# DeepSeek“吸睛”破局背后

## 颠覆了什么？搅了谁的局？

一家人工智能初创企业浅浅扇动两下翅膀，掀起全球科技界的一阵“海啸”。

短短30天，中国初创企业深度求索（DeepSeek）先后发布两款性能比肩GPT-4o的大模型，“1/18的训练成本、1/10的团队规模、不分伯仲的模型性能”令硅谷大受震撼。

事实上，这匹黑马的贡献绝非“低成本”这一个标签所能概括。它不仅重新定义了大模型的生产函数，还将重新定义计算。

不论开源与闭源未来的优势如何，这股冲击波都将迫使全球科技界重新思考：当“规模定律”与“生态壁垒”不再绝对，什么才是下一赛季AI竞争的核心？或许我们能从中获得新的启示。

### 1 击穿三大定式

1月下旬，DeepSeek在中区、美区苹果App Store下载榜单中登顶，超越ChatGPT、谷歌Gemini等全球顶尖科技巨头研发的模型产品。具体而言，它颠覆了什么？

**——打破“越强越贵”的成本诅咒**

价格感人是让DeepSeek快速出圈的第一个标签。DeepSeek-R1的API服务定价为每百万元输入tokens1元（缓存命中）/4元（缓存未命中），每百万元输出tokens16元，而o1模型上述三项服务的定价分别是55元、110元、438元。

凡是使用过几款大模型的用户很快就能形成这样一个共识：就推理能力而言，DeepSeek直逼OpenAI的o1、Meta的Llama-3等一流模型，甚至在回答问题之前还能给出它的推理过程和思考链路。AI投资机构Menlo Ventures负责人Deedy对比谷歌Gemini和DeepSeek-R1后表示，DeepSeek-R1更便宜、上下文更长、推理性能更佳。低成本比肩o1模型，令硅谷的“烧钱模式”一时间遭到猛烈质疑。

**——超越“性能-成本-速度”的不可能三角**

当硅谷仍在为GPU万卡集群豪掷千亿资金时，一群土生土长的中国年轻人用557.6万美元证明：AI大模型的比拼或许并不只靠规模，更重要的是看实际效果。有句话形象地概括出DeepSeek的优势：“不是GPT用不起，而是DeepSeek更具性价比。”

传统模型训练，需要在性能、成本、速度之间权衡，其高性能的获得，需要极高的成本投入、更漫长的计算时间。而DeepSeek重构了大模型的“成本-性能”曲线，同时压缩了计算周期。

根据DeepSeek技术报告，DeepSeek-V3模型的训练成本为557.6万美元，训练使用的是算力受限的英伟达H800 GPU集群。相比之下，同样是开源模型的Meta旗下Llama-3.1模型的训练成本超过6000万美元，而OpenAI的GPT-4o模型的训练成本为1亿美元，且使用的是性能更加优异的英伟达H100 GPU集群。而使用过程中，DeepSeek给出反馈的时长也大部分控制在5秒至35秒之间，通过算法轻量化、计算效率最大化、资源利用率优化，成功压缩了计算时间，降低了延迟。

**——走出“参数膨胀”陷阱**

ChatGPT横空出世后700多天里，全球人工智能巨头不约而同走上了一条“大力出奇迹”的“暴力美学”路线，参数越“炼”越大，给算力、数据、能耗带来了极大压力。很长一段时间，参数几乎成为大模型厂商比拼的最大焦点。

而另辟蹊径的DeepSeek恰巧处于对角线的另一端：并不盲目追求参数之大，而是选择了一条通过探索更高效训练方法以实现性能提升的“小而精”路线，打破了“参数膨胀”的惯性。

例如DeepSeek-R1（4B参数）在数学推理、代码生成等任务上具有比肩70B参数模型（如Llama-2）的能力，通过算法优化、数据质量提升，小参数模型一样能实现高性能，甚至能够“四两拨千斤”。

### 2 实现三大跃升

“DeepSeek出圈，很好地证明了我们的竞争优势：通过有限资源的极致高效利用，实现以少胜多。中国与美国在AI领域的差距正在缩小。”面壁智能首席科学家刘知远说。

算力封锁下的有力破局，得益于DeepSeek技术架构、数据策略、工程实践三方面的关键突破。

**——技术架构：重新定义参数效率**

大模型的千亿参数不应是冰冷的数字堆砌，而应是巧夺天工般重组整合。

传统大模型Transformer架构好比一条承载车辆的高速公路，当车辆（数据）数量足够多的时候，每辆车必须和前后所有车沟通完成才能继续行驶（计算），导致堵车（计算慢、能耗高）。而DeepSeek创新的架构则把一条串行的高速路，变成了一个辐射状的快递分拣中心，先把货物（数据）按类型分类打包，再分不同路线同时出发开往不同目的地，每辆货车（计算）只需选择最短路径。因此既能提高速度又能节约能耗。

