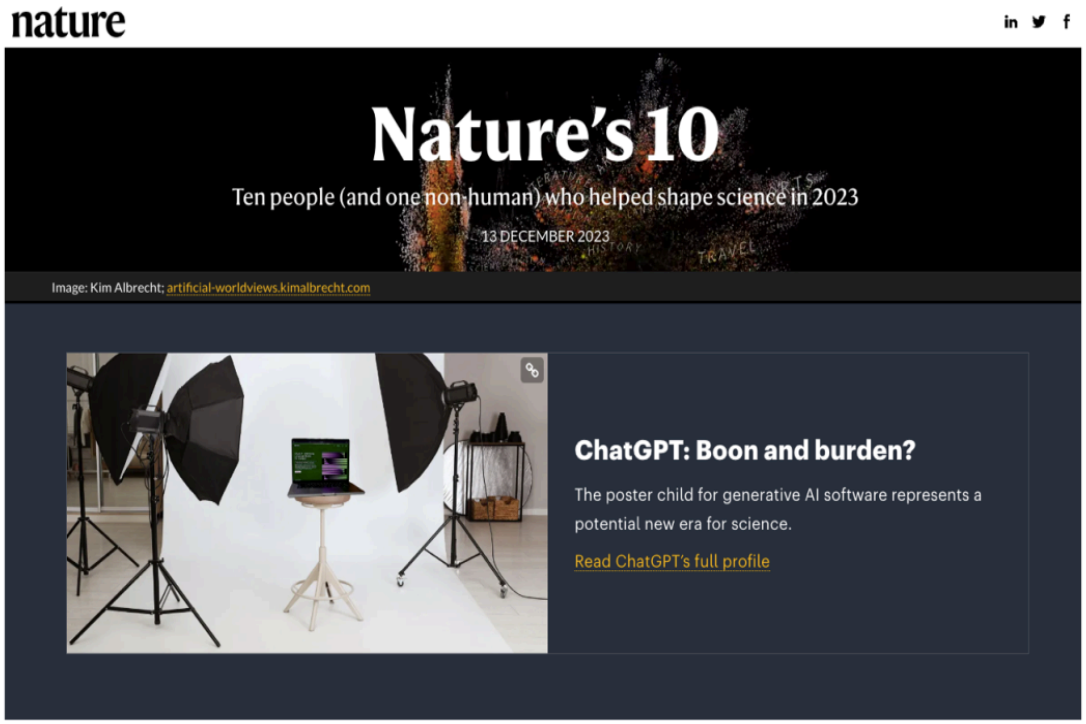
迈向2024，我们如何思考AI创业投资

2023年12月，《自然》杂志发布年度“科学十大影响人物”榜单，今年有史以来首次有“非人类”入选了——名单中包括了ChatGPT。《自然》指出：“尽管ChatGPT并非个体，也不完全符合评选标准，但我们决定破例列入，以承认生成式人工智能正在从根本上改变科学的发展轨迹。”



▲ 图片来源：Nature

在2023年的科技版图上，生成式AI无疑标志着一个重要的转折点。它的发展不仅

■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↓

引起了业界广泛的关注，也对全球经济、社会结构乃至我们对未来的预期产生了深远的影响。

这是每个普通人都可以参与的AI革新。从大型语言模型的持续发展，到AI技术在不同行业的广泛应用，再到开源与闭源策略之间的持续较量，AI的每一步发展都在描绘着未来趋势的轮廓。

面对滔滔浪潮，国家先后在《“十四五”国家信息化规划》《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》中出台了一系列支持AI发展的政策措施。中国的人工智能产业规模也快速壮大，涌现出一批具有国际竞争力的AI企业。

时逢年末，我们回顾2023年生成式AI的发展，从技术对人类的影响、行业格局和未来发展趋势、创业和投资机会等方面展开论述。这不仅是对AI领域过去一年的发展回顾，更是对AI发展走向的思考。

先分享核心结论：

- 在真正有价值的AI应用生态繁荣之前，押注大模型这样的核心技术源头及“卖铁锹”公司是有一定道理的。但目前正在蓬勃发展中的AI应用，同样是价值创造的源泉和我们要追求的星辰大海。

- 像OpenAI这类闭源大语言模型，会向接入其端口的APP应用收取流量费。应用公司为了降低流量费用的负担，一种方法是利用开源模型，自己训练出一个中小模型，另一种方法是优化商业模式，从而平衡流量费用。

- 随着AI技术的进步，工作方式也会发生变革。AI技术既可能重构人们的工作流，也可能重构语言模型本身的工作流。

- 如何用好AI这样极其智能的工具，对人类来说无疑是巨大的挑战。但是，我们也不要那么悲观，AI的能力是有边界的。

- 在AI技术领域，美国和中国的发展路径各有特色。美国的头部大语言模型阵营已基本确立，中国的大型语言模型呈现了百花齐放的态势。对于中国来说，更重要的是大力发展AI应用生态。

