爱建证券 AIJIAN SECURITIES

行业研究 / 行业点评

2025 年 02 月 19 日

适合投资人的 DeepSeek 分析报告

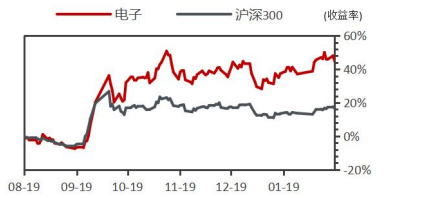
——人工智能专题报告（1）

行业及产业

电子 /

强于大市

一年内行业指数与沪深 300 指数对比走势：



资料来源：爱建证券研究所、聚源数据

相关研究

本期投资提示：

- 2025 年 1 月 20 日，DeepSeek 发布自研模型 R1 在全球科技行业引起的震动，被西方媒体称为“DeepSeek Shock”。DeepSeek 不仅在媒体圈迅速爆火，同样也成为了资本市场的宠儿，其概念指数仅仅诞生 10 日后板块成交金额就超过了全部 A 股成交额的 20%。这一切现象究其原因，是因为 DeepSeek 通过技术微创新，以更低的硬件成本和更短的时间实现了可以与市场领先产品竞争的能力。更重要的是，**DeepSeek 打破了行业“限制中国企业对于最先进 GPU 获取，将能够阻止中国 AI 技术发展”的一贯认知。**
- 本文有别于市场上大部分研究报告冗长的技术细节描述，我们针对投资人短时间内客观理解 DeepSeek 的需求，加入了不同于市场的思考和量化的分析比较。以下是报告的核心观点：
- DeepSeek 的技术创新在哪里？**1）首创 DeepSeekMoE 架构，专门设计用于实现终极专家专业化。DeepSeekMoE 通过降低激活参数比例，实现了训练效率 3.6X 的提升和训练吞吐量 3.6X 的提升。2）通过引入 MLA 机制，DeepSeek-V2 实现了显著增强的性能，节省了 42.5% 的训练成本、减少了 93.3% 的 KV 缓存、并将最大生成吞吐量提升至 5.76 倍。**
- DeepSeek-V3 实际开发成本几何？公司官宣正式训练成本为 580 万美元，但是并没有披露隐性成本。**DeepSeek-V3 是建立在前期模型基础上开发的，前期研发投入约为 2000-3000 万美元。其他数据获取成本和硬件折旧成本未披露，实验试错成本约为 500 万美元，因此预计实际总成本超过 4000 万美元。实际成本虽然高达公开口径成本的 7 倍左右，但是仍然相对 Llama 3-405B 降低了约 69%；相对于 GPT-4o 降低了 95%。**
- 未来 GPU 算力需求会大幅下降吗？短期内云服务大厂资本开支持续处于上升通道，这是由于 Scaling Law 导致行业对于算力军备竞赛的恐慌性投资仍然存在。但是 Scaling Law 中，数据资源同样限制着模型性能的提升。**根据 EPOCH AI 预测，到 2028 年人类生成的公共文本数据总有效库存量约为 300T token 将被全部耗尽。换句话说，在现有模型框架和数据资源供给下，2028 年之后单纯算力提升将难以继续推动模型性能的升级。**

核心结论：

- **给予行业“强于大市”评级。**随着 DeepSeek 的横空出世，低成本高性能的模型训练部署成为可能。我们预计接入 DeepSeek API 的细分领域推理服务商将会快速涌现，同时终端设备将会能够实现蒸馏小模型的本地部署能力，包括服务器，智能手机和智能驾驶系统都将会迎来新一轮的升级浪潮。我们看好以下细分板块在 DeepSeek 推动下的未来发展：
  - 先进算力芯片制造产业链：算力芯片设计，先进制程晶圆代工，先进封测等等。
  - 专业咨询服务：包括 AI+医疗，法律，金融，会计，教育，政务等咨询服务领域。
  - 2C 智能终端：AI+智能手机，智驾汽车等终端产业链。
  - 2B 本地部署设备：服务器 OEM 及产业链。
- 风险提示：大模型开发进度缓慢，模型升级不及预期。AI 行业应用落地迟缓，商业模式难以实现良性循环。贸易摩擦加剧，先进芯片及半导体技术受限。

证券分析师

许亮

S0820525010002

0755-83562506

xuliang@ajzq.com

联系人

目录

1. DeepSeek 从何而来 4

1.1DeepSeek 公司诞生 4

1.2DeepSeek 里程碑事件 4

1.3DeepSeek 的爆火 5

2. 深入剖析 DeepSeek 8

2.1DeepSeek 为什么被称为 shock? 8

2.2DeepSeek 的技术创新在哪里? 11

2.3 技术细节之外的信心提升 12

3. 市场对于 DeepSeek 最关心的问题 15

3.1DeepSeek-V3 的模型实际成本几何? 15

3.2GPU 为代表的算力需求是否会大幅下降? 16

4. DeepSeek 对产业发展的影响 19

5. 风险提示 20

图表目录

图 1：DeepSeek 里程碑事件 4
图 2：DeepSeek-R1 发布后话题爆火 5
图 3：DeepSeek 概念占 A 股成交额比例超过 20% 5
图 4：DeepSeek 通用模型 V3 性能达到行业领先水平 8
图 5：DeepSeek 通用模型 V3 成本优势明显 9
图 6：DeepSeek 推理模型 R1 性能达到行业领先水平 9
图 7：DeepSeek-R1 价格优势明显 10
图 8：DeepSeekMOE 架构可以用更少的训练参数实现更好的性能表现 11
图 9：DeepSeek-V2 通过 MLA 机制实现了性能优势和成本降低 12
图 10：Scaling Law 中算力与 AI 能力的关系 13
图 11：DeepSeek 带动中国资本市场信心提升 14
图 12：美国主要互联网企业资本支出金额（亿美元） 16
图 13：人类生成的公共文本数据将在 2028 年耗尽 17
图 14：OpenAI 模型参数规模快速靠近公共文本数据上限 18

表 1：DeepSeek-V3 和 DeepSeek-R1 比较 8
表 2：DeepSeek-R1 蒸馏小模型本地化部署 10
表 3：DeepSeekMoE 的效率创新 11
表 4：国内外高端 GPU 芯片性能比较 13
表 5：DeepSeek 开发具备成本优势 14
表 6：DeepSeek-V3 正式训练阶段成本拆分 15
表 7：DeepSeek-V3 隐性成本拆分 15
表 8：DeepSeek-V3 实际成本对比 16
表 9：主要细分行业模型参数规模 18

1. DeepSeek 从何而来

1.1DeepSeek 公司诞生

2023 年 7 月，DeepSeek 公司由幻方量化创始人梁文锋主导创立，其团队依托幻方投资的资金与“萤火超算”万卡级算力资源（万张 A100 GPU），致力于 AGI 技术探索。2023 年 7 月 17 日，杭州深度求索人工智能基础技术研究有限公司（DeepSeek）正式注册，定位为技术驱动的开源 AI 公司。

