大模型行业加速分化

图为中国大模型企业参展2025世界人工智能大会

本报记者 陈存

AI大模型企业DeepSeek近期频频进入大众视野。

这家曾在2024年掀起模型开源大潮的企业，在此后一年多的时间里一度显得十分低调。市场的目光更多转向豆包、千问等互联网大厂研发的大模型，或是智谱、MiniMax这类率先上市的AI公司。

最近，两则消息让DeepSeek重回焦点位：一是其正在开展首轮融资谈判，在短短一个多月内，预计估值飙升至450亿美元；二是其最新发布的V4-Pro系列模型，在经历多轮降价后，宣布将永久降价75%。

资本动作和价格调整，看似是企业经营层面的常规操作，背后折射出的却是大模型行业的新变化——当技术能力逐渐拉平，独立大模型公司正围绕资本与生态开启新一轮竞速，而这也将重新划分企业“座次”。

大模型进入“头部玩家”时代？

大模型行业就此进入“头部玩家”时代，留在“牌桌”上的企业不多了。

今年以来，大模型行业单轮融资纪录不断刷新，其融资速度可称“疯狂”。

未上市企业中，此前长期拒绝外部融资的DeepSeek，被曝首次启动融资接触，估值约450亿美元，近期还在推进一笔规模达700亿元的融资。Kimi（月之暗面）完成20亿美元的最新融资，投后估值升至200亿美元。阶跃星辰也即将完成近25亿美元的融资，并已拆除红筹架构，加速赴港IPO准备。

而于今年1月港股上市的智谱和MiniMax，涨幅更是超过300%。5月22日港股收市后，恒生指数公司公布季度检讨结果：将MiniMax-W及智谱这两只AI概念股纳入恒生科技指数。这意味着，香港最大的30家科技主题上市公司中，开始出现AI原生大模型企业。研究显示，智谱或将因此吸引510亿元至920亿元的南向资金流入，MiniMax则可能吸引约470亿元的资金。

头部大模型企业不仅融资金额动辄高达百亿元，其投资方也是阵容豪华。DeepSeek首轮投资即吸引腾讯与阿里巴巴；Kimi叠加了阿里、腾讯、美团龙珠等知名资方的加持；阶跃星辰则获腾讯的三次重仓。近期，更有国智投、北京人工智能基金、中国移动等国资背景企业与机构入局。

这一场景与2023年“百模大战”之时高度相似，却又存在微妙的差别。彼时，国内大模型公司数量快速膨胀，只要有技术团队、有模型能力，就有机会获得融资，“AI六小虎”成为一级市场最受关注的创业群体之一。

而到了2025年，据统计，AI模型层公司全年仅完成22笔融资，单轮融资规模在10亿元以上的大模型公司仅有MiniMax、智谱和Kimi三家。钱并没有消失，而是在快速向头部集中。

即使是盛极一时的“AI六小虎”亦有分化。智谱和MiniMax率先敲钟，月之暗面和阶跃星辰分别押注深度思考与端侧模型；而零一万物与百川智能已悄然放弃基础模型的竞逐，转身扎进更为垂直的AI应用赛道。

2023年，百川智能创始人王小川曾放话“在年底做出国内最好的大模型”“3年内追上GPT-4”。2025年，王小川在全员信中反思，过去两年“战线拉得过长，不够聚焦”“过早进入商业化”，并称“接下来将围绕百小应、AI儿科、AI全科、精准医疗四个方向专注聚焦，减少多余的动作”。零一万物也停止了超大基模（万亿参数以上）训练业务，全面聚焦To B垂直场景，转向了轻量化产业大模型与AI Agent研发。

