中国宏观经济月报

2025/02/11

中国宏观经济研究员 **杨曦**
010-66555831
xi_yang@chiefgroup.com.hk

DeepSeek 对英伟达长期股价的潜在影响

1 月 27 日，DeepSeek 在中国区和美国区的苹果 App Store 免费榜上同时登顶，成为下载量第一的应用程序。与此同时，美国科技股市场却遭遇了大幅下跌，费城半导体指数（SOX）下跌了 9.2%，创下自 2020 年 3 月以来的最大单日跌幅。其中，英伟达股价下跌了近 17%，市值蒸发了近 6000 亿美元，成为美股历史上最大规模的单日市值缩水之一。甚至 WTI 原油价格也在盘中一度下跌了 3%。一些交易员认为，如果大模型的训练和推理不再需要大量算力，数据中心的电力需求也会随之减少，进而减少对石油发电的依赖。

这场科技股的大幅波动，主要归因于 DeepSeek 在训练和推理成本上的显著优势。

推理成本（API 报价）：每百万 Token 的输入成本仅为 1 元。

模型[1]	上下文长度	最大思维链长度[2]	最大输出长度[3]	百万tokens 输入价格 (缓存命中)[4]	百万tokens 输入价格 (缓存未命中)	百万tokens 输出价格 输出价格
deepseek-chat	64K	-	8K	~~0.5元~~[5] 0.1元	~~2元~~[5] 1元	~~8元~~[5] 2元
deepseek-reasoner	64K	32K	8K	1元	4元	16元[6]

1. `deepseek-chat` 模型已经升级为 **DeepSeek-V3**；`deepseek-reasoner` 模型为新模型 **DeepSeek-R1**。

ERNIE 4.0	ERNIE-4.0-8K ERNIE-4.0-8K-0613 ERNIE-4.0-8K-Latest ERNIE-4.0-8K-Preview	推理服务	输入	0.03元/千tokens
			输出	0.09元/千tokens
		搜索增强	触发	0.008元/次
	ERNIE-4.0-8K-Latest	推理服务	输入	0.03元/千tokens
			输出	0.03元/千tokens
		搜索增强	触发	0.008元/次

训练成本：根据 DeepSeek 发布的技术报告，他们共使用了约 2000 张 H800 GPU 进行训练，整个 V3 模型的训练成本不超过 600 万美元。

1、在预训练阶段，每万亿 Token 的训练使用了 2048 个 H800 GPU 集群，仅需 180K 个 GPU 小时，大约 3.7 天即可完成。

2、整个预训练过程总计耗时 2664K GPU 小时（不到两个月），加上上下文扩展和后训练，总耗时约为 2788K GPU 小时。

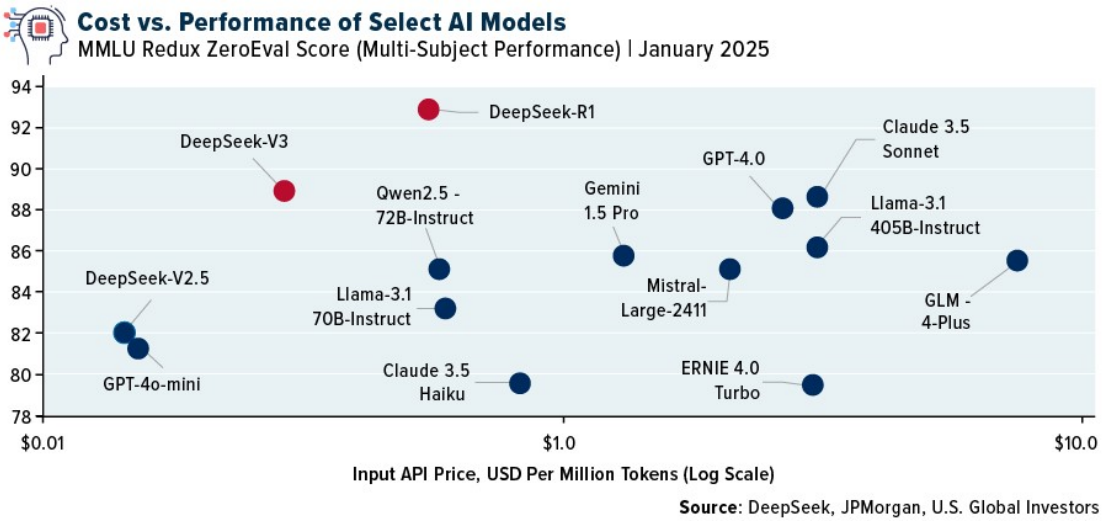
3、按照 H800 每小时 2 美元的租赁价格，总训练成本不超过 600 万美元。

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is $2 per GPU hour.