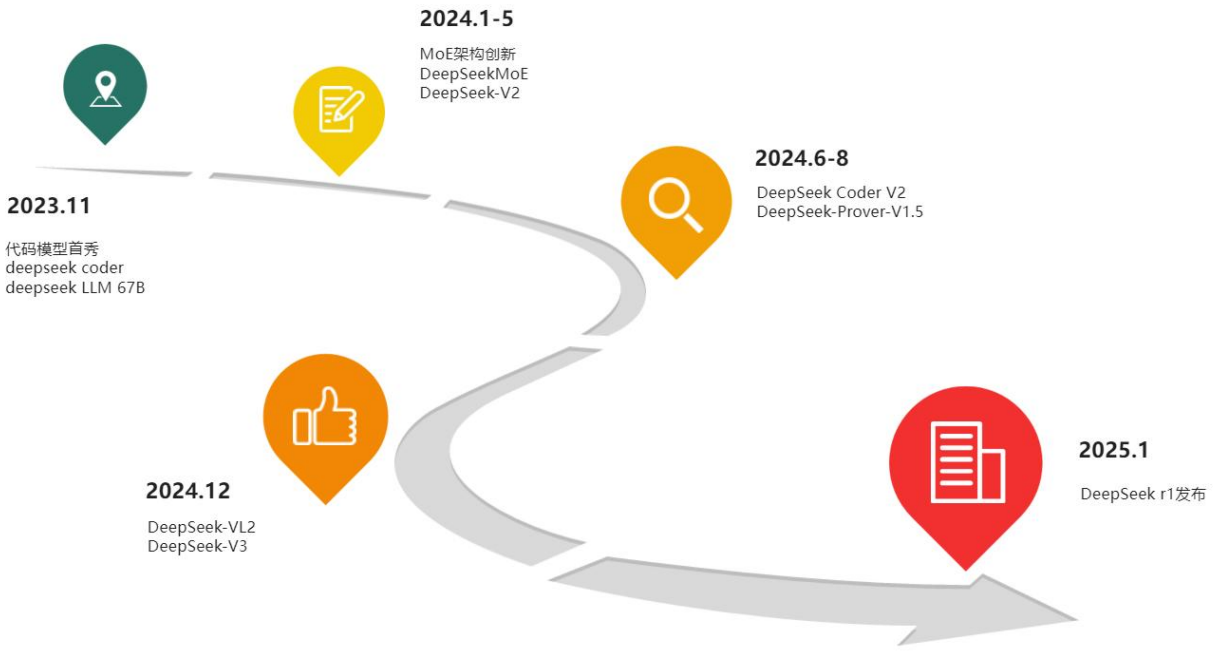
1.2DeepSeek 里程碑事件

DeepSeek 整个发展历程可以分为五个阶段：

阶段一：2023 年 11 月，DeepSeek 代码模型首秀。主要包括：DeepSeek Coder：首个开源代码大模型，支持多语言生成与调试，且性能超越 CodeLlama，奠定了技术口碑。DeepSeek LLM 67B：通用大模型开源，对标 LLaMA2 70B，中英文任务表现领先。

阶段二：2024 年 1 月-5 月，DeepSeek 实现了 MoE 架构创新。发布 DeepSeekMoE 国内首个开源 MoE 模型，采用细粒度专家共享架构。DeepSeek-V2 第二代 MoE 模型，引入 MLA（多头潜在注意力）技术，推理成本仅为 LLaMA3 的 1/4，API 定价低至 GPT-4 Turbo 的 1/70，大幅拉低 AI 使用成本。

图 1：DeepSeek 里程碑事件



资料来源：DeepSeek，爱建证券研究所

阶段三：2024 年 6-8 月，多领域拓展与性能跃升。DeepSeek 发布垂直领域模型 DeepSeek Coder V2（2024 年 6 月）：代码能力超越 GPT-4 Turbo。DeepSeek-Prover-V1.5（2024 年 8 月）：数学推理模型，覆盖初等数学至研究生水平。

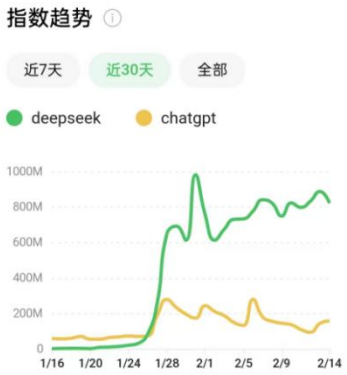
阶段四：2024 年 12 月，实现通用模型的迭代。DeepSeek-V3 发布，公司宣称训练成本仅 550 万美元，性能对标国际闭源模型，生成速度提升 3 倍。DeepSeek-VL2（2024 年 12 月）：多模态 MoE 模型，视觉能力显著提升。

阶段五：2025 年 1 月 20 日，DeepSeek 正式发布第一代推理模型 DeepSeek-R1-Zero 和 DeepSeek-R1。

1.3DeepSeek 的爆火

随着 DeepSeek-R1 发布，相关话题在媒体网络爆火，微信指数迅速超越 ChatGPT。2025 年 2 月 5 日，DeepSeek 同样也成为了资本市场的宠儿。**DeepSeek 概念指数仅仅诞生 10 日后，2025 年 2 月 14 日 DeepSeek 板块成交金额就超过了全部 A 股成交额的 20%。**同时期内，沪深 300 指数涨幅约为 3.8%，DeepSeek 成为了这一轮市场上涨的重要推动力量。

图 2：DeepSeek-R1 发布后话题爆火



资料来源：微信指数，爱建证券研究所

图 3：DeepSeek 概念占 A 股成交额比例超过 20%

资料来源：iFind，爱建证券研究所

除了在媒体和资本市场的火爆，DeepSeek 是国内首个获得各行各业认可并实际接入的大模型。具体来看，目前已经接入 DeepSeek 的已经包括云服务领域，网络安全领域，办公&教育，医疗，政务等等行业的多家国内外知名企业。

■ 云服务领域

华为云：2 月 1 日，硅基流动与华为云团队联合首发并上线基于华为云昇腾云服务的 DeepSeek 推理服务，让模型能够在大规模生产环境中稳定运行。腾讯云：2 月 2 日宣布在高性能应用服务“HAI”上支持一键部署 DeepSeek-R1 模型，开发者仅需 3 分钟即可接入调用。阿里云：2 月 3 日宣布阿里云 PAI Model Gallery 支持云上一键部署 DeepSeek-V3、DeepSeek-R1。百度智能云：2 月 3 日宣布

DeepSeek-R1 和 DeepSeek-V3 模型已在百度智能云千帆平台上架，并推出了超低价格方案。京东云：2 月 4 日宣布正式上线 DeepSeek-R1 和 DeepSeek-V3 模型，支持公有云在线部署、专混私有化实例部署两种模式。联通云：2 月 3 日宣布联通云已基于星罗平台实现国产及主流算力适配多规格 DeepSeek-R1 模型。天翼云：2 月 6 日，电信天翼云自主研发的“息壤”智算平台率先完成国产算力与 DeepSeek-R1/V3 系列大模型的深度适配优化。火山引擎：支持 V3/R 等不同尺寸的 DeepSeek 开源模型，提供高性能推理服务。