- AI Agent是个值得关注的创业方向。AI Agent是一种能够自主执行任务、独立决策、主动探索、自我迭代并能相互协作的智能软件。

- 虽然大语言模型领域已经实现了众多技术突破，但仍然有不少可以迭代、提升的板块，比如减少“幻觉”、增加上下文长度、实现多模态、具身智能、进行复杂推理以及自我迭代等等。

- AI应用领域创业的几个要点：要做出优质的原生新应用体验；更前瞻、发现非共识、有颠覆性；关注用户增长和商业化潜力；把握宏观趋势红利；跟大模型保持安全距离，有自己的业务纵深；最重要的还是团队。

- 创业公司要敢于在非共识的领域，做正确而非容易的事。

/ 01 /
2023年，AI领域有哪些新变化?

AI发展至今，从业界的角度来看，可以分为两个阶段：1.0阶段主要集中于分析和判断，而2.0阶段更侧重于生成。2.0阶段的代表模型是大型语言模型和图像生成模型，Transformer和Diffusion Model这两个算法模型推动着生成式AI的发展。

2023年的大部分时间，OpenAI这家初创公司的产品稳居大型语言模型高性能的榜首，特别是在3月OpenAI发布GPT-4语言模型之后，几乎是一骑绝尘。但Google在12月成功发布最新的大型语言模型Gemini，与GPT-4形成双雄割据的格局。

在AI领域，开源模型社区一直没有缺席。开源模型社区在Meta（原Facebook）

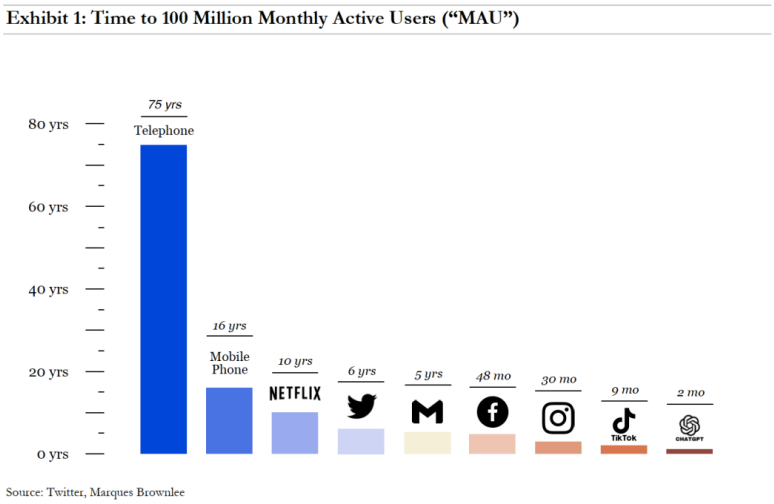
的开源大型语言模型LlaMa及LlaMa2的支持下，进行着密集的科研和工程迭代，比如：试图用更小的模型，释放出与大模型类似的能力；支持更长的上下文；采用效率更高的算法和框架来训练模型等等。

多模态（图像、视频等多媒体形式）已经成为AI领域研究的热点。多模态分为输入和输出两个方面。输入是指让语言模型能够理解图像和视频中蕴含的信息，输出是指除文本之外，生成其他媒体形式，比如文生图。考虑到人类生成和获取数据的能力是有限的，未必可以长期支撑人工智能的训练，未来可能需要用AI自己合成的数据，来训练语言模型。

在AI基础设施领域，英伟达凭借其GPU的巨大市场需求，成为行业的领导者，跻身1万亿美元市值俱乐部。但是它也将面临来自老对手AMD、英特尔等芯片制造商，以及Google、微软、OpenAI等大厂和语言模型新贵的激烈竞争。

除大模型外，业界对各种类型的AI应用有强烈的需求。生成式AI在图像、视频、编程、语音以及智能协作应用等多个领域取得了显著进展。

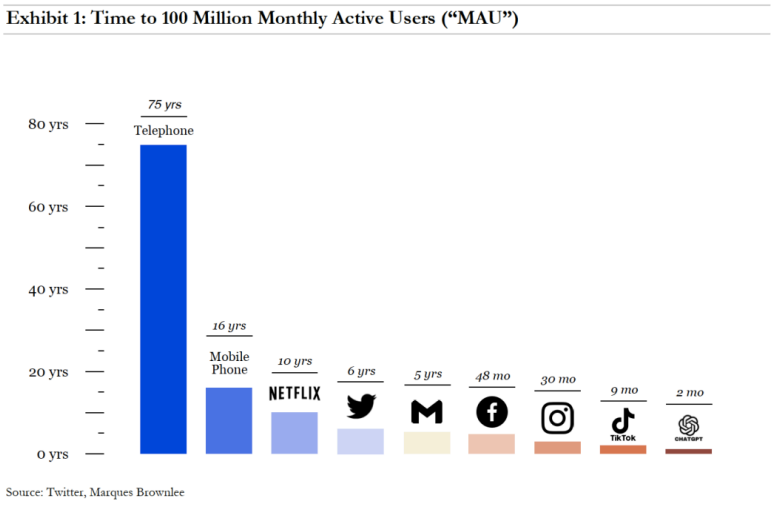
全球用户对生成式AI展现出极大的热情。ChatGPT在短短2个月内，达到1亿月活跃用户数。相比智能手机时代的超级应用们，在大量推广预算之下，TikTok用了9个月，Instagram用了2.5年，WhatsApp用了3.5年，YouTube和Facebook用了4年。