来源：DeepSeek-V3 Technical Report

在推理成本方面，OpenAI 的 o1 模型每百万输入和输出 Token 分别收费 15 美元和 60 美元，而 DeepSeek 的 R1 模型在相同输入和输出下的价格仅为 OpenAI 的 3%。



DeepSeek 是如何实现低成本训练的？

DeepSeek 团队通过创新的训练策略显著降低了成本，尤其是在监督微调（SFT）环节进行了优化。他们最初尝试完全跳过 SFT 步骤，仅通过强化学习（RL）进行训练，推出了名为 DeepSeek-R1-Zero 的模型版本。尽管这种方法在初期需要更多的计算资源（因为模型需要更多的探索），但研究人员发现，通过引入少量冷启动数据，可以显著提升训练的稳定性并增强模型的推理能力。

在 R1 系列模型之前，业界普遍采用 RLHF（基于人类反馈的强化学习）方法，依赖大量由人类撰写的高质量问答数据，以帮助模型在奖励信号不明确的情况下做出复杂决策。而 R1 系列模型则摒弃了 RLHF 中的人类反馈（HF）部分，仅保留了纯粹的强化学习（RL）机制。DeepSeek 团队表示，经过数千次“纯强化学习”训练步骤后，DeepSeek-R1-Zero 在推理基准测试中的表现已经与 OpenAI 的 o1-0912 模型相当。

然而，纯强化学习训练存在一个显著问题：模型过度关注答案的正确性，而忽视了语言流畅性等基础能力，导致生成的文本出现了中英混杂的现象。为了解决这一问题，DeepSeek 团队使用数千条链式思考（CoT）数据对 V3-Base 模型进行微调。这些数据包含了规范的语言表达和多步推理示例，帮助模型初步掌握了逻辑连贯的生成能力。随后，团队启动了强化学习流程，生成了约 60 万个与推理相关的样本和 20 万个与推理无关的样本。最终，这 80 万个样本数据被用于微调 V3-Base 模型，从而得到了 R1 模型。

通过这种创新的训练方案，DeepSeek 不仅大幅降低了训练成本，还成功提升了模型的推理能力和语言生成质量。

DeepSeek 的成果将如何影响 AI 产业的未来？

DeepSeek 的突破性成果将对 AI 产业的未来产生深远影响。尽管英伟达的股价因此受到了一定冲

击，但从长远来看，DeepSeek 的最大影响可能并非直接作用于英伟达，而是对那些依赖自研大模型并通过模型调用构建商业模式的公司，如 OpenAI、Anthropic、月之暗面、字节跳动等，产生更为显著的影响。

据 Meta 员工在匿名平台透露，DeepSeek 仅以 1%的成本投入便实现了超越 Llama 3 的性能表现，这一成就已引发公司内部 AI 团队的担忧，尤其是考虑到他们正在研发的下一代模型 Llama 4 的预期投入将比 Llama 3 高出数倍。技术媒体 The Information 随后报道称，Meta 已成立四个专门的研究小组，深入分析 DeepSeek 的技术原理，并计划将其应用于 Llama 模型的优化中。在 DeepSeek V3 发布之前，Llama 曾是全球最强大的开源模型，而 V3 的推出无疑对其地位构成了挑战。

然而，对于美国的大型科技企业而言，保持技术领先地位仍是首要目标，而非单纯的成本优化。这些公司不仅致力于在行业内保持领先，还希望在全球范围内维持技术优势，尤其是在面对中国等国家的竞争时。尽管这些巨头可能会借鉴 DeepSeek 的方法来优化部分成本，但这并不会成为他们的核心战略方向。