■ **网络安全领域**

360 集团：2 月 2 日宣布其安全大模型正式接入 DeepSeek，将以 DeepSeek 为安全大模型基座，训练出“DeepSeek 版”安全大模型。安恒信息：2 月 4 日发布信息称，旗下恒脑·安全垂域大模型正式集成 DeepSeek，完成基于 DeepSeek-R1 的安全大模型的训练。奇安信：2 月 5 日宣布自主研发的 QAX 安全大模型已全面完成了 DeepSeek 的深度接入。亚信安全：2 月 5 日宣布基于 DeepSeek-V3/R1 构建智能体，能够在海量的安全告警中，快速、低成本地挖掘到不同来源、不同类型安全告警中的关联关系。安博通：2 月 7 日，安博通下一代 AI 防火墙与人工智能大模型强强联合，搭载 DeepSeek-R1-Distill-Qwen-32B 模型。

■ **办公&教育领域**

视觉中国：完成 DeepSeek-R1 接入与本地化部署，并在多个产品中深度应用其能力。钉钉：钉钉 AI 助理接入 DeepSeek，支持深度思考。用友：通过引入 DeepSeek 进一步优化 YonSuite 的 AI 能力，并将以更多数据开源模型训练 YonGPT 模型。飞书：飞书多为表格、飞书智能伙伴已接入 DeepSeek。ima：正式接入 DeepSeek-R1 模型，在使用搜、读、写和知识库的时候，可以选择腾讯混元大模型或 DeepSeek-R1 模型。网易有道：全面接入 DeepSeek-R1，AI 全科学习助手“有道小 P”结合 DeepSeek-R1 超长思维链所提供的思考及分析能力，进一步优化了个性化答疑功能。云学堂：已全面接入 DeepSeek-R1/V3 大模型，云学堂的 AI 制课专家、AI 学习地图、AI 学习专家、AI 对练等产品均能够使用包括 DeepSeek 在内的多种大模型能力。万兴科技：已完成与 DeepSeek 最新推理大模型 DeepSeek-R1 的深度适配，旗下产品如万兴喵影、亿图图示、亿图脑图 MindMaster、万兴 PDF 等均已集成该模型。知乎：DeepSeek-R1 模型能力已经在知乎直答网页和知乎 App 双端集成上线，实现了搜索结果和解答质量的全面提升。科大讯飞：讯飞开放平台宣布 DeepSeek 全系大模型现已正式上线，支持公有云 API 调用和专属模型一键部署。

■ 政务，医疗等其他

深圳全面启用 DeepSeek 打造智慧政务新体验：2 月 16 日，深圳市基于政务云环境面向全市各区各部门正式提供 DeepSeek 模型应用服务，实现了基于 DeepSeek 的人工智能政务应用一体化赋能升级。此前，深圳已于 2 月 10 日完成 DeepSeek-R1（671B）满血版模型在政务云上的部署，并于 2 月 13 日组织开展全市使用操作培训，成为全省首个基于政务云信创环境下全市范围部署应用 DeepSeek 的城市。据“北京海淀”微信公众号 2 月 11 日消息，海淀区正式为区内企业提供 DeepSeek 全量模型服务，为辖区内行业提供大模型开发、大模型推理 API 服务、大模型应用开发服务等。

浙江省卫健委和蚂蚁集团联合推出“安诊儿”：2 月 16 日，由浙江省卫健委和蚂蚁集团联合推出的“安诊儿”宣布融合 DeepSeek-R1，升级大模型底座能力。多家医疗企业接入 DeepSeek ：2 月 6 日，智云健康宣布将 DeepSeek-R1 模型接入公司自研医疗人工智能系统“智云大脑”。2 月 7 日，医渡科技宣布，已将 DeepSeek 人工智能模型整合至公司自主研发的“AI 医疗大脑” YiduCore。2 月 7 日，鹰瞳 Airdoc 自主研发的万语医疗大模型完成焕新升级，接入 DeepSeek R1 模型。

■ 海外知名企业

微软已将 DeepSeek-R1 模型纳入其 Azure AI Foundry，这标志着 AI 经济格局的重大转变。这一整合不仅挑战了现有的 AI 服务定价，还使更多企业能够以更低的成本采用 AI 技术。英伟达支持 DeepSeek：作为国际巨头之一，英伟达已正式宣布支持 DeepSeek 模型服务。英特尔支持 DeepSeek：英特尔是另一家正式宣布支持 DeepSeek 模型服务的国际巨头。

2. 深入剖析 DeepSeek

2.1DeepSeek 为什么被称为 shock?

DeepSeek 近期在全球科技行业引起的震动，被西方媒体称为“DeepSeek Shock ‘’ 。这不仅造成了研究人员的兴奋，也引起了资本市场的高度关注。

究其原因，是因为 DeepSeek 以更低的硬件成本和更短的时间实现了可以与 OpenAI 和 Anthropic 等美国公司的尖端产品竞争的能力。具体来讲，引起轰动的产品主要是通用大模型 V3 和推理大模型 R1。

表 1：DeepSeek-V3 和 DeepSeek-R1 比较

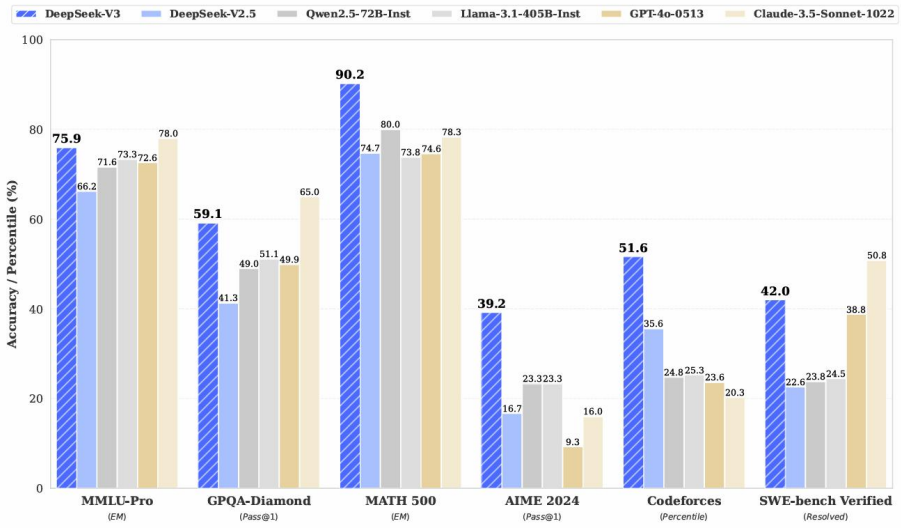
特性	DeepSeek V3	DeepSeek R1
架构	混合专家（MoE）	混合专家（MoE），优化推理能力
参数规模	671B 每次仅激活 37B 参数	671B
计算优化	大幅节省计算资源并提高响应速度	采用动态门控机制，适应推理任务
应用场景	自然语言处理 NLP	复杂逻辑推理
特色	由于其优秀的性价比，适用于实时变化的商业和研究需求	可以蒸馏出参数规模不同的开源小模型，可以迅速部署在不同应用场景的深度推理