**——数据策略：质量驱动的成本控制**

DeepSeek研发团队相信，用“炼数据”取代“堆数据”，能使训练更具效率。

传统的数据策略好比去农场随便采捡，常有价值不高的烂菜叶（低质量数据）。而DeepSeek创新的数据蒸馏技术，有针对性地筛选掉质量不高的烂菜叶：一方面自动识别高价值数据片段（如代码逻辑推理链），相比随机采样训练效率提升3.2倍，另一方面通过对抗训练生成合成数据，将高质量代码数据获取成本从每100个tokens的0.8元降低至0.12元。

**——工程实践：架起“超级工厂”流水线**

大模型传统的训练方式好比手工造车，一次只能装配一台，效率低下。而DeepSeek的3D并行相当于一方面通过流水线并行把造车流程拆分为10个步骤，同时组装10辆车（数据分块处理），另一方面通过张量并行，把发动机拆成零件，分给10个工厂同时生产（模型分片计算）。

至于推理过程，传统模型好比现点现做的餐厅，客户等菜时间长，推理过程慢。而DeepSeek采用的INT4量化，能把复杂菜品提前做成预制菜，加热（计算）时间减半，口味损失不到5%，实现了大模型的低成本工业化。

网友使用DeepSeek提问

纵深

## “聚光灯之外”的安全问题

从技术到愿景，DeepSeek坚定选择的始终是一条难且正确的路。这也是为什么，即便别国在人工智能领域已坐享先发优势，后发者依然有机会凭借技术创新、成本革命打破大模型竞争的传统逻辑，打破人工智能行业竞争格局，打破“他国更擅长从0到1的原始创新，而中国更擅长从1到10的应用创新”的成见，重塑竞争优势的奥秘。

正如梁文锋此前接受采访时所说，“中国要逐步成为贡献者，而不是一直‘搭便车’。”

DeepSeek爆火背后，人工智能等前沿技术不断迭代、尽显锋芒，也让其知识产权、网络安全风险分散机制成为“聚光灯之外”的重要议题。

“近期DeepSeek线上服务受到大规模恶意攻击，注册可能繁忙，请稍等重试。”近日，当记者打开DeepSeek网页版时，在其醒目位置看到了这样的提示。这也让高科技产品背后的风险分散机制受到关注。

多位业内人士表示，科技创新面临很大的不确定性，风险高。而科技保险作为一种有效的风险管理手段，如果安排得当，可以有效发挥风险分担作用。目前，在政策鼓励下，市场上的科技保险产品已覆盖包括研发、成果转化及市场推广等科技项目全生命周期。

据记者了解，目前政策激励主要集中在保费补贴等需求端，对保险公司的补贴激励和数据资源支持较少。因为“看不懂”“算不清”，保险公司一般不会贸然开展此类业务，科技创新领域仍有大量的风险保障空白待填补。

清华大学五道口金融学院副教授周臻建议，针对不同发展阶段和不同行业的企业，完善覆盖科技发展全生命周期的扶持体系，设计富有针对性的补贴政策，完善补贴体系，保持政策的一致性与连贯性，促进重点领域的科技创新。

多位专家呼吁，可由政府部门牵头，设立有专业服务能力的第三方机构，为科技保险的核保、定价、定损、理赔制定标准、提供依据，让科技企业的风险被“看得懂”“算得清”，推动更多满足科技企业特定需求的产品面世。

本报综合经济参考报、中证报等

相关

### 被DeepSeek刺激到了？文心一言、ChatGPT宣布免费

随着文心大模型的迭代升级和成本不断下降，文心一言在官网宣布将于4月1日0时起全面免费，所有PC端和APP端用户均可体验文心系列最新模型，以及超长文档处理、专业检索增强、高级AI绘画、多语种对话等功能。

文心一言官网还同步透露即日起上线深度搜索功能，具备更强大的思考规划和工具调用能力，可为用户提供专家级内容回复，并处理多场景任务，实现多模态输入与输出。当前，用户可在文心一言官网上体验深度搜索功能，APP端也即将上线。

2月13日凌晨3点，OpenAI首席执行官Sam Altman公布了GPT-4.5/5将很快陆续发布，免费版ChatGPT将在标准智能设置下无限制使用GPT-5进行对话。

从公开表态来看，免费版ChatGPT能在标准智能设置下无限制地使用GPT-5进行对话，不过会防止滥用。

主流判断认为，这意味着OpenAI“火力全开”，背后是DeepSeek的“刺激效应”。

据证券时报