市场仍旧繁荣，但资本不再雨露均沾。大模型行业就此进入“头部玩家”时代，留在“牌桌”上的企业不多了。

IPO，下半场的入场券？

行业出现一个极其矛盾的现象：用户越多，大模型公司可能越容易亏钱。

那么，尚且留在“牌桌”上的企业就安全了吗？

对未上市的企业而言，跻身下一个时代的通道正在收窄，不完成惊险一跃或许就将跌落谷底；即便成功上市，大模型公司真正的压力，可能也才刚刚开始。

资本“看人下菜碟”的背后是强烈的市场信号：讲好故事就能获得充沛资金的时代结束了，只有持续创新、创造收益，才能获得回报。

这与大模型产业的特性密不可分。过去的互联网行业，遵循的是“用户越多、边际成本越低”的商业模式。微信新增一个用户，腾讯的成本不会同步增长；抖音新增一个用户，字节跳动反而能获得更多广告收入。但大模型公司的每一次Token调用、每一次复杂任务、每一次长文本生成，都对应真实的算力消耗。尤其是在Agent、AI生成逐渐普及后，成本还在进一步增加。

与此同时，模型本身却越来越像一种“公共能力”，开始迅速“贬值”。2023年时，长文本、多模态被视为稀缺能力；到了今天，模型之间虽然仍有差距，但已经很难再形成碾压态势，“花小钱办大事”成为多数用户的优先选择。DeepSeek的开源与低价策略，进一步拉低了用户的成本预期。

这也导致行业出现一个极其矛盾的现象：用户越多，大模型公司可能越容易亏钱。

字节跳动有广告业务输血，腾讯的游戏和社交业务足够赚钱，阿里拥有电商和云计算体系作为支撑，但它们同样要考虑变现：阿里曾常年贴钱做AI，刚刚进入回报期；字节旗下的豆包也开始探索收费。独立大模型公司没有可背靠的母公司，处境更加艰难，也更依赖外部资金的注入。

这一逻辑对已上市的MiniMax和智谱也同样适用。估值疯狂上涨的同时，其营收与利润数据却反映出另一重现实。2025年，智谱经调整后净亏损31.82亿元，毛利率从2024年的56.3%下降至41.0%；MiniMax毛利率从12.2%改善至25.4%，经调整净亏损约17.3亿元。

这也是为什么IPO变得越来越重要。对于许多大模型公司而言，上市就意味着获得了一个长期、公开、持续的融资渠道。

多维博弈，谁主沉浮？

单一的模型技术对决，正演变为赛道差异化、生态立体化、落地场景化的多维博弈。

正因如此，独立大模型公司已经越来越少单纯宣传“谁的模型最强”，而开始强调另一件事：自己到底扮演着什么角色。单一的模型技术对决，正演变为赛道差异化、生态立体化、落地场景化的多维博弈。

DeepSeek锚定开源生态与极致低价的技术路线，抢占开发者与中小企业市场。近日，DeepSeek官宣V4-Pro模型API永久降价75%，同等业务量下，其调用成本仅为GPT、Claude等海外模型的几十分之一。爆出融资消息后，梁文锋在投资者会议上仍强调研发开源模型和实现通用人工智能（AGI）的目标。

阶跃星辰深耕端侧AI，锚定实体场景。2025年世界人工智能大会期间，阶跃星辰创始人兼CEO姜大昕宣布最新一代多模态推理大模型Step-3发布，多模态推理能力落地汽车和手机两大智能终端。目前，在手机端，阶跃星辰已与OPPO、荣耀、中兴等主流品牌达成深度合作；在汽车端，携手千里科技、吉利打造智能座舱。

Kimi押注编程能力和Agent集群两大核心方向，彻底砍掉了此前分散资源的海外C端产品和视频生成业务。今年2月，Kimi推出Kimi Claw，定位为“云端化”的OpenClaw，并直接配置了5000多个ClawHub社区技能。

智谱主打政企市场与产业落地能力，聚焦To B、To G产业赛道，深耕金融、政务、能源、工业、教育等重点领域，打造标准化行业解决方案，从去年起慢慢弱化面向C端产品的资源投入，基本叫停智谱清言在C端的宣传和投放。