资料来源：DeepSeek，爱建证券研究所

■ 通用大模型 DeepSeek-V3

DeepSeek V3 采用混合专家(MoE)架构，主要面向自然语言处理(NLP)任务，旨在提供高效、可扩展的解决方案。其优势在于高效的多模态处理能力(文本、图像、音频、视频)和较低的训练成本。

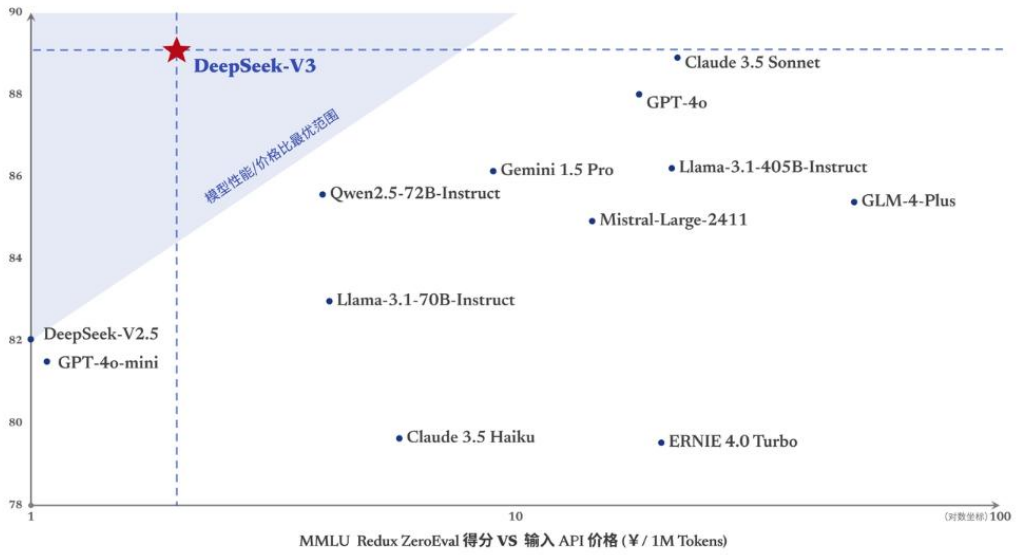
图 4：DeepSeek 通用模型 V3 性能达到行业领先水平



资料来源：DeepSeek，爱建证券研究所

DeepSeek-V3 在实现领先性能的同时，还保证了更加经济的训练成本。

图 5：DeepSeek 通用模型 V3 成本优势明显



资料来源：DeepSeek，爱建证券研究所

■ 推理大模型 DeepSeek-R1

DeepSeek-R1:专注于复杂推理任务设计，强化在数学、代码生成和逻辑推理领域的性能。通过大规模强化学习(RL)和冷启动技术，R1 在无需大量监督微调(SFT)的情况下，实现了与 OpenAI O1 系列相当的推理能力。

图 6：DeepSeek 推理模型 R1 性能达到行业领先水平

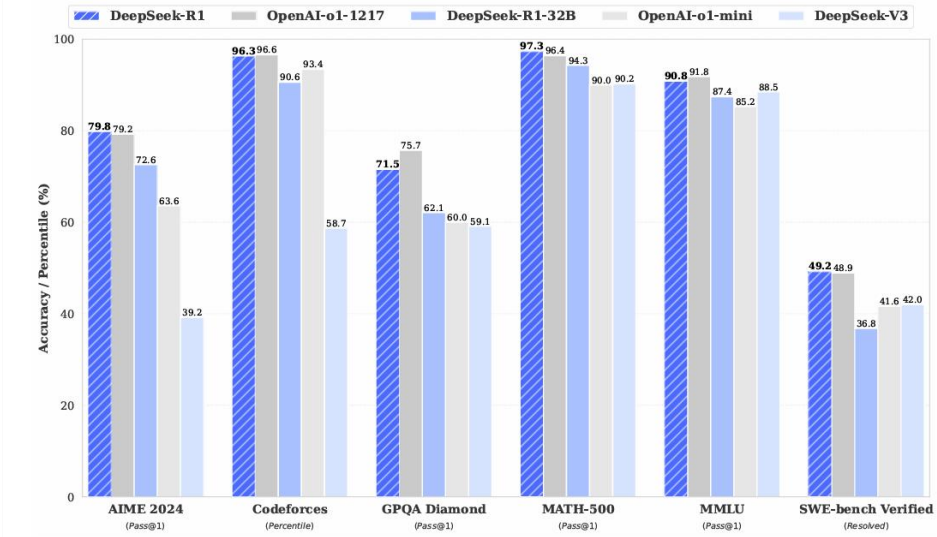


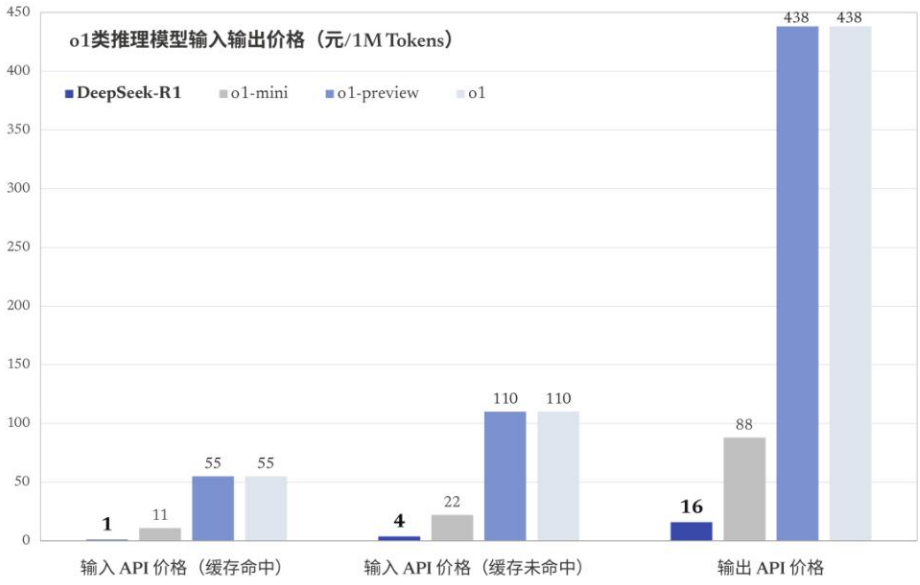
Figure 1 | Benchmark performance of DeepSeek-R1.

