多模态到高模态

由卡内基梅隆、密西根和 DeepMind 组成的团队，研究了高模态场景的高效表示学习，结果是一个单一模型 HighMMT，扩展到 10 种模态（文本、图像、音频、视频、传感器、本体感知、语音、时间序列、集合和表格）和来自 5 个不同研究领域的 15 项任务。HighMMT 表现出至关重要的行为放大效应：每增加一种模态，性能就会继续提高，并且在微调期间，它将转移到全新的模态和任务。[3]

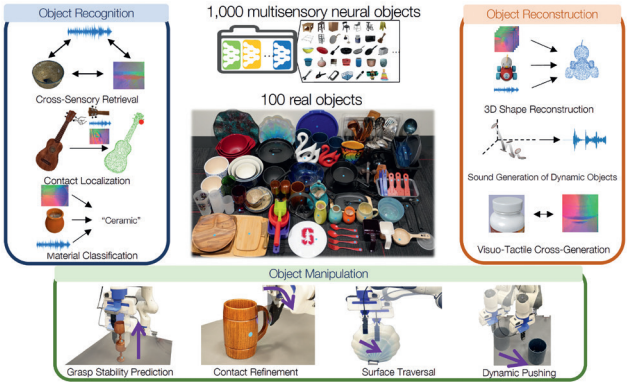
1. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models 2. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language 3. HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Representation Learning

具身智能

具身智能指的是 AI 系统能够通过与环境的物理交互来理解和学习的能力。这对于生成式 AI 来说非常重要，因为它扩大了 AI 系统可以处理和生成的数据类型和范围。与处理抽象数据相比，具身智能可以帮助生成式 AI 更好地理解和处理现实世界的复杂性和多样性。然而，具身智能也带来了一些挑战，包括如何在物理环境中进行高效的学习，如何处理和解决实际环境中的不确定性，以及如何保证在与环境交互过程中的安全性等。这些都是当前和未来研究的重要课题。

一个具身的多模态大语言模型

谷歌团队提出了一个具身的语言模型，以直接将真实世界的连续感知模态纳入语言模型，从而建立词语和感知之间的联系。这个具身语言模型的输入是多模态句子，其中交错了视觉、连续状态估计和文本输入编码。研究团队将这些编码进行端到端的训练，结合预训练的大型语言模型，用于多个实体任务，包括顺序机器人操控规划、视觉问题回答和字幕添加。评估表明，这个被称为 PaLM-E 的单一的大型实体多模态模型，可以处理各种具身推理任务，来自各种观察模态和在多个具身上，并且进一步表现出积极的转移：该模型从跨互联网规模的语言、视觉和视觉 - 语言领域的多样化联合训练中获益。除了在机器人任务上进行训练外，还是视觉 - 语言通用型模型。[1]



OBJECTFOLDER BENCHMARK 测试套件包含 10 个用于多感官对象中心学习的基准任务，围绕对象识别、重建和操纵展开。作为对 OBJECTFOLDER 中 1000 个多感官神经对象的补充，团队还引入了 OBJECTFOLDER REAL，它包含从 100 个真实世界物体中收集的真实多感官数据，包括它们的 3D 网格、视频录制、冲击声音和触觉读数。

视觉、听觉、触觉，真实世界的多维感知

李飞飞等研究人员提出了 OBJECTFOLDER BENCHMARK，这是一个围绕物体识别、重构和操作的 10 个基准任务的套件，旨在推动多感官物体为中心学习的研究。团队还介绍了 OBJECTFOLDER REAL，这是第一个包含 100 个真实室内物体的视觉、声音和触觉实际测量数据的数据集。团队希望其新数据集和基准套件能够作为基石，促进多感官物体建模和理解方面的进一步研究和创新。[2]

人类已经训练出机器人打败了李世石，可以训练出一个机器人胜过梅西吗？

Google Deepmind 和牛津的团队，使用深度强化学习训练了一个具有 20 个驱动关节的人形机器人，使其能够玩简化的一对一足球比赛。首先独立训练了各个技能，然后在自我对抗的环境中将这些技能端到端地组合起来，展示了运动技能，如快速摔倒恢复、行走、转身、踢球等，并以平稳、稳定和高效的方式在动作之间进行过渡，远远超出了对机器人的直观预期。机器人还发展出了对游戏的基本战略理解，学会了预测球的移动并封堵对手的射门等。这些行为全都是从一组简单的奖励中出现的。训练在模拟环境中进行，并在实际机器人上进行了零样本迁移。[3]

