从通用大模型到金融大模型

“它不是一种新武器，而是一个新世界的开始。”今年8月，电影《奥本海默》上线。围绕原子弹的争议，影片曾出现这样一句话。

一个月后的外滩大会上，这句话套用给了大模型。蚂蚁集团董事长兼CEO井贤栋判断，大模型的出现，“不是一个新技术，而是一个新世界”。

以ChatGPT发布为起点，以大模型席卷全球为标志，大模型重构了一个流量红利触顶后，以智能化主导经济增长的新世界，一个可能改变社会关系以及生产关系的新世界。

科技的力量迅速传导，压力给到国产大模型这边。上半年“卷”模型，下半年“卷”应用，国产大模型经历了带着些焦虑的狂热，也经历了些透过现象看本质般的冷静。

规模、参数的近身肉搏之后，落地逐渐成为“百模大战”的共识。得益于数据密集型的行业优势，辅以强劲的数字化基础及行业本身对新技术的天然追逐，金融在一众场景中脱颖而出。

也是因为这些优势，金融行业得以更快过渡到“冷静期”，将重点调整至价值——应用的价值。如何调用大模型？用大模型做什么？又要实现怎样的效果？遵循着这些问题的指引，金融大模型更容易找到落地的答案。

“金融行业将是大模型技术落地的最佳领域。”度小满CEO朱光曾做出了这样一个判断。只不过，在应用的“星辰大海”面前，当下的金融大模型仍然面临着复杂的多维挑战。

第一波浪潮

2012年，深度学习领域“宗师”Geoffrey Hinton带着他的两个学生一举拿下了当年ImageNet比赛的冠军。那场比赛中，三人密切配合地创建了一个新的神经网络用以进行图像识别，最终的准确率达到了惊人的84%。

在后来的介绍中，Hinton提到，相比谷歌1.6万颗CPU的累积，夺冠的AlexNet算法只用了4颗GPU。

当年的尝试，开启了用GPU训练AI模型的序幕，亦开启了OpenAI万卡算力的“暴力美学”。

催生ChatGPT的历史轨迹——Hinton的其中一个学生叫Ilya Sutskever，也就是后来OpenAI的联合创始人和首席科学家、灵魂技术人物。

在近期全球科技圈瞩目的OpenAI“逼宫”大戏中，秉持着对AI安全的恐惧和对技术发展的信念，Ilya亦承担起了除CEO阿尔特曼外最重要的一个角色。

这十年里发生了什么？

Google收购DeepMind，后者于2016年推出了震惊全球的AlphaGo。同年，英伟达把全球第一台AI超算送给了成立不足一年的OpenAI；三年后，微软与OpenAI开启合作，向OpenAI提供10亿美元投资，并与其建立独家云计算合作关系。

故事的另一条线索在大洋这边。2013年初，百度成立深度学习研究院。2016年百度创始人李彦宏宣布，百度将彻底转型为一家人工智能公司。

两条线索在2023年交汇。2023年3月，百度率先推出“文心一言”，成为国内首个生成式AI产品。自此，也揭开了国产大模型的诸神之战。

今年5月，中国科学技术信息研究所所长、科技部新一代人工智能发展研究中心主任赵志耕提到，根据不完全统计，当时中国已发布79个大模型。14个省市地区都有大模型研究，但北京和广东非常突出，分别有38个大模型和20个大模型。

市场有最新消息称，截至10月国内已经发布了238个大模型，如果按此数据推算，相当于不到半年时间就翻了3倍。

过去一年，国产大模型的发展可以看见一条清晰的界限。在11月11日的金融街论坛年会“金融科技创新与合规安全”平行论坛上，度小满CTO许冬亮总结称，底层的模型发展趋势上，早期的ChatGPT和文心一言为代表的通用大模型是第一波浪潮。

而第二波可以称之为“+AIGC”。许冬亮解释称，即在现有企业产品服务的基础上，将生成式人工智能技术应用其中，以提供更好的产品服务。

2023年7月，华为盘古大模型3.0发布，“行业”成了关键词。百度千帆大模型、腾讯云行业大模型、讯飞星火大模型、360智脑大模型等面向B端市场的落地行业大模型，在短时间内异军突起。大模型竞赛从通用大模型转向行业大模型的趋势越发明显。

对于大模型在不同产业中产生的效果，企业增长咨询公司弗若斯特沙利文发布的《2023年中国AI技术变革企业服务白皮书》曾做了一个简单的归类。比如在服务型产业中，AI能够实现5.8%的显著成本降低，主要集中在客户营销、客户运营、客户服务等获取和转化客户的成本方面，具有高替代潜力。

国产大模型如火如荼，全球科技竞赛的你追我赶也仍旧激烈。2023年11月，恰逢ChatGPT推出一年的时间节点，OpenAI举办首届全球开发者大会，其中GPT-4升级为GPT-4 Turbo成为最重磅的“炸弹”。

第三方排行榜SuperCLUE基于SuperCLUE通用大模型综合性中文测评基准，对GPT-4 Turbo进行了全方位测评。结果显示，GPT-4 Turbo在10项基础能力中有8项满分，相比上一代GPT-4模型，GPT-4 Turbo有10.33分的巨大提升。

“对国内大模型而言，差距在进一步扩大。GPT-4 Turbo总分领先国内最强模型有30分以上。”最后一条结论中，SuperCLUE如此写道。

中国大模型和美国距离多远？半年前，出门问问CEO李志飞给出更“直观”的计算答案是16个月。

他解释，2022年1月，谷歌发布指令学习大模型FLAN，之后的2022年10月ChatGPT发布，2023年3月GPT-4发布。中国企业目前发布的一批大模型与FLAN水平相近，如此推算，中美大模型的差距为16个月。

当下的美国，由OpenAI、微软和英伟达组建的软硬件联盟，正加速推动AI 2.0时代到来。与此同时，国产大模型也在遭遇芯片供应的不确定性风险。

10月，美国商务部工业与安全局（BIS）发布更新针对人工智能（AI）芯片的出口管制规定，芯片与半导体行业首当其冲。

国内GPU设计企业壁仞科技总裁徐凌杰预测称，未来一年，中国绝大部分的算力需求将集中在大模型训练上；此后，推理的场景会变得越来越多。

一个广泛的共识是，中国的特点是应用丰富、落地更快，大部分企业并非靠技术取胜，而是靠在落地中挖掘新的生产力。