资料来源：DeepSeek，爱建证券研究所

DeepSeek R1 属于性能与成本平衡的中小规模模型（~7B 参数）。训练成本主要涵盖算力（如 GPU 集群）、数据准备、算法调优等核心环节。相较于

千亿参数模型的数千万甚至上亿美元成本（如 GPT-3 估算约 1200 万美元），DeepSeek R1 的设计更注重实际落地效率。这也使得 DeepSeek-R1 目前的 API 服务价格也远低于行业水平。

图 7：DeepSeek-R1 价格优势明显



资料来源：DeepSeek，爱建证券研究所

同时让行业兴奋的是，基于 DeepSeek-R1 蒸馏出的开源小模型，可以适配于不同的应用环境，这让本地化部署的浪潮迅速展开。通过参数与场景的精准匹配，DeepSeek R1 蒸馏模型可最大化性价比，覆盖从嵌入式设备到企业级服务的全场景需求。

表 2：DeepSeek-R1 蒸馏小模型本地化部署

参数规模	0.1–0.5B	0.5–1B	1–3B	3–7B
核心优势	超低功耗、毫秒级响应	平衡轻量化与基础语义理解	支持中等复杂度生成任务	接近原模型能力，高效推理
主要限制	仅支持简单任务、短文本处理	输出质量中等，推理可控性弱	长文本生成不连贯	显存需求较高（需 GPU 加速）
适用硬件	手机/嵌入式芯片（Cortex-M）	树莓派/边缘计算盒（4GB RAM）	低端 GPU 或 CPU 服务器（8GB RAM）	中端 GPU（如 T4/A10）
任务复杂度	单一分类/检测	简单生成+分类	多轮对话/摘要	长文本生成/推理
硬件成本	<$50/设备	$100–$300	$500–$2000	$3000+
开发周期	1–3 天	1–2 周	2–4 周	4–8 周
适用行业	智能家居、农业传感器	零售客服、教育工具	金融合规、电信运维	医疗、法律、营销

资料来源：DeepSeek，爱建证券研究所

2.2DeepSeek 的技术创新在哪里?

DeepSeek 仍然是基于 Transformer 框架下的大语言模型，这一点和 OpenAI 等其他大模型并无二致。但是 DeepSeek 在模型训练和推理过程中实现了开创性的创新。**DeepSeek 的创新主要在于 MoE 和 MLA 两个技术突破。**

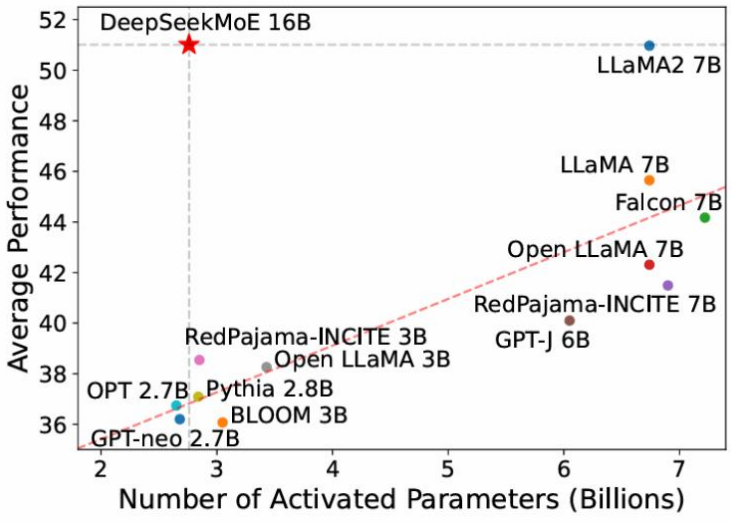
表 3：DeepSeekMoE 的效率创新

技术维度	传统 MoE	DeepSeekMoE	效率提升倍数
单 token 计算量	激活 20%总参数	激活 5.5%总参数	3.6×
设备间通信量	无限制跨节点路由	最多 4 节点路由	2-5×
KV 缓存/Token	384KB（Llama 100B）	135KB	2.8×
训练吞吐量	1.2 样本/秒/GPU（基线）	4.3 样本/秒/GPU	3.6×

资料来源：DeepSeek，爱建证券研究所

- **DeepSeekMoE**：2024 年 1 月，DeepSeek 发表论文***《DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models》***。论文提出了 DeepSeekMoE，这是一种创新的 MoE（混合专家）架构，专门设计用于实现终极专家专业化。论文中介绍，DeepSeek 通过降低激活参数比例，实现了训练效率 3.6X 的提升和训练吞吐量 3.6X 的提升。

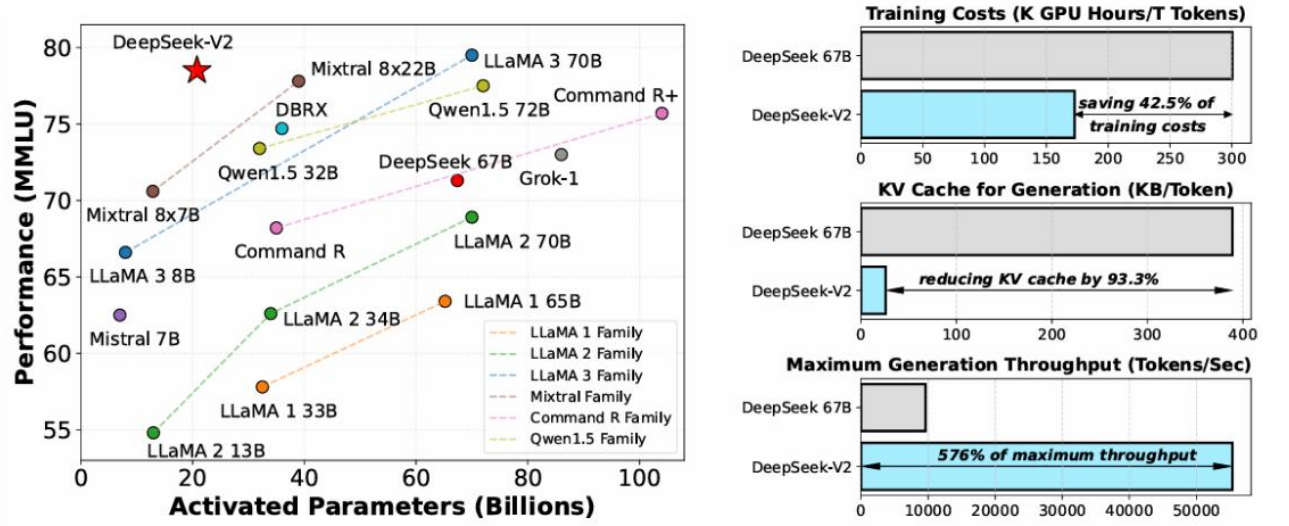
图 8：DeepSeekMOE 架构可以用更少的训练参数实现更好的性能表现



资料来源：DeepSeek，爱建证券研究所

- **MLA**：2024 年 5 月，DeepSeek 发布论文***《DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model》***。论文在 DeepSeek-V2 模型中，引入了 MLA（多头潜在注意力机制），MLA 通过将 KV 缓存显著压缩为潜向量来保证高效推理。

图 9：DeepSeek-V2 通过 MLA 机制实现了性能优势和成本降低



资料来源：DeepSeek，爱建证券研究所

与 DeepSeek 67B 相比，DeepSeek-V2 实现了显著增强的性能，同时节省了 42.5% 的训练成本、减少了 93.3% 的 KV 缓存、并将最大生成吞吐量提升至 5.76 倍。

除此之外，DeepSeek 在训练时利用了多个 token 预测的技术，即在预测下一个 token 时会基于之前所有 token 的上下文，这样在后续预测时可以减少部分计算量，加快训练速度。此外，DeepSeek-V3 还支持动态的 FP8 混合精度，在底层运算中使用 FP8 浮点数，大幅减少了计算量。同时，它对 GPU 底层部署进行了优化，提升了 GPU 通讯效率，节省了时间。

