

更多AI工具可直接访问：<https://www.faxianai.com/>

Deepseek满血版入口：<https://www.faxianai.com/ai/6039.html>

SemiAnalysis报告：对这次DeepSeek事件的分析

整理者：AI君主小亦

The screenshot shows the top portion of a SemiAnalysis article. The navigation bar includes links for Menu, Institutional, Tools, About, and Archive, along with a search icon, Log in, and Subscribe buttons. The article title is "DeepSeek Debates: Chinese Leadership On Cost, True Training Cost, Closed Model Margin Impacts // H100 Pricing Soaring, Subsidized Inference Pricing, Export Controls, MLA". It is dated January 31, 2025, and is 15 minutes long with 8 comments. The authors are Dylan Patel, AJ Kourabi, Doug O'Laughlin, and Reyk Knuhtsen. The main image depicts a large blue whale breaching a city, causing destruction. Below the image is a "Table of Contents" section with links to "The DeepSeek Narrative Takes the World by Storm", "DeepSeek and High-Flyer", "The GPU Situation", and "DeepSeek's Cost and". The main heading is "The DeepSeek Narrative Takes the World by Storm". A cookie consent banner is visible at the bottom right.

原文地址：<https://semianalysis.com/2025/01/31/deepseek-debates/>

深度求索的故事席卷全球

深度求索（DeepSeek）的故事在全球引起了轰动。在过去的一周里，深度求索成了全球各界唯一的热议话题。目前，深度求索的日访问量远超 Claude、Perplexity，甚至超过了 Gemini。

但对于密切关注这一领域的人来说，这其实并非什么“新鲜事”，令人瞩目的是人们对它的疯狂炒作。长期以来，SemiAnalysis 一直认为深度求索极具天赋，但美国大众此前并不关注。当全世界终于开始关注时，却陷入了一种脱离现实的疯狂炒作。几个月来，我们一直在谈论深度求索（每个链接都是例证）。这家公司并不新。

我们想强调的是，舆论风向与上个月相比发生了逆转。上个月，当规模定律被打破时，有人认为这对英伟达（Nvidia）和 GPU 不利；如今，又有人说算法改进速度过快。我们已经破除了这些谬论。

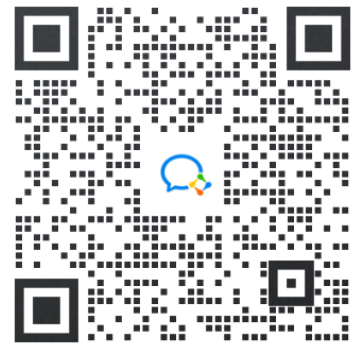
免责声明：

1. 本资料来源于网络公开渠道，版权归属版权方；
2. 本资料仅限会员学习使用，如他用请联系版权方；
3. 会员费用作为信息收集整理及运营之必须费用；
4. 如侵犯您的合法权益，请联系客服微信将及时删除



行业报告资源群

1. 进群福利：进群即领万份行业研究、管理方案及其他学习资源，直接打包下载
2. 每日分享：6份行研精选报告、3个行业主题
3. 报告查找：群里直接咨询，免费协助查找
4. 严禁广告：仅限行业报告交流，禁止一切无关信息



微信扫码，长期有效

更多AI工具可直接访问：<https://www.faxianai.com/>

Deepseek满血版入口：<https://www.faxianai.com/ai/6039.html>

知识星球 行业与管理资源

专业知识社群：每月分享8000+份行业研究报告、商业计划、市场研究、企业运营及咨询管理方案等，涵盖科技、金融、教育、互联网、房地产、生物制药、医疗健康等；已成为投资、产业研究、企业运营、价值传播等工作助手。



微信扫码，行研无忧

如今的舆论认为，深度求索效率极高，我们不再需要更多计算资源，而且由于模型的变化，现在一切都存在大量过剩产能。虽然杰文斯悖论（Jevons paradox）也被过度炒作了，但它更接近现实情况。这些模型已经对 H100 和 H200 的定价产生了实际影响，刺激了需求。