▲ 不同类型的科技应用达到1亿月活跃用户的时间。

图片来源：7 Global Capital

创投机构也在投入重金，支持AI领域的进展。根据美国投资机构COATUE的统计，截止至2023年11月，风险资本投资机构向AI领域投入了近300亿美元，其中约60%投向OpenAI等大型语言模型新贵，约20%投向支持和交付这些模型的基础设施（AI云服务、半导体、模型运营工具等），约17%投向了AI应用公司。



▲ 图片来源：COATUE

在真正有价值的AI应用生态繁荣之前，这种押注核心技术源头及“卖铁锹”公司的投资逻辑是有一定道理的。但目前正在蓬勃发展中的AI应用，同样是价值创造的源泉和我们要追求的星辰大海。

多模态生成领域出现多项技术突破

2022年，在Stable Diffusion开源之后，我们见证了大量“文生图”（由文字生成图像）产品面世。这一年可以被视为图像生成问题的解决之年。

紧接着在2023年，用AI识别声音、生产音频的技术也取得了显著进展。如今，AI的语音识别和合成技术已经非常成熟，合成声音与人类声音很难被区分。

随着技术的持续发展，视频的生成和处理将是下一个阶段AI发展的重点。目前在“文生视频”（由文字生成视频）领域已经出现了多项技术突破，AI在视频内容生成方面展现出了潜力和可能性。借助AI视频新秀Runway Gen-2、Pika以及斯坦福大学的W.A.L.T等模型及应用，用户只需输入对图像的描述，就能得到一段视

频片段。

英伟达知名工程师Jim Fan认为，2024年，AI大概率要在视频领域取得进展。

▲ 图片来源：X.com

如果我们换一种维度，来思考不同形态的媒体格式，那么一张二维的图像，如果增加一个时间的维度，就变成了视频。如果增加一个空间的维度，就变成了3D。如果将3D模型经过渲染，我们就能得到更加可精确控制的视频。可能未来AI也能逐渐攻克3D模型，但还需要更长的时间。

“压缩即智能”

2023年，OpenAI的首席科学家伊利亚·苏茨克维（Ilya Sutskever）在某次外部分享中，提出一种“压缩即智能”的观点，即语言模型对文本的压缩比越高，就说明它的智能程度越高。

压缩即智能，可能不一定严谨，但却提供了符合人类直觉的解释：最极致的压缩算法，为把数据压缩到极致，势必需要在充分理解的基础上，抽象出更高层次的意义。

以Llama2-70B这个Meta研发的语言模型为例，它是Llama2模型700亿参数版

本，是目前最大的开源语言模型之一。

Llama2-70B使用大约10T（10万亿）字节的文本作为训练数据，训练出来的模型是一个140GB大小的文件，压缩比大约是70倍（10T/140G）。

在日常的工作中，我们通常把大的文本文件压缩成Zip文件，其压缩比大约是2倍左右。对比之下，可以想见Llama2的压缩力度。当然Zip文件是无损压缩，语言模型是有损压缩，不是一个标准。

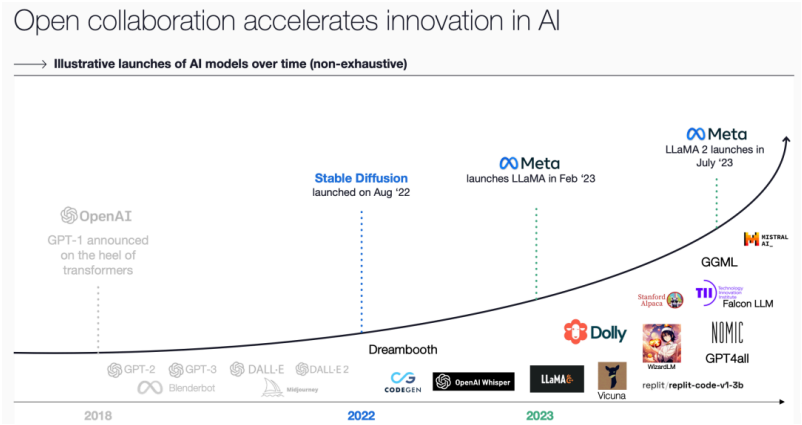
▲ OpenAI副总裁Andrej Karpathy分享截图。

图片来源：Web3天空之城

神奇的地方在于，一个140GB的文件就可以把人类的知识和智能给保存下来。大部分的笔记本电脑都可以装得下140GB的文件。当笔记本电脑的算力和显存足够大，只要再加上一个五百行的C代码程序，就可以运行大语言模型。