2.3 技术细节之外的信心提升

DeepSeek 在开源社区和行业中引起了轰动，主要原因在于其采用纯粹的强化学习方式训练模型，并取得了显著的效果，这颠覆了之前认为强化学习训练复杂且难以实现高效模型的认知。

除了在技术细节的因素，其之所以能够称为“Shock”，主要在于打破了行业***“限制中国企业对于最先进 GPU 获取，将能够阻止中国 AI 技术发展”***的一贯认知。

- **AI 行业的 Scaling Law**

Scaling Law（规模定律） 是深度学习领域的重要规律，它描述了 AI 模型性能（如准确率或生成质量）与模型规模、训练数据量及计算资源（如算力）之间的数学关系。**简单来说，“模型变大、数据变多、算力变强，效果会系统性提升”。**这使得所有人产生了一种算力军备竞赛的恐慌，所有的行

业参与者都会不遗余力的投资算力，行业内普遍认为算力投资需求看不到上线，同时 AI 性能提升与算力投资成线性或者指数关系。

图 10：Scaling Law 中算力与 AI 能力的关系

资料来源：NVIDIA，爱建证券研究所

与此同时，早在 2022 年 10 月：拜登政府就首次出台限制措施，明确针对高性能计算芯片 A100/H100 及其制造设备。2023 年 10 月：拜登政府升级管制，将英伟达 A800/H800 等降级版 GPU 纳入禁售范围，并扩大限制国别。

表 4：国内外高端 GPU 芯片性能比较

芯片型号	NVIDIA A100	NVIDIA H100	NVIDIA A800	NVIDIA H800	华为昇腾 910B	寒武纪 MLU370
FP32 算力 (TFLOPS)	19.5	42.8	19.5	30.4	~12	~8
FP16 算力 (TFLOPS)	312	2000	312	1450	256	128
INT8 算力 (TOPS)	1248	4000	1248	2900	1024	512
显存容量	40/80GB	80GB	40/80GB	80GB	32GB	16GB
显存带宽 （GB/s）	1935 (HBM2e)	3350 (HBM3)	1935	2500 (HBM2e)	~900 (HBM2)	~600
单卡互连带宽	600GB/s	600GB/s	400GB/s	400GB/s	200GB/s	150GB/s
集群扩展性	支持数千节点无缝扩展	支持数千节点无缝扩展	扩展效率下降约 30%	扩展效率下降约 30%	最多支持 4096 节点	实验室环境最大 64 节点

资料来源：NVIDIA，华为，寒武纪，爱建证券研究所

根据上表中数据，国产 GPU 芯片在性能方面相较于 NVIDIA 产品在性能上有显著的差距。根据 Scaling Law，这将会直接导致中国大模型研发进度延后，与国外企业的差距也将越来越大。例如，百度、腾讯、阿里巴巴等企业在训练大模型时，因无法获得英伟达 A100/H100 GPU，改用降级版 A800 或国产芯片，导致训练时间增加 30%-50%。部分企业被迫通过“堆量”弥补单卡性能差距，显著推高成本。

■ DeepSeek 的成功打破了中国 AI 产业的 Scaling Law 困境

DeepSeek 成立于 2023 年，虽然受制于高端 GPU 的限制，但是在极端的时间内就开发出性能领先且具备成本优势的国产大模型。

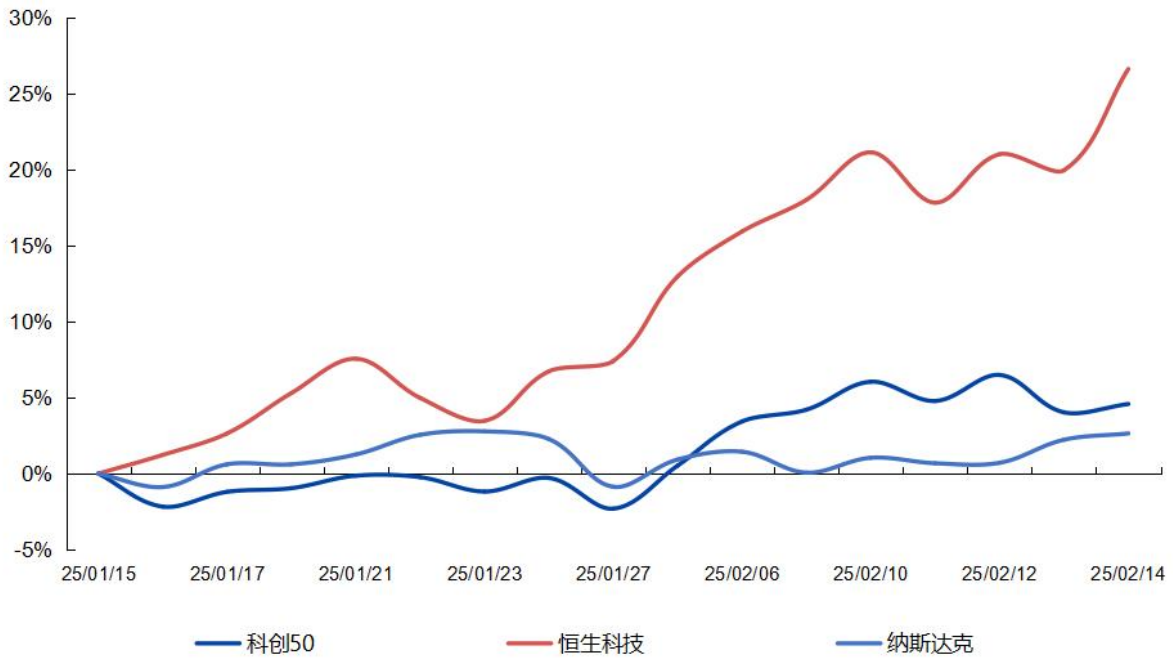
表 5：DeepSeek 开发具备成本优势

模型	参数量	官方成本估算	备注
Meta Llama 2-13B	13B	600 万–800 万美元	开源模型，基线参考
Mistral-8x7B	45B (MoE)	约 1200 万美元	MoE 架构降低激活参数量
Gemma-7B	7B	400 万–600 万美元	Google 高效调优技术
DeepSeek-R1	7B	550 万美元	前期版本对比基准

资料来源：DeepSeek，爱建证券研究所

总而言之，DeepSeek 通过微创新实现了低成本高性能的模型开发。同时在高端芯片受限制的情况下，用短短两年时间实现了对国外先进企业的追赶，这从根本上提振了国内企业对于 AI 产业的发展信心。

图 11：DeepSeek 带动中国资本市场信心提升



资料来源：iFind，爱建证券研究所

3. 市场对于 DeepSeek 最关心的问题

关于 DeepSeek 的讨论很多，主要围绕在 DeepSeek 实际开发成本，以及未来 GPU 算力投资是否会大幅减少等等。下面，我们将会对这些问题进行详细分析。

3.1DeepSeek-V3 的模型实际成本几何?