深度求索与 High-Flyer（幻方）

High-Flyer 是一家中国对冲基金，也是将人工智能应用于交易算法的早期 adopters。他们很早就意识到了人工智能在金融领域之外的潜力，以及规模化的关键意义。因此，他们不断增加 GPU 的储备。在使用数千个 GPU 集群对模型进行试验后，High-Flyer 在 2021 年出口限制实施前投资购买了 10000 个 A100 GPU，这一举措取得了回报。随着 High-Flyer 的发展，他们在 2023 年 5 月决定分拆出“深度求索”，目标是更专注地追求人工智能能力的进一步提升。当时，由于缺乏商业模式，外部投资者对人工智能兴趣寥寥，High-Flyer 便自行出资成立了这家公司。如今，High-Flyer 和深度求索经常共享人力和计算资源。

深度求索如今已发展成为一项认真且协同的事业，绝非许多媒体声称的“副业”。我们确信，即便考虑到出口管制因素，他们在 GPU 上的投资也超过 5 亿美元。

DeepSeek AI TCO						
	Unit	A100	H20	H800	H100	Total
Years	#	4	4	4	4	
# of GPUs	#	10,000	30,000	10,000	10,000	60,000
NVDA \$ ASP	\$	\$13,500	\$12,500	\$20,000	\$23,000	
Server CapEx / GPU	\$	\$23,716	\$24,228	\$31,728	\$34,728	
Total Server CapEx	\$m	\$237	\$727	\$317	\$347	\$1,629
Cost to Operation	\$m	\$157	\$387	\$170	\$230	\$944
Total TCO (4y Ownership)	\$m	\$395	\$1,114	\$487	\$577	\$2,573

Note: TCO assumes server capital costs are amortized over 4 years at a 13.3% WACC
Note: NVDA \$ ASP only attributable to NVDA

GPU 情况

我们认为他们拥有约 50000 个 Hopper GPU，但这并不等同于 50000 个 H100，一些人存在这样的误解。英伟达为遵守不同规定，生产了 H100 的多种变体（H800、H20），目前中国的模型供应商仅能获得 H20。需要注意的是，H800 的计算能力与 H100 相同，但网络带宽较低。

我们认为深度求索拥有约 10000 个 H800 和 10000 个 H100。此外，他们还订购了更多 H20。在过去 9 个月里，英伟达生产了超过 100 万个专供中国的 GPU。这些 GPU 由 High-Flyer 和深度求索共享，并在一定程度上进行了地理分布。它们被用于交易、推理、训练和研究。如需更具体的详细分析，请参考我们的《加速器模型》。

深度求索人工智能的总拥有成本

我们的分析显示，深度求索的服务器总资本支出接近 13 亿美元，运营这些集群的成本高达 7.15 亿美元。同样，所有人工智能实验室和超大规模数据中心为了各种任务（包括研究和训练），拥有的 GPU

数量比单次训练所需的更多，因为资源集中存在一定挑战。X.AI 作为一个人工智能实验室比较独特，它所有的 GPU 都集中在一个地方。

深度求索只从中国招聘人才，不看重过往资历，高度关注能力和求知欲。他们经常在北京大学和浙江大学等顶尖大学举办招聘活动，招聘广告中甚至吹嘘员工能无限制使用数万个 GPU。他们极具竞争力，据说为有潜力的候选人提供超过 130 万美元的年薪，远超中国其他大型科技公司和人工智能实验室，如 Moonshot。他们目前约有 150 名员工，且规模还在迅速扩大。岗位角色不一定预先设定，招聘人员有一定灵活性。

历史表明，资金充足且专注的小型初创公司往往能突破极限。深度求索不像谷歌那样官僚主义，由于是自筹资金，他们能迅速将想法付诸实践。不过，和谷歌一样，深度求索（在很大程度上）运营自己的数据中心，不依赖外部机构或供应商。这为实验开辟了更多空间，使他们能够在整个技术栈上进行创新。