智能驾驶即智能规划

现代自动驾驶系统的特点是顺序性的模块化任务，即感知、预测和规划。上海人工智能实验室等组成的研究团队，追求最终目标，即自驾车的规划。他们重新审视了感知和预测中的关键组件，并优先考虑了这些任务，以使所有这些任务都能为规划做出贡献。他们引入了统一的自动驾驶（UniAD），这是一个全面的框架，它在一个网络中集成了全栈驾驶任务。它精心设计以利用每个模块的优点，并从全局视角提供互补特性抽象以进行代理人交互。任务通过统一的查询接口进行沟通，以便互相促进规划。他们还实例化了 UniAD。（获 CVPR 2023 最佳论文）[4]

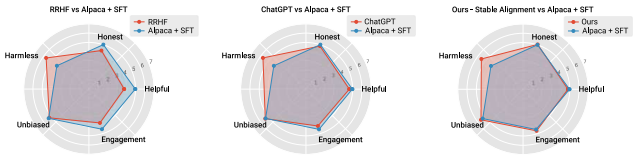
1. PaLM-E: An Embodied Multimodal Language Model 2. The OBJECTFOLDER BENCHMARK: Multisensory Learning with Neural and Real Objects
3. Learning Agile Soccer Skills for a Bipedal Robot with Deep Reinforcement Learning 4. Planning-oriented Autonomous Driving

安全与可信

ChatGPT，展示了机器生成的文本与人类生成的内容无法区分的能力，但这一“后图灵测试”时代的来临，带来了新的挑战。通用人工智能成功记忆人类知识，并不能保证模型能按照人类的期望来执行。其实早在 ChatGPT 推出之前，已经有研究揭示了一些大模型内部的行为异常，包括生成有害内容、强化偏见和传播虚假信息。提高期望的社会行为和抑制不期望的社会行为，这一过程，通常被称为“社会对齐”（social alignment）。

大模型的模拟社会互动

与人类不同，人类通过社会互动达成关于价值判断的共识，而当前的语言模型则是在孤立中训练以僵化地复制其训练语料库，导致在不熟悉的场景中的泛化表现不佳，并容易受到对抗性攻击。这项工作提出了一种新的训练范式，允许大模型从模拟的社会互动中学习。与现有的方法相比，这一方法更具可扩展性和效率，在对齐基准测试和人类评估中表现出卓越的性能。这种在大模型训练中的范式转变，有助于开发能够稳健准确地反映社会规范和价值观的 AI 系统。[2]

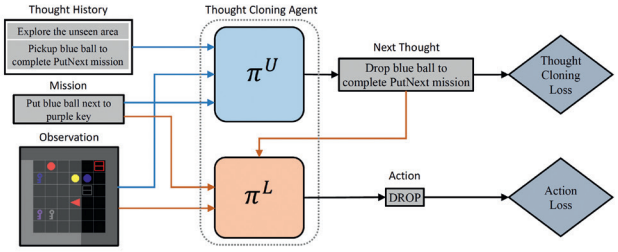


人类评估结果。参与者（n = 206）被要求在 7 点 Likert 量表上对回应在有帮助、诚实、无害、无偏见和参与度方面进行评分。

学说人话，也要学习用人话思考

我们不仅训练大模型说人话，更重要的是训练大模型像人一样，用语言思考。这个研究团队的成员之一，来自辛顿的 Vector 研究所，并且是加拿大 CIFAR AI 的主席。研究团队认为，强化学习（RL）代理远未达到人类在这些能力上的水平。假设这种认知缺陷的一个原因是他们缺乏用语言思考的好处，可以通过训练它们像人类一样思考来提升 AI 代理的能力。

研究团队引入了一种新的模仿学习框架，称为“思维克隆”，其理念不仅是克隆人类示范者的行为，而且还要克隆人类在执行这些行为时的思维。虽然研究者认为在互联网规模的数据集上，人们在行动时大声思考（例如，带有文字记录的在线视频），思维克隆将真正出彩，但在这里，他们的思维和行动数据都是在人工生成的领域进行实验。结果表明，思维克隆的学习速度远超行为克隆，其性能优势在分布测试任务的情况下越发显著，凸显了其更好地处理新情况的能力。



思维克隆（TC）的总体框架。TC 代理有两个部分：上层部分和下层部分。在每个时间步，TC 代理接收观察结果、任务以及思维历史作为输入。上层部分产生思维，下层部分根据这些思维产生行动。生成的思维和行动与示范数据集中的实际情况进行比较，以计算损失。