/ 02 /
开源生态和大语言模型的流量税

开放研究和开源生态是推动AI发展的重要力量



▲ 开源生态促进了AI技术创新。图片来源：Coatue.com

开放研究是AI技术发展的基础。全球最顶尖的科学家和工程师在Arxiv等网站发表大量论文，分享他们的技术实践。无论是早期的AlexNet卷积神经网络模型，还是奠定算法基础的Google的Transformer，抑或是OpenAI、Meta等公司发表的模型实践论文，都是科研和技术上的重大突破，引领着AI技术发展。

开源社区的发展和迭代尤其值得关注。在开源大语言模型的支持下，科研人员和工程师可以自由地探索各种新的算法和训练方法。即使是闭源的大语言模型，也能向开源社区学习和借鉴。

可以说，开源社区实现了某种程度上的科技平权，让全球的人们都能共享AI领域的最新技术成果。

大型语言模型的“流量税”

回归商业本质，大型语言模型的训练成本是非常昂贵的。以GPT为例，据远川研究所统计，训练GPT-3用了超过1000万美金，训练GPT-4用了1亿多美金，下一代模型的训练成本可能要达到10亿美金。此外，运行这些模型并对外提供服务的时候，其算力和能源的消耗也是很昂贵的。

大型语言模型的商业模式是MaaS（Model As a Service），它输出智能的计费方法是按照输入输出的流量（或称token，词元）来收费。由于大语言模型昂贵的训练和运行成本，它收取的流量费大概率会“水涨船高”。

GPT-4

With broad general knowledge and domain expertise, GPT-4 can follow complex instructions in natural language and solve difficult problems with accuracy.

Learn about GPT-4

Model	Input	Output
gpt-4	$0.03 / 1K tokens	$0.06 / 1K tokens
gpt-4-32k	$0.06 / 1K tokens	$0.12 / 1K tokens

GPT-3.5 Turbo

GPT-3.5 Turbo models are capable and cost-effective.

`gpt-3.5-turbo-1106` is the flagship model of this family, supports a 16K context window and is optimized for dialog.

`gpt-3.5-turbo-instruct` is an Instruct model and only supports a 4K context window.

Learn about GPT-3.5 Turbo ↗

Model	Input	Output
gpt-3.5-turbo-1106	$0.0010 / 1K tokens	$0.0020 / 1K tokens
gpt-3.5-turbo-instruct	$0.0015 / 1K tokens	$0.0020 / 1K tokens

▲ 图片来源：openai.com

以OpenAI为例，上图是其官网所展示的部分模型的流量计费方案。有人做过粗略估计，按照AI应用调用GPT-3.5 Turbo流量的中位数水平，只要有一个用户每天使用该应用（DAU），用户背后的APP公司约需要向OpenAI支付0.2元人民币左右的流量费用。以此类推，如果是千万级别日活的APP应用接入了GPT的端口，那么每天要向OpenAI支付200万人民币的流量费。

2023 年 9 月中国大模型 API 定价

厂商	模型名称/版本	价格(元/千 token)	免费额度
阿里系	通义千问-turbo	0.012	30 万 token
	通义千问-plus	0.14	
腾讯系	混元大模型-标准	0.01	10 万 token
	混元大模型-高级	0.10	
百度系	文心一言(ernie-bot)-turbo	0.008	未提及
	文心一言(ernie-bot)-正常	0.12	
	llama2 系列-70B	0.044	
	chatglm6B	0.006	
讯飞星火	1.5 版本星火模型	0.018	300 万 token
	2.0 版本星火模型	0.036	套餐 1800 元起
清华智谱	ChatGLM-Pro	0.01	18 元额度
	CHATGLM-LITE	0.002	
参照对比			
OpenAI	GPT3.5	0.014	2 美金/100 万 token