我们认为他们是目前最好的“开放权重”实验室，超过了 Meta 的 Llama 项目、Mistral 等。

深度求索的成本与性能

本周，深度求索的价格和效率引发了热潮，主要焦点是深度求索 V3 的“600 万美元”训练成本。但这是错误的。这就好比只看产品物料清单上的某一部分，却将其视为整个产品的成本。预训练成本只是总成本中很小的一部分。

训练成本

我们认为预训练成本远非该模型的实际投入。我们确信，在公司发展历程中，他们在硬件上的花费远高于 5 亿美元。为了开发新的架构创新，在模型开发过程中，需要投入大量资金来测试新想法、新架构思路，并进行消融实验。开发和实现这些想法需要整个团队投入大量人力和 GPU 计算时间。深度求索的关键创新——多头潜在注意力机制（Multi-Head Latent Attention），就耗费了数月时间。

论文中提到的 600 万美元成本仅指预训练运行的 GPU 成本，这只是模型总成本的一部分。研发费用和硬件本身的总拥有成本等重要部分并未计算在内。参考一下，Claude 3.5 Sonnet 的训练成本高达数千万美元，如果这就是 Anthropic 所需的全部成本，他们就不会从谷歌筹集数十亿美元，也不会从亚马逊筹集数百亿美元了。这是因为他们必须进行实验、提出新架构、收集和清理数据、支付员工工资等等。

那么深度求索是如何拥有如此庞大的集群的呢？出口管制的滞后是关键，下面在出口管制部分会详细讨论。

缩小差距——V3 的性能

V3 无疑是一款令人印象深刻的模型，但值得注意的是，要明确它是相对于什么而言令人印象深刻。许多人将 V3 与 GPT-4o 进行比较，并强调 V3 如何超越 4o 的性能。这确实没错，但 GPT-4o 于 2024 年 5 月发布。人工智能发展迅速，从算法改进的角度来看，2024 年 5 月恍如隔世。而且，经过一段时间后，用更少的计算资源实现相当或更强的能力，这并不令人意外。推理成本的下降是人工智能进步的一个标志。

DeepSeek-V3 Competitive Analysis						
Model	Price / 1M Input Tokens	Price / 1M Output Tokens	MMLU (Pass@1)	SWE Verified (Resolved)	AIME 2024	MATH-500
Claude-3.5-Sonnet-1022	\$3.00	\$15.00	88.3	50.8	16.0	78.3
GPT-4o-0513	\$2.50	\$10.00	87.2	38.8	9.3	74.6
DeepSeek-V3 (TogetherAI)	\$1.25	\$1.25	88.5	42.0	39.2	90.2
DeepSeek-V3 Median Provider ⁴	\$0.90	\$1.10				
DeepSeek-V3 (Normal Price) ^{1,2}	\$0.27	\$1.10				
DeepSeek-V3 (Discount Price) ^{1,2,3}	\$0.14	\$0.28				
Gemini 1.5 Pro	\$1.25	\$5.00	86.0		20.0	88.0
GPT-4o-mini	\$0.15	\$0.60	82.0	33.2	6.7	79.0
Llama 3.1 405B	\$3.50	\$3.50	88.6	24.5	23.3	73.8
Llama 3.2 70B	\$0.59	\$0.73	86.0		20.0	64.0

1. Hosted by DeepSeek.
2. Cache Miss Input Token Pricing.
3. DeepSeek-V3 pricing discounted through 8 Feb 2025.
4. Median price across providers.