思维克隆还为 AI 的安全性和可解释性提供了重要的好处，并使得调试和改进 AI 变得更加容易。因为我们可以观察到代理的思维，我们就能（1）更容易地诊断问题所在，使得修复问题变得更加容易，（2）通过纠正其思维来引导代理，或者（3）防止其执行不安全的计划。总的来说，通过训练代理如何思考以及行为，思维克隆创造出更安全、更强大的代理。[2]

保护版权，数据来源透明，水印基本可靠

随着大语言模型变得普遍，机器生成的文本有可能充斥互联网，带来垃圾邮件、社交媒体机器人和无价值的内容。水印技术使大模型生成的文本可以被检测和记录，可以减轻这些危害。然而，一个关键的问题仍然存在：在实际情况下，水印技术的可靠性如何？水印文本可能被修改以适应用户的需求，或者被完全重写以避免检测。研究发现，即使在经过人工和机器转述后，水印仍然可以被检测出来。虽然这些攻击稀释了水印的强度，但转述在统计上很可能会泄露出原文的 n-grams（词汇序列模式）或者更长的片段，当观察到足够的词元时，会产生高置信度的检测。例如，即使对大量的人类转述，平均观察 800 个词元后，可以检测到水印，误报率设置为 1e − 5。研究还考虑了一系列新的检测方案，这些方案对嵌入在大型文档内的短跨度水印文本敏感。[3]

1. Training Socially Aligned Language Models in Simulated Human Society 2. Thought Cloning 3. On the Reliability of Watermarks for Large Language Models

第三章 监管、安全与人才

所有监管的核心，都是在充分利用生成式人工智能技术造福人类能力、提升竞争优势的同时，对其风险加以管理和控制。

中美欧监管

进入 2023 年，中国、美国、欧盟都加快了人工智能的监管和立法进程。主要是因为 ChatGPT 的推出和 GPT-4 为代表的大模型开启了通用人工智能的大门，大模型更加强大，创新加速，新技术向各行各业渗透，人工智能给人类带来福祉和风险会同步放大。所有的监管的核心，都是在充分利用人工智能技术造福人类能力、提升竞争优势的同时，对其风险加以管理和控制。

中国非常重视人工智能的监管，已经针对人工智能应用的不同场景和突出问题，推出了数个管理条例，如算法推荐的透明、内容的深度合成、生成式人工智能管理等。关于人工智能的立法，已经提上 2024 年的议程。

中国在上半年开始施行《互联网信息服务深度合成管理规定》，网信办发布了《境内深度合成服务算法备案清单》，《生成式人工智能服务管理暂行办法》于 8 月 15 日施行，全国信息安全标准化技术委员会公布国家标准《信息安全技术人工智能计算平台安全框架》征求意见稿。科技部也公布了《科技伦理审查办法（试行）》，目前已经完成征求意见。

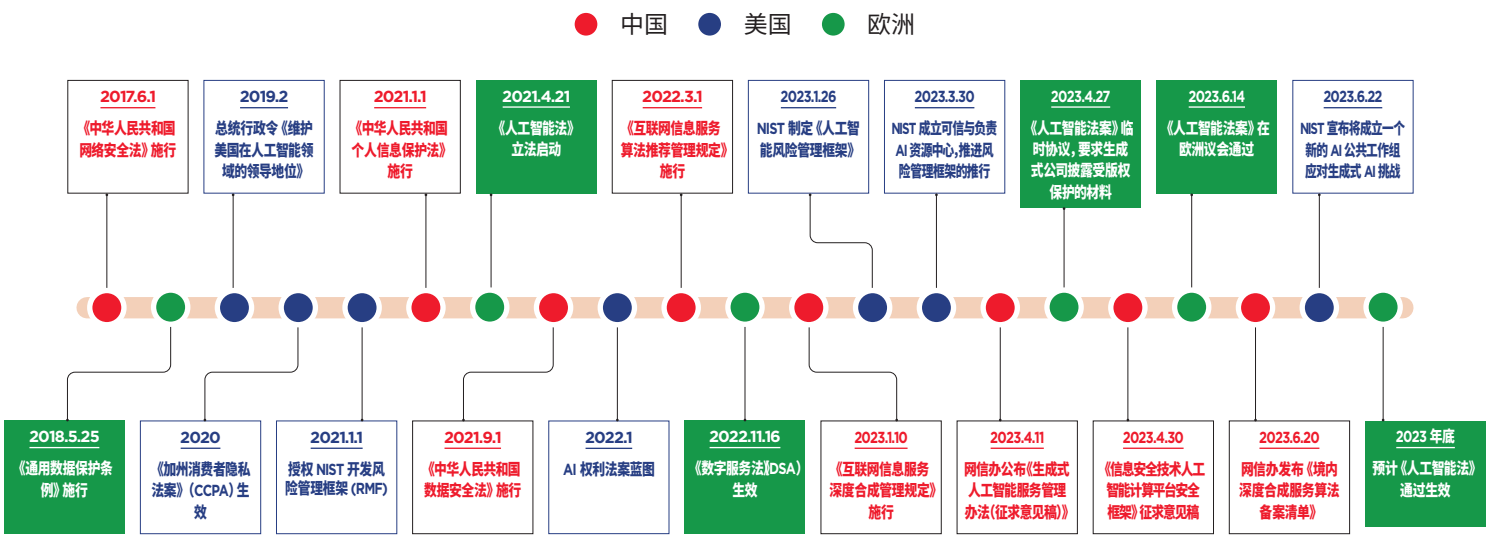
如果参照此前有关个人信息和数据安全的相关立法进程，对人工智能的监管和立法，正在从针对性地回应特定人工智能领域的治理难题阶段，进入到推出全国综合性的立法阶段。

