# **大模型时代AI领域机会的初步思考**

原创 Eagle 当下瞬间进入历史 *2024年11月09日 14:17* *北京*

今年是以大模型为特征的人工智能爆火出圈的第二年，和去年相比，街谈巷议和让人眩晕的冲击震撼似乎少了一些，但大家的共识是，行业前进的步伐一点也没有放缓。大家在经历了一轮炫目冲击波后迅速学习、理解并接受了这一新生事物，转而深入思考这一巨大变化带来的深远和现实影响，或投身其中，或谋篇布局。

而在今年10月份的诺奖周，诺奖委员会先后将物理奖和化学奖两个最重要的诺贝尔科学奖颁给了人工智能科学家，是前所未有的科学史事件，也是人工智能发展史的标志性事件，尤其考虑到颁发物理奖的理由和物理学的突破基本无关。

诺奖的认可通常具有滞后性，尤其在人工智能这个日新月异的领域。今年整个行业最重要的突破也许可以用Sora代表的长视频生成，Gemini为代表的超长上下文和o1代表的增加推理时计算来大幅提升模型推理能力为代表。

在这个快速变化、影响深远且未来难料的人工智能时代回答现实的机会在哪是非常挑战的问题，也是一个行业内几乎所有人都要思考并回答的问题。红杉资本的David Cahn先后用两篇题为“AI的 2000亿美元问题”和“AI的6000亿美元问题”这种刺激性标题的文章提出了AI行业投资与回报之间巨大差距的问题，也触发了大家的讨论和深入思考。

首先被关注的机会是模型，原因是这轮AI突破的根本推动力是以GPT4为代表的模型能力的大幅提升。在OpenAI通过ChatGPT引发海啸关注后，主要行业巨头都宣布下场开发大模型，大量大模型创业公司成立，得到天文数字的风险投资支持，很快形成行业巨头和多家独角兽激烈角逐竞争的局面，以Meta为代表的大模型开源力量也让竞争和发展局面复杂而难以预测。训练前沿模型需要的资金从1亿美金级别很快增加到10美金级别，也让这个赛道变成让参与者和投资人都压力重重的游戏。一个有点恐怖的共识是，即便像OpenAI这样的领先公司，也很难在模型领域构建真正的护城河。另外如OpenAI的产品负责人所说，即使在OpenAI内部也很难预测两个月后模型的能力会提升到什么水平，前沿模型水平在难以预测的快速演进，这个领域完全是一场昂贵的“军备竞赛”。今年微软收购Inflection和谷歌收购Character都是模型赛道的变化标志。

目前确定性最高的机会出现在加速计算芯片领域，英伟达股价暴涨最近成为市值最高的公司就是市场给出的一个直接的反应，它的新款GPU芯片经常供不应求。随着模型参数规模、训练数据和训练时间的增加，模型的能力随之增长，规模扩展定律（Scaling law）成为这一轮AI革命背后的法则，强化学习大师Rich Sutton的文章“AI研究中苦涩的经验”（The Bitter Lesson）成为AI研究人员的指导，o1带来的推理时间计算扩展定律又将对算力的需求增加一个新的维度，在可以预测的未来，算力芯片的需求将继续不断提升。

困难的问题是目前英伟达几乎垄断了AI加速算力的市场，这个领域其他希望竞争的厂家有机会吗？或者对于英伟达来说，护城河足够宽吗？谷歌早已在加速计算芯片领域用新一代架构TPU布局，Meta和其它几个巨头也开始行动。中美也都出现几家明星加速芯片创业公司并获得投资者青睐，但围绕芯片构建生态就不是一朝之功了，几家传统芯片巨头转型的困难也在于此。

应用领域的机会最具多样性，Open AI的ChatGPT一经发布就成为史上最爆款应用，无论使用户数的增长还是收入都是现象级的。和模型及芯片相比应用需要的投资少收益快，成为创新创业最活跃的领域。聊天，编程，写作，图像视频生成，文档处理以及大量垂直类的行业应用呈现百花齐放的局面，大家都在期待AI原生的杀手级应用出现。

应用领域机会最大的挑战在于如何选择模型策略，自主开发还是使用模型服务成为两难选择，大模型公司随时可能进入有潜力的应用领域，核心大模型的生态具有高度动态和复杂性。目前领先的几个应用独角兽一方面依靠产品的客户价值构建护城河，另一方面也会利用模型中立的策略保持灵活性，并不断探索基于开源通用模型构建自主领域模型的能力。大多数应用玩家没有构建模型的能力，这会导致产品容易同质化，如何围绕用户的痛点构建产品成为关键；目前模型的能力还在高速迭代演进，也会让产品的设计非常困难，需要不断回顾产品设计的逻辑，基于当前模型能力和未来能力的预测规划产品。

除了模型、芯片和应用三个大的领域外，围绕三者周围的软硬件系统和工具也有很多新的问题需要解决，出现了大量的创新技术和公司。

支持大模型原生应用开发框架和工具是一个全新的赛道，LangChain和LlamaIndex的快速流行就是一个代表，正如ML系统领域的高手贾扬清博士的评论所说，在这个模型快速演进的阶段，目前的应用框架可能是“过早的抽象”，反映了这个赛道的挑战。

模型开发最大的挑战是如何支持超大模型的训练，目前前沿的训练平台已经从万卡级别扩展到十万卡级别，这是一个新型的分布式系统构建的问题，和传统的高性能计算有相似性也有本质的区别，有多个分布式训练框架脱颖而出，新的优化方案也不断出现。除了训练平台，推理领域的问题也会随着模型应用范围不断扩展而凸显，尤其o1代表的范式迁移也会增加推理领域的需求。大模型引入的一个新的问题是企业私有数据的处理问题，有两个竞争性的思路，将企业的数据加入模型参数或者利用模型能力处理企业私有数据，两者具有不同的挑战需要解决。模型整个生命周期不同阶段有不同的管理问题，也是可以创新的机会。

围绕加速芯片算力的基础设施领域也有大量的挑战问题成为创新的机会，算力集群的组成需要极致的网络连接，算力资源的昂贵和超大系统组件的高故障要求训练状态不断保存，这一要求直接推动了新型高性能存储的创新，基础设施资源的高效利用和运营对系统软件的需求也是前所未有，这些都是随大模型演进出现的新型机会。

所有这些机会目前以各种形态呈现，未来是否会演变成水平化的平台机会或者垂直性整合的平台机会，是现有玩家延展为主的机会还是创新玩家崛起的机会，都是需要思考并有待时间回答的。

​