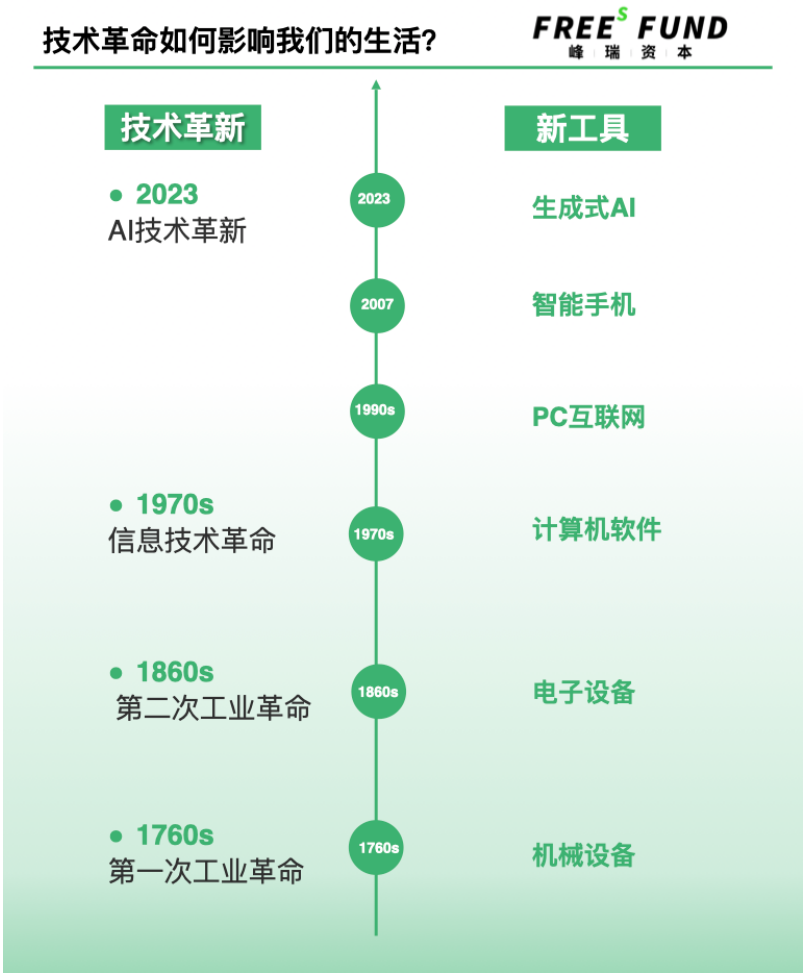
▲ 图片来源：微信公众号@AI赋能实验室

国内大模型的流量费报价如上图所示，跟OpenAI的价格基本相当。部分中小模型会便宜一些，但性能有差距。

流量费用会影响到AI应用如何设计商业模式。为了降低流量费用负担，有些创业公司会考虑利用开源生态的能力，自己做一个中小模型，去承接大部分的用户需求。如果碰到超出中小模型能力范围的用户需求，再调用大型语言模型。

这类中小模型，可能会被直接部署在离用户最近的终端侧，成为“端侧模型”。端侧模型很考验硬件的集成水平，未来我们的电脑和手机上，可能会更广泛地集成GPU之类的硬件芯片，具备在终端侧运行小模型的能力。Google、微软已经推出可以在终端侧运行的小模型。Nano是Google发布的Gemini大模型中最小的一款，专门为在移动设备上运行而设计的，无需联网，可以直接在设备上本地和离线运行。

/ 03 /
AI如何影响了人类社会?



每一次技术革命，都会带来新的效率工具

人类历史上有几次大的技术革命。1760年左右兴起的第一次工业革命，产生了机械设备；1860年之后的第二次工业革命，产生了电子设备；1970年之后，我们又经历了计算机软件、PC互联网和智能手机这三次技术革新，有人统称其为第三次工业革命即信息革命。

2023年开始的生成式AI革命，或许可以被称为第四次工业革命，我们创造了新的智能。生成式AI是人类认知和改造世界的新工具，已经成为新的抽象工具层。

根据历史经验，每一次技术革命都会极大提升人类生产效率。第一次和第二次工业革命后，自然世界形成了两个抽象工具层，即机械和电子设备层。20世纪70年代，以计算机为代表的信息技术革命引入了新的抽象层——软件。通过软件，人们开始以更高效的方式理解、改造世界，并与之互动。随后，PC互联网和智能手机崛起，进一步推动了软件技术的发展。

AI如何影响人们的工作?

除了关注AI带来的效率提升，我们还要关注机器如何替代了人类的工作。据统计，英国第一次工业革命之前，农业人口占比约为75%，而工业革命之后降至16%。美国信息革命之后，工业人口从38%降至8.5%，当时那些工业人口大多转变成白领人口。而这次AI的智能革命，首当其冲的正是白领人群。

随着AI技术的进步，商业社会中的组织形式和协作方式可能会发生一系列变化。

首先是，公司可能往小型化发展。商业外包可能会变得非常普遍。比如，公司可以把研发、营销等板块外包出去。

其次是工作流的重构，也就是标准操作程序（SOP）可能会发生变化。每个人的能力和精力有所不同，因此，工作流能够让人们提高效率，各司其职。研究人员正在探索在AI可能替代某些职能的情况下，人们的工作流该如何调整。当前的语

言模型也存在可以提升效率、增强能力的地方，语言模型可能也需要借助工作流的编排，进行协作。

除了技术技能之外，提高其他能力也变得至关重要。例如，提升鉴赏力和品味，才能让AI辅助你生成更好的方案或者作品。再比如，增强批判性思维，能帮助你更好地判断、鉴别AI生成的内容。