现阶段，各大公司全力投入研发的大语言模型（LLM），依旧处于蓬勃发展的上升阶段，这需要持续投入大量算力支持。即使未来 LLM 的算力需求趋于饱和，其他形式的机器学习（ML）模型仍可能对算力有巨大需求。算力如同能源，几乎无处不在。DeepSeek 若想进一步提升性能，也需要更多 GPU 资源来支持其训练，以满足日益增长的计算需求。

一直以来，中文大模型的发展相较于其他语言的大模型，明显处于滞后状态，主要原因之一是高端芯片的短缺。DeepSeek 创始人梁文峰在接受《暗涌》采访时也强调了这一问题，直言 Deepseek “面临的问题从来不是钱，而是高端芯片被禁运”。在未来相当长的一段时间内，英伟达的高端芯片仍将是 AI 科技企业的核心需求。

英伟达在一份声明中指出，DeepSeek 的成果实际上证明了市场对英伟达芯片的需求将会增加，而非减少。这一观点有其合理性，因为随着模型训练和推理成本的降低，人工智能的商业化进程将加速，从而推动对算力的需求增长。

依据经济学中著名的杰文斯悖论，技术的进步虽说能够有效提高资源的使用效率，使得单个应用在运行过程中所需的资源量显著减少。然而，令人意想不到的是，随着资源使用成本的持续下降，市场对于资源的总体需求不但没有如预期般减少，反而会逆势上扬，导致总的资源消耗量不降反升。

DeepSeek 的成功显著降低了 LLM 的开发门槛，这将促使更多中小型企业甚至个人开始训练私有模型。如果这一趋势能够引发中小型企业、家庭和个人用户对推理需求的“第二波”增长，那么这些增量需求将远超 AI 巨头减少的 GPU 采购量。回首往昔，类似的场景也曾上演，当年个人电脑价格从令人咋舌的 2 万美元一路暴跌至亲民的 2 千美元后，微软公司便紧紧抓住这一历史机遇，迎来了发展的黄金时期。。而且，值得注意的是，相较于模型训练环节，在商业化之后的推理环节，对于算力的消耗将会更为惊人。更何况，即便研发出了更高效的计算方法，也并不意味着算力需求就会随之减

少，二者并非简单的线性关系。

总体而言，DeepSeek 开创的低成本模式，虽说在短期内确实对英伟达的股价造成了一定冲击，让市场泛起一阵涟漪。但如若我们站在更为宏观的长远视角审视，随着 AI 技术的日益普及，以及商业化进程的加速前行，英伟达的芯片需求极有可能会迎来进一步显著增长，届时，AI 产业又将迈向一个全新的发展阶段。

研究员声明

主要负责撰写本研究报告全部或部分内容的研究分析员在此声明：（1）该研究员以勤勉的职业态度，独立、客观地出具本报告，本报告清晰准确地反映了该研究员的研究观点；（2）该研究员所得报酬的任何组成部分无论是在过去、现在还是将来均不会直接或间接地与研究报告所表述的具体建议或观点相联系。

一般声明

本报告由香港致富证券有限公司（以下简称“致富证券”）制作，报告中的信息均来源于公开资料，本公司对这些信息的准确性和完整性不作任何保证。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价和征价。本公司及本公司员工对使用本报告及其内容所引发的任何直接或间接损失概不负责。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。

市场有风险，投资需谨慎。投资者不应将本报告为作出投资决策的惟一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向本公司或其他专业人士咨询并谨慎决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本报告内的所有意见均可在不作另行通知之下作出更改。本报告的作用纯粹为提供信息，并不应视为对本报告内提及的任何产品买卖或交易的专业推介、建议、邀请或要约。

香港	北京	上海	深圳
香港致富证券有限公司	**香港致富证券有限公司北京代表处**	**香港致富证券有限公司上海代表处**	**香港致富证券有限公司深圳代表处**
香港德辅道中 308 号 富卫金融中心 11 楼	北京市朝阳区国贸写字楼 1 座 6 层	上海市陆家嘴东路 161 号招商局大厦 1309 室	深圳福田区福华路 399 号中海大厦 6 楼
电话：（852）25009228	电话：（8610）66555862	电话：（8621）38870772	电话：（86755）33339666
传真：（852）25216893	传真：（8610）66555831	传真：（8621）58799185	传真：（86755）33339665

网址：www.chiefgroup.com.hk