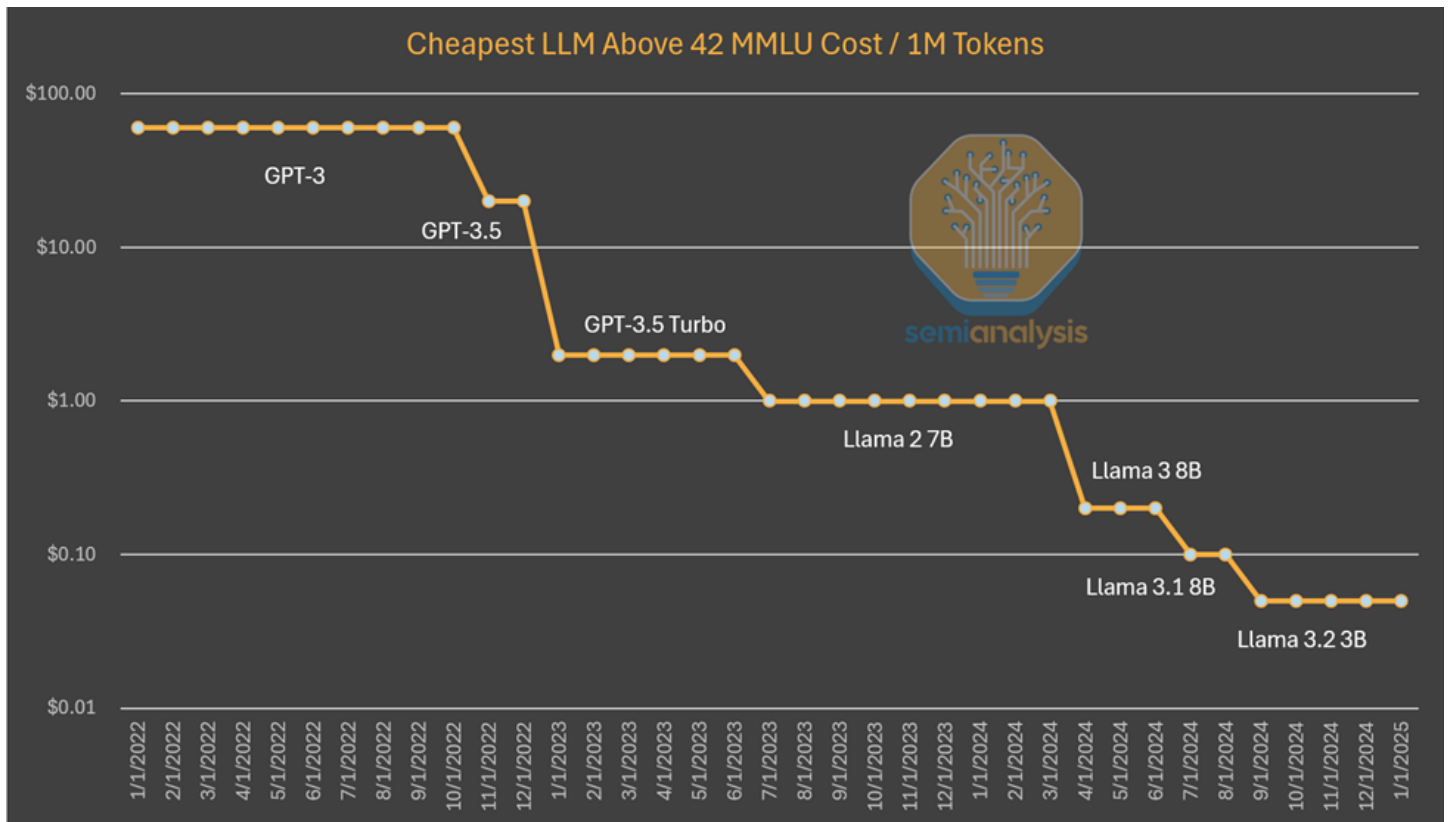
Source: SemiAnalysis

深度求索 V3 的竞争分析

例如，能在笔记本电脑上运行的小型模型，其性能可与 GPT-3 相媲美，而 GPT-3 的训练需要超级计算机，推理则需要多个 GPU。换句话说，算法的改进使得用更少的计算资源来训练和推理具有相同能力的模型成为可能，这种模式反复出现。这次全世界之所以关注，是因为它来自中国的一个实验室。但小型模型性能提升并非新鲜事。