根据《国务院 2023 年度立法工作计划》，人工智能法草案等预备提请全国人大常委会审议。在 2017 年 7 月国务院发布的《新一代人工智能发展规划》中，预计 2025 年初步建立人工智能法律法规、伦理规范和政策体系，形成人工智能安全评估和管控能力。现在看来，这一规划中的监管立法进程可能会提前完成。

欧盟继率先实施《通用数据保护条例》（GDPR) 之后，又率先启动综合性的人工智能的立法，《人工智能法案》（AI ACT）目前已经在欧洲议会获得通过，再经过欧盟理事会和欧盟委员会，得到各成员国的批准，可能在 2023 年底或 2024 年初正式生效。这将对全球的人工智能的监管产生重大影响，《人工智能法案》建立起了三级风险体系。而这种风险体系，也是围绕着对个人权利和人类福祉可能造成的侵害程度来制定的。不过，由于美国与中国在人工智能领域发展领先于欧盟，欧洲业界担心，以目前试图无所不包的立法草案，在人工智能这个人类尚未真正理解其含义的领域，是否会影响产生真正的创新。

美国短期内不会出现全国性的人工智能综合立法，目前主要依靠既有的相关法律进行监管。美国商务部的国家标准与技术研究所（NIST）年初发布了第一版《人工智能风险管理框架》（AI RMF），目标是帮助设计、开发、部署或使用人工智能系统的组织和机构，提高人工智能风险管理的能力，并促进发展可信和负责任的人工智能。AI RMF 只是一个自愿性的技术框架。美国联邦政府正在把 NIST 定为人工智能风险管理的资源中心和国际合作机构。

中美欧立法进程对比



地方的 AI 雄心

人工智能产业政策正在各地密集出台。北京、上海与深圳处于第一梯队，杭州、南京、苏州、成都等多个城市处于第二梯队。

北京明确在文件标题中提到了“通用人工智能”。到 2025 年，北京计划基本建成可有力支撑数字经济高质量发展的通用人工智能产业发展格局，以及具有全球影响力的人工智能创新策源地。届时，算力芯片等基本实现自主可控，通用人工智能雏形显现。

深圳最为国际化，充分利用区内跨境的优势，多份文件提到了要依托前海深港现代服务业合作区、河套深港科技创新合作区、光明科学城等区域，建立与国际接轨的科研管理制度，探索实施更加开放、便捷的国际组织注册制度，吸引港澳台以及国际人工智能高端创新要素聚集。

上海重视对战略产业的 AI 赋能，计划加大在战略性新兴产业项目中对人工智能产业技术创新的布局，促进智能机器人、智能网联汽车、无人机、无人船、医疗器械、药物研发，以及金融与物流等产业应用。

第一梯队城市均围绕算力、数据、产业上下游生态加以布局，成为跨区域协作的中心。第二梯队城市的能级就要稍弱一些。随着全国范围内的算力统筹与数据开放，它们同样拥有机会。创新人才与应用场景会成为制约这些城市做大人工智能产业规模的关键因素。

成都是首个在这波 AI 大模型浪潮中公布政策的西部地区城市，计划到 2025 年，全市人工智能产业产值突破 1500 亿元；南京则提出到 2025 年，全市人工智能核心产业收入超过 500 亿元。