我们要更积极地利用AI，把它当作工作和生活中的辅助工具，或者说副驾驶，充分利用其潜力和优势。

AI的能力是有边界的

在AI发展迅猛的当下，不少人提出了AI威胁论，担心AI对人类造成的负面影响。确实，人类目前发明出了看起来比自己还聪明的工具。如何控制好AI这样的“硅基生物”，对人类来说无疑是巨大的挑战。科学家们正在尝试解决这个问题，OpenAI也曾发表探讨类似问题的论文。

但是，我们也不要那么悲观，至少目前人类社会的数字化程度，可以限制AI的能力边界。

如今的大语言模型主要是用大量文本数据训练出来的。文本的数字化程度很高，又经过人类的抽象，信息密度大，所以AI训练的效果很好。

但是离开了文本空间，AI的智能会受到诸多限制，因为它没有经过相应的数据训练。所以我们暂时不用太担心，AI并没有那么厉害和全面。我们有充足的时间去熟悉和适应它，找到跟硅基生物友好相处的方法。

/ 04 /
展望2024，
大语言模型与AI应用会如何发展？

头部大语言模型阵营

在全球范围内，大型语言模型呈现出显著的区域化发展特征。比如，美国和中国的发展路径各有特色。美国的头部大语言模型阵营已基本确立，主要集中在几家大型科技公司，或者它们跟几家头部模型创业公司的联合体。可以说，美国的AI领域已进入高成本的军备竞赛阶段，新的参与者比较难入局。

而中国的大型语言模型则呈现了百花齐放的态势，目前有百余个项目声称正在开发大型模型。中国可能更依赖于开源生态，二次开发出新的语言模型。

目前，除美国以外的其他国家，都还没有开发出与GPT-4相当的大型语言模型。在大模型技术领域，中国和美国仍然存在差距。

但全球在AI领域的较量还未到终局。对于中国来说，最重要的是大力发展AI应用生态。在互联网和数字经济时代，中国就是应用领域的优秀生，也向海外输出了相关的应用实践。在紧跟大模型最新技术的前提下，等应用生态繁荣起来之后，我们再反向去做技术突破，可能是一种解决思路。

大语言模型会如何发展?

虽然大语言模型领域已经实现了众多技术突破，但仍然有不少可以迭代、提升的板块，比如减少“幻觉”、增加上下文长度、实现多模态、具身智能、进行复杂推理以及自我迭代。

首先，我们来讨论“幻觉”现象。幻觉可以理解为一种错误的输出，Meta将其定义为“自信的假话”。幻觉的产生最常见的原因是语言模型采集的知识或数据的密度不够。不过，幻觉也可以被视为创造力的体现，就像诗人在酒后能写出美妙的诗篇，AI的幻觉可能也会给我们带来奇妙的内容。

减少幻觉的方法有很多种，比如使用更高质量的语料库进行训练；通过微调和强化学习来提高模型的准确性和适应性；在模型的提示词中加入更多背景信息，让模型基于这些信息更准确地理解和回应问题。

第二，增加上下文长度。上下文长度相当于语言模型的脑容量，现在通常是

32K，最高的是128K，也就是不到10万字或者英文单词。如果想让语言模型理解复杂的语言文本、处理复杂的任务，这个长度还远远不够。下一代的模型大概率会努力扩大上下文长度，以提高处理复杂任务的能力。

第三是多模态。人类主要依靠视觉来获取信息，而当前语言模型主要依靠文本数据来做训练。视觉数据能够帮助语言模型更好地认知物理世界。在2023年，视觉数据被规模化地加入到模型的训练过程中。比如，GPT-4引入了多模态数据，Google的Gemini模型据说也使用了大量的图像和视频数据。从Gemini演示视频的表现来看，它的多模态交互似乎有明显提升，但复杂推理等智力的提升还没看出来。

第四是具身智能，是指一种基于物理身体进行感知和行动的智能系统，能够从环境中获取信息、理解问题、做出决策并行动。这个概念并没有那么复杂，地球上所有的生物，都可以说是具身智能。比如人形机器人，也被认为是具身智能的一种形式。具身智能相当于给AI延展出了能活动的“手脚”。

第五是复杂推理。通常，GPT会一次性地给出回答，没有太明显地多步推理或回退迭代。而人类在思考复杂问题的时候，会在纸上列出一些步骤，反复推演和计算。研究人员想了一些方法，比如借助思维树等思考模型，试图让GPT学会复杂的多步骤推理。

最后是自我迭代。现在的语言模型主要还是依靠人给它设计算法，提供算力，给它喂数据。畅想未来，语言模型能够实现自我迭代吗？这可能要依赖于新的模型训练和微调方法，例如强化学习等。据说OpenAI正在尝试一种代号为“Q*”的训练方法，研究如何让AI自我迭代，但具体进展尚未知晓。