到目前为止，我们从这种模式中看到，人工智能实验室为了获得更高的智能水平，在绝对金额上的投入越来越多。据估计，算法的进步意味着每年实现相同能力所需的计算资源减少 4 倍。Anthropic 的首席执行官 Dario 认为，算法定价在朝着 GPT-3 质量发展，成本已下降 1200 倍。就推理而言，甚至可以实现 10 倍的改进。

在研究 GPT-4 的成本时，我们也看到了类似的成本下降趋势，不过处于曲线的更早期阶段。虽然随着时间推移成本差异的缩小，不能像上面的图表那样通过保持能力不变来解释。在这种情况下，我们看到算法改进和优化使成本降低了 10 倍，同时能力也有所提升。



需要明确的是，深度求索的独特之处在于他们率先达到了这样的成本和能力水平。他们发布开放权重的做法也很独特，不过之前 Mistral 和 Llama 模型也有过类似举措。深度求索达到了这样的成本水平，但到今年年底，如果成本再下降 5 倍，也不要感到惊讶。

另一方面，R1 能够取得与 o1 相当的结果，而 o1 直到 9 月才发布。深度求索是如何这么快就追赶上的呢？

答案是，推理是一种新范式，与之前的预训练范式相比，它的迭代速度更快，且更容易实现较小计算量下的显著提升，而之前的预训练范式成本越来越高，且难以取得稳健的进展。如我们在报告中所所述，之前的范式依赖于规模定律。

新范式通过在现有模型的训练后阶段，利用合成数据生成和强化学习来提升推理能力，能够以更低的成本实现更快的进步。较低的进入门槛和易于优化的特点，使得深度求索能够比往常更快地复制 o1 的方法。随着参与者在这种新范式中找到更多扩展方法，我们预计实现相同能力所需的时间差距将会扩大。

需要注意的是，R1 的论文中并未提及所使用的计算资源。这并非偶然——为训练后的 R1 生成合成数据需要大量计算资源，更不用说强化学习了。我们并不否认 R1 是一款非常优秀的模型，能如此迅速地在推理能力上追赶上令人钦佩。深度求索作为一家中国公司，用更少的资源实现了追赶，这更是令人赞叹。

但 R1 提到的一些基准测试也具有误导性。将 R1 与 o1 进行比较很棘手，因为 R1 特别没有提及那些自己不领先的基准测试。虽然 R1 在推理性能上与 o1 相当，但它并非在所有指标上都是明显的赢家，在很多情况下甚至不如 o1。

我们还没有提到 o3。o3 的能力明显高于 R1 和 o1。事实上，OpenAI 最近公布了 o3 的结果，其基准测试成绩直线上升。“深度学习遇到了瓶颈”，但却是另一种情况。

谷歌的推理模型与 R1 相当

在人们为 R1 疯狂炒作时，一家市值 2.5 万亿美元的美国公司——谷歌，提前一个月发布了一款推理模型 Gemini Flash 2.0 Thinking，且价格更低。这款模型可供使用，通过 API 调用时，即使其上下文长度更长，价格也比 R1 便宜得多。

在已公布的基准测试中，Flash 2.0 Thinking 的表现优于 R1，尽管基准测试并不能说明全部情况。谷歌只公布了 3 个基准测试结果，所以这只是一个不完整的画面。不过，我们认为谷歌的模型很可靠，在很多方面都能与 R1 抗衡，却没有得到任何炒作。这可能是因为谷歌的市场推广策略平淡无奇，用户体验也不佳，但也可能是因为 R1 来自中国，令人感到意外。