第一梯队城市在大模型领域的优势与布局

	算力	数据	产业	生态	跨区域协同
北京	• 海淀区、朝阳区建设北京人工智能公共算力中心、北京数字经济算力中心； • 提高环京地区算力一体化调度能力	• 归集高质量基础训练数据集； • 谋划建设数据训练基地； • 搭建数据集精细化标注众包服务平台； • 推动公共数据和社会数据定向有条件开放 • 用好北京国际大数据交易所社会数据专区	• 政务服务、医疗领域、科学研究、金融领域、自动驾驶、城市治理； • 聚焦本市虚拟数字人、数字医疗、电商零售等创新活跃的数据优势领域； • 发挥中关村先行先试“试验田”作用	通用人工智能产业创新伙伴计划	加强与天津市、河北省、山西省、内蒙古自治区等区域的算力合作
上海	• 临港新片区算力产业生态； • 此外，三角枢纽节点（青浦区为起步区）、G60 科创走廊、金山等枢纽型数据中心集群建设	• 推动人工智能领域高质量数据集建设； • 在经济发展、民生服务、城市治理等领域建立公共数据动态开放清单； • 鼓励企业通过上海数据交易所开展数据产品交易	• 加大在战略性新兴产业项目中对人工智能产业技术创新的布局； • 推动智能机器人、智能网联、无人机、人工智能医疗器械关键技术研发； • 制定并定期更新人工智能示范应用清单； • 浦东新区应当发挥人工智能创新应用先导区的作用	临港新片区智算产业联盟	长三角人工智能产业协同
深圳	• 建设城市级智能算力平台，鹏城云脑III项目 2023 年年底前启动建设； • 打造大湾区智能算力枢纽，谋划共建粤港澳大湾区智能算力统筹调度平台。	• 2023 年年底前出台公共数据开放管理办法、公共数据资源目录，制定公共数据开放计划； • 进一步做大深圳数据交易所交易规模	聚焦通用大模型、智能算力芯片、智能传感器、智能机器人、智能网联汽车等领域	组建深圳市 AI 教育联盟和 AI 讲师团	鼓励企业依托河套深港科技创新合作区、前海深港现代服务业合作区或海外研发中心，研发基于国际主流大模型的创新产品，积极拓展国际市场。

说明：根据公开资料整理，不完全列举。

安全与伦理

有关人工智能安全与伦理的争论，一直伴随着人工智能的发展，即使在人工智能停滞不前的“黑暗时代”，关于人工智能将统治人类的科幻小说照样畅销。而当人工智能真的“通用”起来，其安全与伦理问题，需要从科幻式的高谈阔论，落地为具体可行的政府监管与企业责任。

科技部的《科技伦理审查办法（试行）》，已经于 5 月结束征求公众意见环节，并开始施行。其中提出了人工智能企业，应该接受科技伦理审查；审查主体，应该设立科技伦理（审查）委员会。目前中国公开宣布设立人工智能伦理委员会的科技公司，只有阿里巴巴一家。

该办法也是遵循风险管理的思路，低风险实行简化程序，而高风险的新兴科技活动，实行清单管理和复核制度。其中提到了四类人工智能的高风险科技活动：

- 侵入式脑机接口用于神经、精神类疾病治疗的临床研究。
- 对人类主观行为、心理情绪和生命健康等具有较强影响的人机融合系统的研发。
- 具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发。
- 面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统的研发。

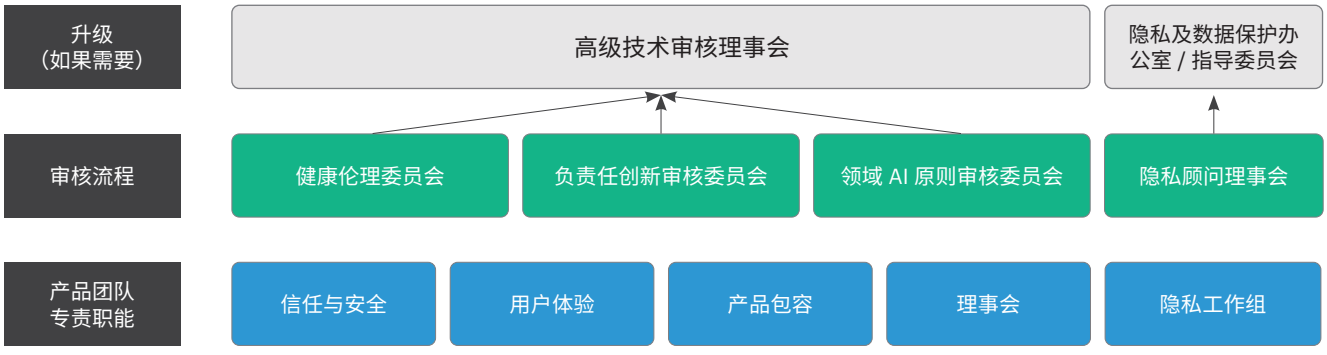
美国较大的人工智能企业，许多都设立部门负责人工智能的安全 / 伦理 / 负责任 / 可信任，尤其重视面向消费者的人工智能产品与服务。但是，去年以来，微软、Meta、谷歌、亚马逊和 Twitter 等公司已经裁减了相关团队的成员。