根据 DeepSeek 官方口径，DeepSeek-V3 模型总参数量达到 671B，每次处理激活 37B 参数，训练数据规模达 14.8T token。综合评估显示，模型维持了极具竞争力的训练成本，完整训练过程（包括预训练、上下文长度扩展和后训练）仅需 2.788M H800 GPU 小时，按照 2 美元/小时成本计算，训练成本约为 558 万美元。

表 6：DeepSeek-V3 正式训练阶段成本拆分

阶段	GPU 小时消耗	占比	成本（美元）	核心目标
预训练	2,664,000	95.5%	532.8 万	基于 14.8T tokens 完成模型基础能力构建，采用 FP8 混合精度与 MoE 架构优化效率
上下文扩展	119,000	4.3%	23.8 万	将上下文长度从 32K 扩展至 128K，优化长文本处理能力
后训练	5,000	0.2%	1.0 万	通过 SFT（监督微调）和 RL（强化学习）提升对齐能力，整合 R1 的推理知识蒸馏
总计	2,788,000	100%	557.6 万	训练周期约 55 天（2048 卡集群）

资料来源：DeepSeek，爱建证券研究所

DeepSeek-V3 的训练成本计算可分为**正式训练阶段**与**隐性成本**两个部分：

预训练效率：每 1T tokens 消耗 18 万 GPU 小时，仅为 Llama 3 同规模模型的 1/10。硬件利用率：通过 DualPipe 流水线并行与通信优化，H800 集群有效利用率达 92%，领先传统架构水平的 60-70%。

表 7：DeepSeek-V3 隐性成本拆分

成本类型	估算范围	说明
前期研发投入	2000-3000 万美元	包含 DeepSeek-V2、R1 等前置模型的研发费用，以及架构验证实验成本
数据成本	未披露	14.8T tokens 的高质量数据清洗与标注（数学/代码数据占比超 30%）
硬件折旧	未计入	若自建 H800 集群（单价约 3.5 万美元/卡），2048 卡硬件成本约 7.168 亿美元
失败实验	约 500 万美元	技术报告中提及的“架构研究”阶段可能消耗的试错成本

资料来源：DeepSeek，爱建证券研究所

DeepSeek-V3 是建立在前期 V2 等模型基础上开发的，前期研发投入约为 2000-3000 万美元。其他数据获取成本和硬件折旧成本未披露，实验试错成本约为 500 万美元，因此预计实际总成本超过 4000 万美元。

表 8：DeepSeek-V3 实际成本对比

模型	参数规模	实际成本	单位 token 成本（$/T）	成本效能比（DeepSeek=1）
DeepSeek-V3	671B	>4000 万$	260	1.0
Llama 3-405B	405B	约 2.5 亿$	833	3.2
GPT-4o	~1.8T	推测>1 亿$	>5400	>21

资料来源：DeepSeek，Meta，OpenAI，爱建证券研究所

如果按照实际成本计算，DeepSeek 的成本水平相对 Llama 3-405B 降低了约 69%；相对于 GPT-4o 降低了 95%。

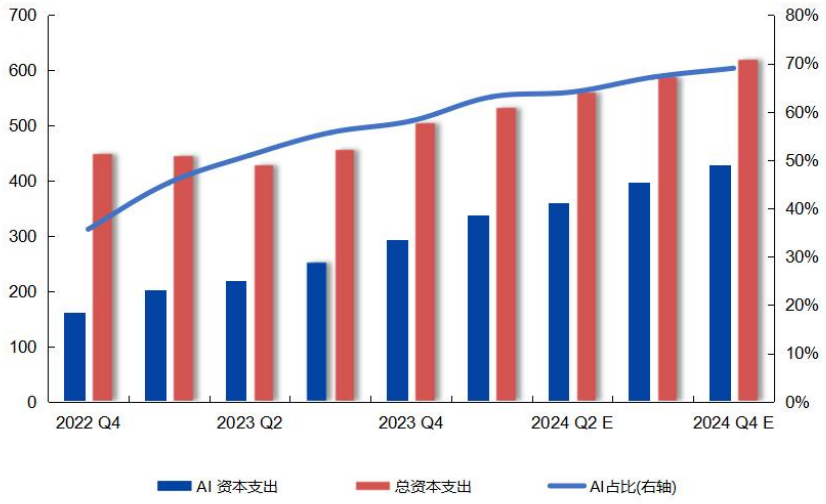
3.2GPU 为代表的算力需求是否会大幅下降?

由于 DeepSeek 模型训练成本的快速下降，直接导致了全球资本市场 GPU 行业的短暂恐慌，市场开始担心未来对于 GPU 等算力硬件的需求被过度放大。我们认为短期内由于各家 AI 巨头公司在大模型领域的“军备竞赛”不会停止，所以行业内算力支出金额的增长仍然难以看到放缓的趋势。

■ 算力军备竞赛持续加剧

从 2022 年 Q4 到 2024 年，我们可以看到美国五大云服务提供商的 AI 投资占比持续增长。2022 年 AI 支出占比普遍在 20%-40%，至 2024 年提升至 50%-85%。此前，市场一致认为对于算力的需求没有上限，这导致全市场都对于算力产生了一种“恐慌性”需求。那算力需求真的会一直线性甚至指数级持续提升吗?

图 12：美国主要互联网企业资本支出金额（亿美元）

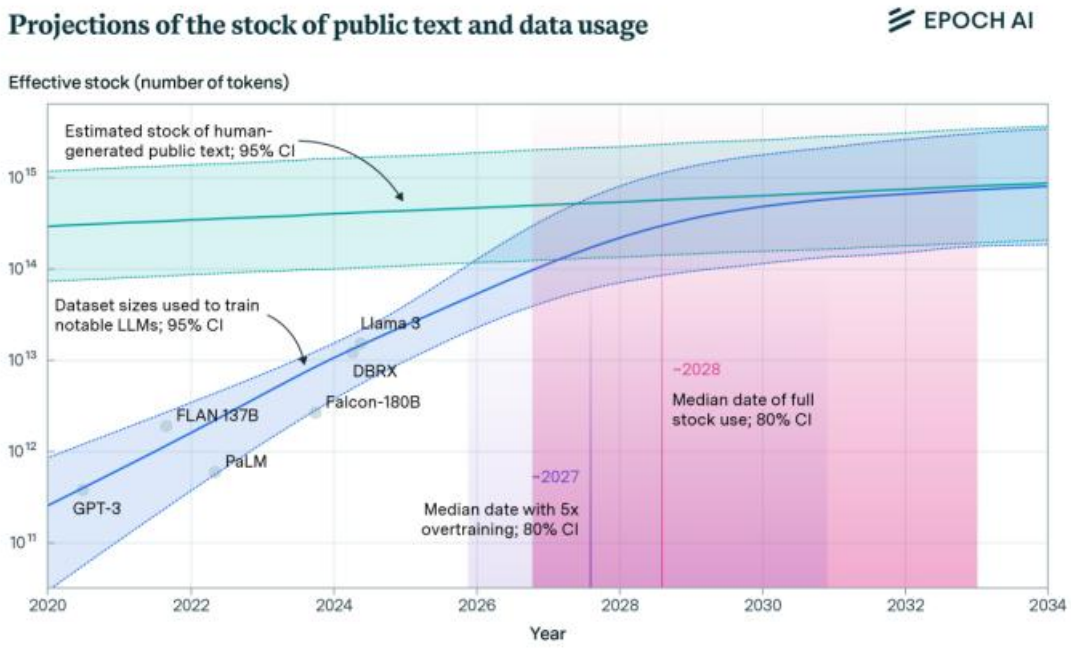


资料来源：Alphabet，Microsoft，Meta，Amazon，Apple，爱建证券研究所

■ **从算力瓶颈到数据瓶颈**

根据此前提到的 Scaling Law，AI 模型性能（如准确率或生成质量）与模型规模、训练数据量及计算资源（如算力）之间成正相关关系。虽然目前提升算力确实能够推动 AI 能力的升级，但是这一切的前提是我们拥有充足的高质量数据投喂到训练过程中。