大模型还处在高速发展期，还有很大的提升空间。除了以上列举的几点之外，还有很多待解决和提升之处，比如可解释性、提升安全性、输出的内容更符合人类的价值观等等。

未来的应用软件——AI Agent

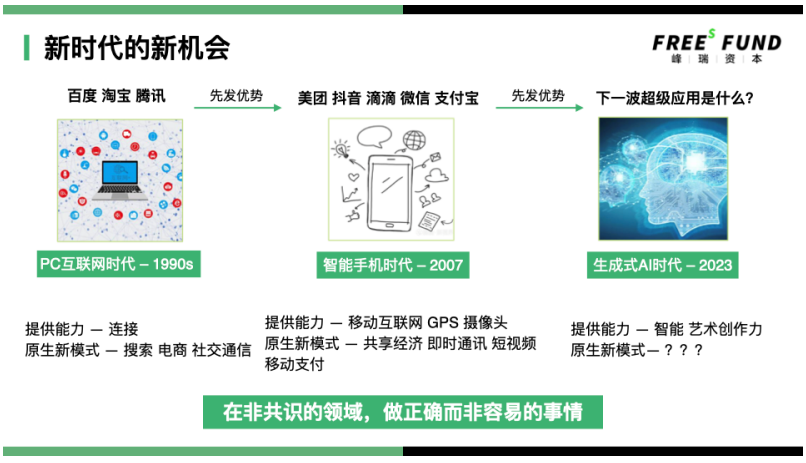
2023年9月，红杉美欧（Sequoia Capital）官网发布了《Generative AI's Act Two》的文章，提到生成式AI已进入第二个阶段。第一个阶段主要集中在语言模型及周边简单应用的开发，第二个阶段的焦点则转向研发真正解决客户需求的智能新应用。

未来的应用软件，可能会逐渐转向AI Agent——一种能够自主执行任务、独立决策、主动探索、自我迭代并能相互协作的智能软件。现有的传统软件可能需要进行相应的调整和改进。和传统的1.0版本软件相比，AI Agent能够提供更接近真实的、高质量的一对一服务体验。

但发展AI Agent的难点在于，语言模型目前还太不成熟和稳定。如果要做出好的应用体验，需要在语言模型基础上，加上一些小模型、一些规则算法，甚至在某些关键环节加入人工服务，从而在垂类的场景或者具体行业中输出稳定的体验。

多Agent协作已经成为热门的研究方向。在标准操作程序的基础上，相互协作的多个AI Agent，能够产生比单独调用语言模型更优的效果。这里有个比较符合直觉的解释，每个Agent可能各有优缺点和专攻方向，跟人类的分工是一样的，大家组合到一起，通过新的标准操作程序（SOP）各司其职、互相启发和监督协作。

/ 05 /
创业和投资机会



在非共识的领域，做正确而非容易的事

在一个新的时代，作为创业公司，需要认真思考，基于这次技术革新，有哪些原生新模式的创业机会。同时还要考虑，哪些是新进入者的机会，哪些是现有行业领先者的机会。

我们可以回看PC互联网和智能手机两次技术变革，如何产生出了新的机会。

PC互联网时代，提供的主要能力是连接，即全球的PC、服务器和一些其他设备实现了联网。PC时代产出的原生新模式包括：搜索、电商和社交通信等，诞生了BAT等各行各业的领先企业。

智能手机时代，提供的主要能力是大部分人都拥有一台手机，具备移动互联、GPS、摄像头等功能。这个基础条件让共享经济、即时通讯、短视频分享、移动金融支付等新模式成为可能。前一时代的行业领先企业是有很强先发优势的，抢占了不少新模式的机会，例如：腾讯和阿里分别做出了微信和支付宝。但是我们也看到美团、抖音和滴滴等一些新势力，获得了巨大成功。它们为什么可以做到?

我认为其成功的关键词是，在非共识的领域，做正确而非容易的事情。

以美团和抖音为例。美团选择的原生新模式叫“餐饮外卖”，属于“共享经济”中的“O2O（线上到线下）”部分，左边大量的餐饮店面，右边是众多各式各样的消费者，中间是成千上万的骑手，是“重模式”，但早期互联网大厂更喜欢和擅长做“轻模式”，切入餐饮行业是“非共识”。外卖的履约服务链条太长、难以数字化，很难进行精细化运营。但最后美团把它做成了，这些难的事情成为其最大的核心优势和竞争壁垒。

再看抖音，它选择的原生新模式叫“短视频分享”，属于当时流行的“创作者经济”的一部分。抖音最大的“反共识”是：它把视频创作者经济跟万亿体量的电商GMV之间的桥梁打通了，形成规模化、有效率的转化。

在电商直播崛起之前，有两类直播，一种叫游戏直播，另一种叫网红直播，变现主要靠观众打赏。这类变现模式的经济体量非常小，容纳不了那么多优秀的创作者。但抖音通过推荐算法、发展创作者生态和商家生态、建立抖音小店闭环、优化内容电商转化等各种努力，把内容往电商转化这个巨大的商业闭环给做通了。做通之后，抖音就可以邀请全国最多最优秀的创作者来抖音平台创作内容，并报之以巨大的电商销售收入作为奖赏。