如 Twitter 在埃隆·马斯克的领导下削减了一半以上的员工人数，其中包括其道德人工智能伦理团队。亚马逊旗下的流媒体平台 Twitch 最近裁减了人工智能伦理团队，让人工智能产品团队直接对与偏见相关的问题负责。2022 年 9 月，Meta 解散了由约 20 名工程师和伦理学家组成的负责任创新团队，该团队负责评估 Instagram 和 Facebook 上的公民权利和道德规范。微软 3 月解散了整个社会与伦理团队，该团队负责人工智能产品的用户体验和整体设计，尤其是将 OpenAI 的大模型技术集成到微软产品中的风险控制。

但微软仍然保留着负责任人工智能团队，制定规则和原则来管理公司的人工智能计划。美国人工智能企业依然重视安全与伦理。微软在这方面的投入实际上增加了。最近的调整，反映出在生成式人工智能发生变革、企业研究与创新竞争加剧的新态势下，人工智能企业正在寻求用更好的研究、技术和更创新的解决方案，来安全和负责地部署新技术。

为了不让大模型对用户和公众完全变成一个“黑箱”，不管是闭源的 OpenAI，还是开源的大模型平台，都把产品的风险披露当成产品发布的标配，就像发布一款新药一样，其中可能的副作用和风险也要在说明书中交待清楚。

在人工智能的安全与伦理问题上，美国的科技企业和社会组织发挥着主导作用。一些人工智能企业在调整内部的安全伦理团队，将其与产品更密切地结合在一起的同时，那些解决大模型的安全与伦理问题的初创企业开始出现了。谷歌拥有一个最完备的人工智能行为原则及治理结构。尽管谷歌的伦理团队内部发生了价值观的冲突，一些人员离职，但谷歌将安全与伦理的原则，内嵌到产品的全生命周期，这样的做法仍然是领先的。