MiniMax聚焦全模态能力与全球化生态，依托多模态融合技术，在文本、图像、语音、视频生成领域形成独特优势。同时，产品兼顾海外市场扩张，面向海外开发者与企业客户开放，旨在构建全球化服务生态。

2025年，时任OpenAI研究员的姚顺雨在自己的博客中提出了“AI下半场”理论，指出行业将从“拼参数、比性能”的上半场，转向“重落地、讲价值”的下半场。时隔一年，必须承认，国内AI行业竞争格局已然改写。资本助推之下，单纯技术尝鲜与模型比拼的时代远去，AI竞争的核心只剩一件事——如何实现真实可衡量的业务价值。

2026智能养老服务机器人应用大赛收官

本报讯 记者**杨鹏岳**报道：5月25日，以“机器人赋能养老，科技温暖夕阳”为主题的2026智能养老服务机器人应用大赛在京津冀大数据创新应用中心成功举办。本次大赛吸引了来自全国57支企业、高校及科研院所的团队参赛，围绕康复机器人与养老机器人两大赛道、八大核心任务赛项展开实战对决。

记者在现场了解到，与以往侧重技术展示的赛事不同，本届大赛摒弃单纯产品展示，立足养老服务真实痛点，设置辅助行走、康复支持、移乘转运、生活照料、二便护理及助浴、健康管理、情感陪护、智慧环境等八大任务赛项，将机器人置于家庭、社区、机构等模拟真实养老场景中实战检验。参赛团队需完成实景任务演示，接受专家评审、安全核验与用户评价。

从现场表现看，参赛产品正从单点功能转向系统化服务。其中，在移乘转运赛项中，床椅一体化设备完整演示了起背、抬腿、翻身、床椅转换、转运洗浴等全流程，将以往需多人协作的高强度护理转化为智能化操作；在二便护理及助浴赛项中，产品展示了自动识别、清洗、烘干、除臭、污物封闭处理等功能，兼顾护理效率与使用者尊严；智慧环境赛项则呈现了物资配送、垃圾回收、巡航清洁、云端调度等能力，推动养老机构日常服务向可调度、可追踪的智能流程演进。

经过激烈角逐，12支团队分获一、二、三等奖。其中，北京大艾机器人科技有限公司凭借下肢外骨骼康复训练机器人摘得康复机器人任务挑战赛一等奖，苏州伊利诺护理机器人有限公司斩获养老机器人任务挑战赛一等奖。大赛奖金合计近40万元，进入现场评审的团队均获创新奖。

当前，我国养老服务面临人力短缺、失能照护压力大、独居老人关怀难等现实挑战，智能养老服务机器人产业正从技术概念迈入场景落地关键期。本次大赛同期举办成果展示、创新发布与供需对接活动，推动技术从赛场走向市场。

“养老机器人产业需以真实需求为导向，在安全性、实用性、经济性上持续突破。”业内专家向记者表示，本次赛事为产学研用搭建了高效对接平台，有望加速智能养老设备进入千家万户与养老机构，以科技力量助力养老服务提质升级。

本次大赛由河北省工业和信息化厅、河北省民政厅、廊坊市人民政府、中国软件评测中心（工业和信息化部软件与集成电路促进中心）、云港陪伴新智人（廊坊）科技有限公司联合主办。

华为推出AI DC数据基础设施全栈方案

本报讯 5月21日，2026华为创新数据基础设施论坛在巴黎举行。论坛上，AI DC数据基础设施全栈方案正式发布。

论坛上，华为公司副总裁、数据存储产品线总裁袁远分享了当下人工智能产业发展的两大方向：一是智能体加速普及，正在成为企业常态化“数字员工”，目前已有超3000万个活跃智能体为人类提供服务；未来五年，这一数字将飙升至22亿；二是AI应用持续深化，Token（词元）将成为新一代通用算力凭证，去年全球每分钟令牌处理量约60亿个，今年已攀升至每分钟150亿个。面向这些发展趋势，袁远指出：“企业要加速AI落地，需推动现有IT架构向AI DC数据基础设施快速演进，围绕数据湖、知识与记忆平台、算力、模型、Agent框架与数据韧性等核心方向，开展系统性规划与建设。”