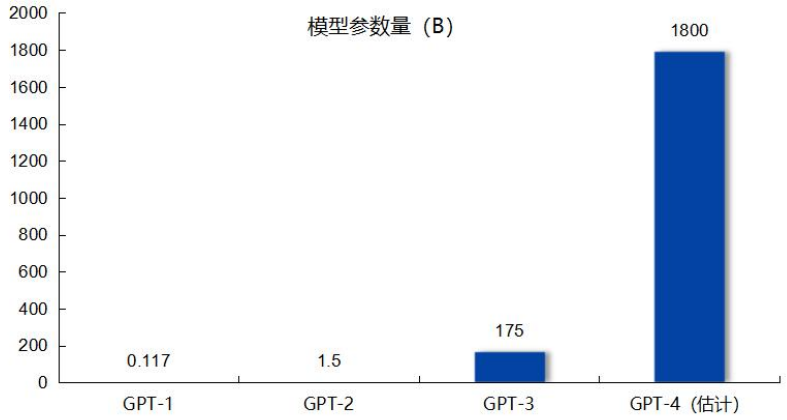
图 13：人类生成的公共文本数据将在 2028 年耗尽



资料来源：Alphabet，Microsoft，Meta，Amazon，Apple，爱建证券研究所

EPOCH AI 在 2024 年 6 月发布论文***《Will we run out of data? Limits of LLM scaling based on human-generated data》***，**人类生成的公共文本数据总有效库存量约为 300T token（90%可能性在 100T 至 1000T 之间）。按照目前大模型算力能力进化的趋势来看，2027-2028 年左右现有的大模型训练就将会耗尽所有的公开文本数据。**

图 14：OpenAI 模型参数规模快速靠近公共文本数据上限



资料来源：OpenAI，爱建证券研究所

而根据目前 OpenAI 最新的模型 GPT4 参数规模的预估值 1.8T，以及 OpenAI 每一代升级后至少一个数量级上升的速度。我们认为，在未来 2-3 代模型升级时间之内，影响 AI 模型性能升级的关键将不再是算力投资。数据资源正在成为模型进化的重要瓶颈。

- **数据瓶颈之后的发展方向在哪里**

虽然公开的文本数据正在以肉眼可见的速度被耗尽，但是仍然有大量的高价值专业数据，以及实时动态数据存在于各个行业公司，政府部门以及媒体领域。我们认为，未来 AI 模型的进化方向除了继续在算力硬件方面投资，更重要的是在专业细分领域利用数据资源优势实现具备差异化优势的细分模型。

表 9：主要细分行业模型参数规模

专业领域	典型数据集大小（Tokens）	常用模型参数量（B）	关键训练技术
医疗	10 亿–100 亿	1.5B–540B	领域自适应微调（Adapter Layers）
金融	50 亿–500 亿	0.34B–50B	时序数据编码（Temporal Attention）
法律	30 亿–200 亿	1.3B–175B	长文本分块训练（Doc-level NSP）
科学	1 亿–10 亿（结构化数据为主）	0.2B–7B	多模态融合（文本+分子图）
动态	实时流数据（持续更新）	10B–50B	在线学习（Real-time Embedding）

资料来源：OpenAI，爱建证券研究所

根据上表中的数据，目前最先进的模型（包括 DeepSeek-R1）已经足够完成训练重点行业的数据集。因此，我们可以相信目前的算力水平已经足够完成细分行业的模型有效训练。也就是说在现有模型框架下，模型训练的算力需求并不会持续的以线性或者指数级方式增长，未来的算力需求主要来自于细分领域的推理需求。

4. DeepSeek 对产业发展的影响

随着 DeepSeek 的横空出世，低成本高性能的模型训练部署称为可能，这对于 AI 应用加速落地产生了极大地推动作用。我们预计接入 DeepSeek API 的细分领域推理服务商将会快速涌现，同时具备足够算力的终端设备将会能够实现蒸馏小模型的本地部署能力，包括服务器，智能手机和智能驾驶系统都将会迎来新一轮的升级浪潮。具体来看，我们看好以下细分板块在 DeepSeek 推动下的未来发展：

1、先进算力芯片制造产业链：算力芯片设计，先进制程晶圆代工，先进封测。

2、专业咨询服务：包括 AI+医疗，法律，金融，会计，办公，教育，政务等等咨询服务领域。

3、2C 智能终端：AI+智能手机，智驾汽车等终端产业链。

4、2B 本地部署设备：服务器 OEM 及产业链。

5. 风险提示

大模型开发进度缓慢，模型升级不及预期。

AI 行业应用落地迟缓，商业模式难以实现良性循环。

贸易摩擦加剧，先进芯片及半导体技术受限。

爱建证券有限责任公司
上海市浦东新区前滩大道 199 弄 5 号
电话：021-32229888　　传真：021-68728700　　服务热线：956021
邮政编码：200124　　邮箱：ajzq@ajzq.com　　网址：http://www.ajzq.com

评级说明

投资建议的评级标准

报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 个月内的相对市场表现，也即：以报告发布日后的 6 个月内的公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A 股市场：沪深 300 指数（000300.SH）；新三板市场：三板成指（899001.CSI）（针对协议转让标的）或三板做市指数（899002.CSI）（针对做市转让标的）；北交所市场：北证 50 指数（899050. BJ）；香港市场：恒生指数（HIS.HI）；美国市场：标普 500 指数（SPX.GI）或纳斯达克指数（IXIC.GI）。

股票评级

买入	相对同期相关证券市场代表性指数涨幅大于 15%
增持	相对同期相关证券市场代表性指数涨幅在 5%～15%之间
持有	相对同期相关证券市场代表性指数涨幅在-5%～5%之间
卖出	相对同期相关证券市场代表性指数涨幅小于-5%

行业评级

强于大市	相对表现优于同期相关证券市场代表性指数
中性	相对表现与同期相关证券市场代表性指数持平
弱于大市	相对表现弱于同期相关证券市场代表性指数

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告采用信息和数据来自公开、合规渠道，所表述的观点均准确地反映了我们对标的证券和发行人的独立看法。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法可能存在局限性，请谨慎参考。

法律主体声明

本报告由爱建证券有限责任公司（以下统称为“爱建证券”）证券研究所制作，爱建证券具备中国证监会批复的证券投资咨询业务资格，接受中国证监会监管。

本报告是机密的，仅供我们的签约客户使用，爱建证券不因收件人收到本报告而视其为爱建证券的签约客户。本报告中的信息均来源于我们认为可靠的已公开资料，但爱建证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供签约客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，爱建证券及其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测后续可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，爱建证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

版权声明