在 Google 的产品团队中嵌入了支持负责任的 AI 实践的专用功能。Google 通过一个三层的内部 AI 原则生态系统，进一步在公司范围内落实这些实践。

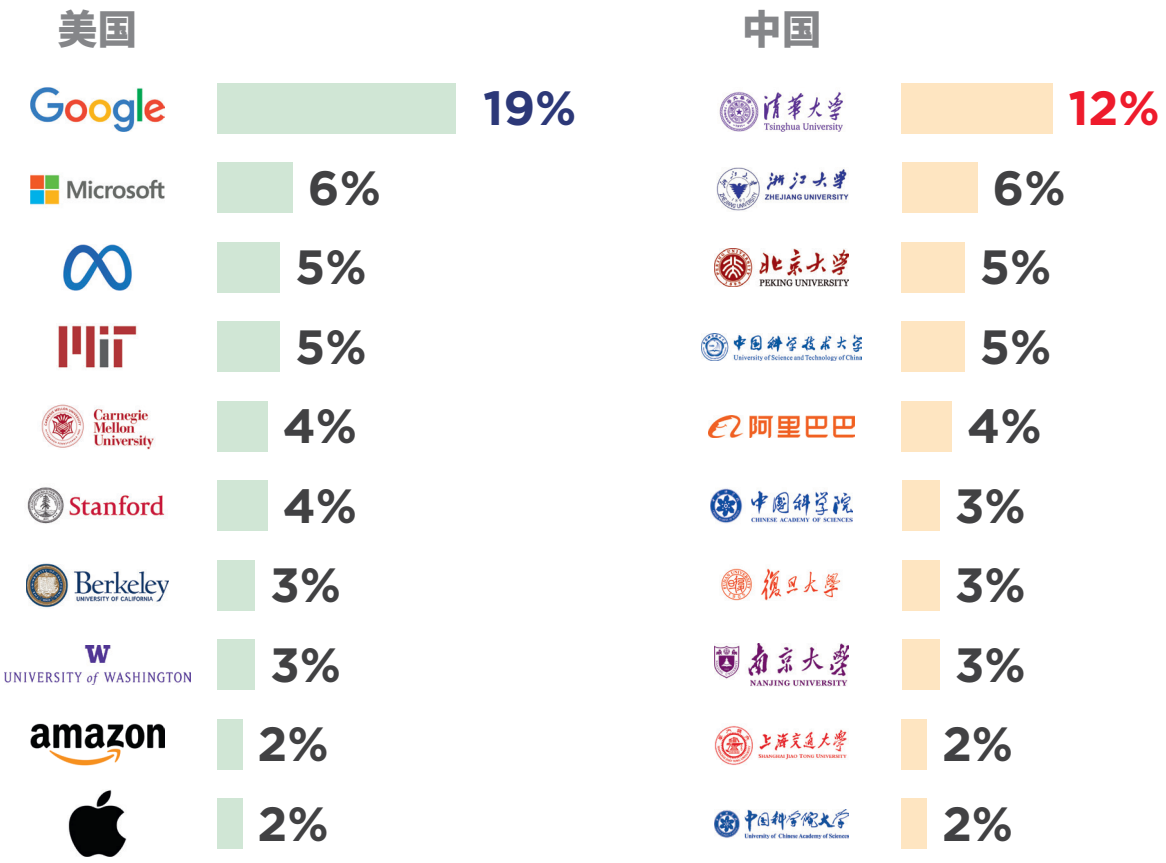
中美塔尖人才

清华大学 AMiner 团队在全球范围内评选过去十年人工智能学科的顶级学者 2000 人次，其中，20 个子领域各 100 名。在全球范围内，人工智能的研究创新，主要由中国与美国引领，英国值得关注。

人工智能的时代才刚刚开始，基础研究的创新主要来自高校。但全球范围内，入选 AI 2000 的来自企业的顶级学者的数量整体呈上升趋势。

在美国，科技企业正在成为推动人工智能研究创新的主要力量。拥有 AI 2000 顶级学者数量排名前 10 的美国机构，5 家企业，5 所高校。谷歌、微软、Meta 位居前三，合计招揽了美国顶级学者的 30%。

中国则由高校承担基础研究的重任。拥有 AI 2000 顶级学者数量排名前 10 的中国大陆机构，9 所高校或相关学术机构。阿里巴巴是国内招揽了最多顶级学者的企业。随着人工智能技术逐步在应用场景落地，未来中国企业也将在全球创新中扮演不可忽视的角色。



说明：占比（%）指该机构入选的 AI 2000 的顶级学者的数量，相对在该国任职的 AI 2000 的顶级学者的比例。仅统计了中国大陆的企业与高校。

从研究到创新

顶尖学者主要在大学等科研机构任职，其次就是为科技巨头的未来服务，还有部分则创办或加入了人工智能相关的初创企业。

DeepMind 与 OpenAI 是最热的初创企业。大量顶尖学者在这两家初创企业公开发表最新进展论文，引领着 AI 社区，推动了创新的扩散。尤其是在 GPT-3 发布后，顶尖学者中排名第 11 的 Noam Shazeer 创办了 Character.ai、排名第 7 的 Manjunath Kudlur 创办了 Useful Sensors Inc.，排名第 98 的 Dario Amodei 创办了 Anthropic、排名第 6 的 Richard Socher 创办了 You.com。这些顶尖学者下场创业，推动了技术创新在应用场景的落地。机器学习是诞生此类初创企业最多的子领域。

在入选了最新的 AI 2000 排名的顶尖学者中，就有 25 人选择了创业作为新的起点。美国仍是创新的中心，约 3/4 的初创企业位于美国。华人仍是创新的核心群体，占了其中的 1/4，与顶尖人才整体占比水平接近。