基于此，华为正式发布AI DC数据基础设施全栈方案，旨在助力企业加速推进AI数据中心建设，实现AI规模化落地。

据介绍，华为通过OceanStor Pacific全闪分布式存储和DME Omni-Dataverse统一数据空间完善AI数据湖体系，实现高质量数据汇聚与供给；面向超大规模推理集群场景和面向企业AI推理场景，华为分别推出业界首个支持异构算力的上下文记忆存储CMS（Context Memory Storage）和首创的“3+1”AI数据平台，前者可实现推理首Token时延降低90%，后者可提升30%推理精准性；在模型工程与资源调度领域，华为Model Engine具备模型开箱即用、模型网关的能力，可支持0码适配新模型、一键部署模型；Agent框架方面，华为Model Engine Nexent智能体平台可通过自然语言交互方式直接生成Agent，大幅降低开发门槛，使Agent上线周期缩短80%；此外，针对Agent、模型、平台与基础设施等不同维度的潜在数据安全风险，华为数据韧性平台通过构建防滥用、防投毒、防篡改、防勒索的端到端数据保护方案，全面守护AI数据资产安全。

袁远表示，人工智能的发展经历了三个阶段：第一阶段是算力，随着GPT普及，通用算力已实现全球普及；第二阶段是模型，此前已经经历了各类大模型百花齐放的阶段；第三阶段是智能体，智能体开发与应用正成为行业热潮。在他看来，人工智能的下半程核心在于数据：“数据既决定AI安全底线，也决定其产业落地价值，拥抱数据、夯实数据底座，已是各行各业拥抱AI的必由之路。” （张琪玮）

《中国“人工智能+”能源发展报告2026》发布

本报讯 近日，国家能源局组织编制的《中国“人工智能+”能源发展报告2026》发布。这是我国能源领域人工智能与能源融合发展方面的首份年度报告。

报告系统回顾了国内外人工智能与能源融合发展进展，深入研判了中国“人工智能+”能源发展形势，并对下一阶段重点方向作出展望。

报告称，“人工智能+”能源已成为世界主要国家抢占未来发展主动权的重要战略方向。当前，人工智能快速发展带动全球算力设施用电需求持续增长。据国际能源署预测，到2030年，全球数据中心用电量将较2025年接近翻番。

在算力设施绿色低碳转型和算电协同方面，2025年，我国已建成42个万卡级智算集群，全国算力中心总用电量达1700亿千瓦时。全国一体化算力网络八大枢纽节点算力用电成为增长的主要来源，近3年平均增速约为39.5%，远高于全社会用电量的平均增速。其中，京津冀枢纽节点、内蒙古枢纽节点用电量近3年平均增速分别达到33.3%和66.5%，反映出算力资源向重点枢纽和能源资源富集地区加快集聚。

同时，国家算力枢纽节点新建数据中心绿电占比超过80%的目标要求，正通过绿电交易、绿电直连、跨省跨区交易、源网荷储一体化等多种方式加快落实。

在行业大模型发展方面，已落地数十个能源行业专用大模型，覆盖电网、新能源、水电、火电、核电、煤炭、油气等领域，大模型场景适配能力持续增强。

国家能源局相关负责人表示，随着我国“人工智能+”能源从概念走向实践、从探索走向推广，产业形态加速演进、创新应用多点突破、融合基础不断夯实，将加快推动人工智能和能源双向赋能，促进能源领域新质生产力跃升发展和生产关系深层次变革。 （文 编）