所以，抖音的海外版TikTok出海后，很多当地的短视频和直播平台都打不过它。因为Tiktok并不仅仅是一个左边创作者右边消费用户的视频内容平台，它更是一个新型的创作者经济和海量电商GMV转化的结合体，是新物种，具备复合型竞争优势。

总结来说，创业公司要敢于选择和进入非共识的领域，在艰难的环境下，努力把事做成。

创业方向和要点

从创业的方向来说，大模型领域巨头林立，大概率不会是创业者的首选方向。而在大模型和应用之间有个“中间层”，大部分是基础设施、应用框架、模型服务等，这个部分容易受到模型和应用的双向挤压，部分领域巨头林立，创业空间不大。

综上所述，我们倾向于认为，结合目前的技术和商业环境，我们应该大力发展AI应用生态。

上图是我们投资的生成式AI相关的创业公司，包括：为语言模型设计的新型DevOps平台、社交游戏平台、智能陪伴服务、AI辅助RNA药物开发、门店自动化营销、服务全球的智能商业视频SaaS、新型线上心理咨询平台和中美工程师远程雇佣工作平台等等。

我们总结了AI应用领域创业的几个要点：

第一，要做出优质的原生新应用。要抓住AI智能时代提供的新能力，即智能供给和艺术创作力供给，做出优质独特的原生新应用体验，这个难度其实不小。我们在上文曾提到，语言模型的智能等还不够成熟稳定，存在明显的能力边界。创业公司可能需要选择相对垂直细分的场景，采用各种技术和运营手段，做出良好体验。

第二，非共识、更前瞻、有颠覆性。非共识指的是在赛道选择上不要随大流，敢于进入艰难的领域，“做正确而非容易”的事情。更前瞻是指选择有挑战的业务和技术路线。

例如，采用当下还在发展、更先进的技术架构，例如：创业者要优先做Agent而不是CoPilot，CoPilot们更像是行业领先者的机会（想想微软和Github）。再比如，创业团队可以考虑提前按照下一代语言模型的能力（如GPT-5），去构思和设计应用。

颠覆性是指最好对所切入的行业产生颠覆效果，例如：颠覆性的产品体验、颠覆原有的商业模式等等。这类颠覆性的好处是有可能跑在行业领先者前面。比如峰瑞投资的Babel（巴别科技），抓住尚未成熟的“Serverless”、大语言模型等技术发展趋势，致力于重构软件开发的生产工具和生产要素，让AI来做编程、调试、部署和运维等工作。

第三，关注用户增长和商业化潜力。用户增长潜力很重要，大家容易理解，即便你从一个细分市场切入，未来也可以做成大的规模。

我们早期为什么要关注商业化呢?

这要回到我们在上文提到的大模型的流量税。如果你选择接入大模型，从创业的第一天开始，你就要给大模型支付流量税。

面向个人用户的应用，当前规模商业化的途径通常有三种：前向收费（如游戏、增值服务等）、广告以及电商。只有极少数应用有可能把电商做起来（例如淘宝、抖音等）。新应用直接向用户收费很难，大多数创业者会有畏惧心理，会考虑选择比较间接的方式，希望做大用户规模后在应用里做广告来商业化。

从智能手机时代的情况来看，除了电商应用之外，中国最头部的几个泛资讯类APP估计每天能够在单个活跃用户上赚到的广告收入大约在0.1元到0.3元之间，这已经是广告商业化的极致水平。而一般规模的APP，可能还远远达不到0.1元。

我们在前面讲过语言模型的“流量税”，每个用户每天的成本约0.2元，广告收入通常很难覆盖得住这样的成本。用户规模越大，亏损反而越严重，除非你通过前面提到的端侧模型等手段把“流量税”降下来。

因此，AI应用在商业模式设计上可能需要优先考虑前向收费。当然，在新的AI智能时代，说不定我们的创业者可以找到除上述三种规模商业化之外的其他商业化途径，让我们拭目以待。

第四，把握宏观趋势红利。要预判和抓住中国的宏观趋势红利，比如商品出海、视频电商、工程师红利等等。我们要努力抓住属于时代的β。

峰瑞投资的创业公司特看科技，也在抓住中国商品出海、新型视频电商等新趋势的机会，立志通过产品创新技术打造世界级的商业视频SaaS平台，赋能海外视频创业者和商家。

第五，跟大模型保持安全距离，有自己的业务纵深。安全距离大家应该有所耳闻，知名的海外反例有不少，比如一些生成文案的商业应用公司，虽然实现了“昙花一现”式的快速增长，终究难以逃脱大模型和其他创业公司的双向冲击。此外，创业项目的业务纵深也很重要，这个业务纵深是指大模型够不着的地方，特别是一些难以数字化或者数字化不充分的场景。

当然，最重要的还是团队，技术要好，团队成员也要懂行业和场景，所谓“技术为先，场景为重”。