AI 2000 顶级学者初创企业（GPT-3 发布后成立）的任职情况

领域	学者	机构	年份	国家
安全与隐私	Felix Schuster	Edgeless Systems GmbH	2020	德国
	Petar Tsankov	LatticeFlow	2020	瑞士
机器学习	Noam Shazeer	Character.AI	2021	美国
	Ashish Vaswani	Stealth Startup	2022	美国
	Soumith Chintala	Voltron Data	2021	美国
	Song Han	OmniML	2021	美国
	Richard Socher	You.com	2020	美国
	Jason Yosinski	Windscape AI	2021	美国
	Dario Amodei	Anthropic	2021	美国
数据挖掘	Francesco Bonchi	SOM S.r.l.	2020	意大利
计算机视觉	Andrew Rabinovich	Headroom Inc.	2020	美国
自然语言处理	Richard Socher	You.com	2020	美国
	Fethi Bougares	ELYADATA	2021	突尼斯
	Myle Ott	Character.AI	2021	美国
语音识别	Guoguo Chen	SEASALT.AI	2020	美国
	Jonathan Shen	Inference.io	2020	美国
计算机系统	Manjunath Kudlur	Useful Sensors Inc.	2022	美国
	Georgios Vlachos	Axelar	2020	加拿大
	Amitabha Roy	Kumo.AI	2021	加拿大
芯片技术与工具链	Song Han	OmniML	2021	美国
	Eriko Nurvitadhi	MangoBoost	2022	美国
	Huizi Mao	OmniML	2021	美国
计算机网络	Kiran Joshi	Oma Robotics	2020	美国
	Hongzi Mao	Hologram Labs	2021	美国
	Pankaj Berde	Snickerdoodle Labs	2021	美国

说明：跨领域学者重复统计。

第四章 十大展望

2024 年中国将出现比肩 GPT-4 的多语言通用大模型。

大语言模型

1

2024 年中国将出现比肩 GPT-4 的多语言通用大模型

2

超长上下文（Long Context）将引领下一次 LLM 技术突破

3

在出现更有前景的大语言模型之前，为实现垂直领域更好的效果，以下三种方式将共存：

- 在不改变数据分布的情况下，利用更多通用数据进行通用大模型预训练，不特别引入行业数据；
- 利用行业专属数据微调（Fine-Tuning）通用大模型；
- 利用行业数据占比更高的数据集进行垂直模型预训练

多模态模型

4

当前 CLIP + Diffusion 的文生图模型是过渡态，未来 2 年内将出现一体化的模型结构

5

下一代 Text-to-Image 模型将具备更强的可控性，它将结合底层模型能力和前端控制方式，对模型的设计将注重与控制方式的结合

6

2025 年之前，Video 和 3D 等模态将迎来里程碑式的模型，大幅提高生成效果

7

以 PALM-E 为代表的具身智能（Embodied AI）展现出在机器人的感知、理解和决策等方向上的巨大潜力，但当前训练和可靠性存在较大挑战

8

短期内 Transformer 正成为多个模态的主流网络结构，但压缩整个数字世界的通用方法尚未出现，Transformer 并不是人工智能技术的终点

商业机会

9

3 年内，颠覆式的 AI 应用的核心驱动力来自于底层模型的创新，两者无法解耦，模型的作用将大于产品设计的作用

10

当前生成式 AI 市场处于技术主导的早期阶段，存在千亿美元市值的平台性企业的机会

关于报告

周志峰
启明创投
合伙人

胡奇
启明创投
副总裁

周健工
未尽研究
创始人

李柯达
未尽研究
研究总监

启明创投

启明创投成立于 2006 年，先后在上海、北京、苏州、香港，西雅图、波士顿和旧金山湾区设立办公室。目前，启明创投旗下管理 11 只美元基金，7 只人民币基金，已募管理资产总额达到 95 亿美元。自成立至今，专注于投资科技及消费（Technology and Consumer, T&C）、医疗健康（Healthcare）等行业早期和成长期的优秀企业。截至目前，启明创投已投资超过 530 家高速成长的创新企业，其中有超过 200 家分别在美国纽交所、纳斯达克，香港交易所，上交所及深交所等交易所上市，及合并等退出，有 70 多家企业成为行业公认的独角兽和超级独角兽企业。AI 1.0 到 AI 2.0 发展的 10 余年中，启明创投一直是 AI 领域最活跃的创投机构之一。秉持预判趋势、提前布局的方法论，启明创投已经投资了二十余家在大模型、视觉、语音、自动驾驶、机器人等领域的领跑企业。

未尽研究

未尽研究是一家独立的机构，研究前沿科技和创新，包括人工智能、新能源、生命科技，以及技术与地缘相关的问题。除了日常的分析和报告，未尽研究在每年结束的时候，发布一份有助于“看到”来年技术趋势的报告《看 DAO XXXX》。

启明创投
QIMING VENTURE PARTNERS

×

未尽研究
WEIJIN RESEARCH