DeepSeek-V3 Competitive Analysis						
Model	Price / 1M Input Tokens	Price / 1M Output Tokens	MMLU (Pass@1)	SWE Verified (Resolved)	AIME 2024	MATH-500
Claude-3.5-Sonnet-1022	\$3.00	\$15.00	88.3	50.8	16.0	78.3
GPT-4o-0513	\$2.50	\$10.00	87.2	38.8	9.3	74.6
DeepSeek-V3 (TogetherAI)	\$1.25	\$1.25	88.5	42.0	39.2	90.2
DeepSeek-V3 Median Provider ⁴	\$0.90	\$1.10				
DeepSeek-V3 (Normal Price) ^{1,2}	\$0.27	\$1.10				
DeepSeek-V3 (Discount Price) ^{1,2,3}	\$0.14	\$0.28				
Gemini 1.5 Pro	\$1.25	\$5.00	86.0		20.0	88.0
GPT-4o-mini	\$0.15	\$0.60	82.0	33.2	6.7	79.0
Llama 3.1 405B	\$3.50	\$3.50	88.6	24.5	23.3	73.8
Llama 3.2 70B	\$0.59	\$0.73	86.0		20.0	64.0

1. Hosted by DeepSeek.
2. Cache Miss Input Token Pricing.
3. DeepSeek-V3 pricing discounted through 8 Feb 2025.
4. Median price across providers.
Source: SemiAnalysis

需要明确的是，这些都无损于深度求索的卓越成就。深度求索作为一家行动迅速、资金充足、人才济济且专注的初创公司，能够在推理模型发布上击败 Meta 等巨头，值得称赞。

技术成就

深度求索已经找到了关键方法，实现了领先实验室尚未取得的创新。我们预计，深度求索公布的任何改进，几乎都会立即被西方实验室效仿。

这些改进有哪些呢？大多数架构上的成就都与 V3 相关，V3 也是 R1 的基础模型。下面详细介绍这些创新成果。

训练（预训练和后训练）

深度求索 V3 大规模应用了前所未有的多令牌预测（MTP）技术，它增加了注意力模块，能够预测接下来的几个令牌，而非单个令牌。这一技术在训练过程中提升了模型性能，且在推理时可舍弃。这是通过算法创新实现低计算量下性能提升的一个范例。

训练过程中还采用了 FP8 精度等技术，不过美国的领先实验室采用 FP8 训练已有一段时间。

深度求索 V3 也是一个混合专家模型，即由多个擅长不同领域的小模型组成一个大型模型，这是一种新兴的模型架构。混合专家模型面临的一个难题是如何确定每个令牌该进入哪个子模型（即“专家”模型）。深度求索通过实施“门控网络”，以一种平衡的方式将令牌路由到合适的专家模型，且不影响模型性能。这意味着路由效率极高，在训练过程中，相对于整个模型的规模，每个令牌仅需改变少量参数。这不仅提高了训练效率，还降低了推理成本。

尽管有人担忧混合专家模型（MoE）带来的效率提升可能并不显著，节省下来的成本会迅速被投入到构建更大规模的模型中，导致总体投入不会减少。但实际上，MoE 提高效率会加速人工智能的规模化发展。企业都在专注于扩大模型的计算规模，并提升算法效率。达里奥指出，更强大的人工智能模型所带来的经济效益十分可观。

就 R1 而言，它极大地受益于强大的基础模型（V3），部分原因在于强化学习（RL）。强化学习主要聚焦两个方面：格式规范（确保输出连贯）以及有用性和无害性（确保模型实用）。在基于合成数据集对模型进行微调的过程中，R1 的推理能力得以提升，这与 o1 的情况类似。需要注意的是，R1 的论文中并未提及计算资源的使用情况，因为提及所用的计算资源会暴露他们实际拥有的 GPU 数量比对外宣称的更多。如此大规模的强化学习，尤其是在生成合成数据时，需要大量的计算资源，正如我们在关于规模定律的文章中所提到的。

此外，深度求索使用的部分数据似乎来自 OpenAI 的模型，我们认为这可能会对输出数据提取相关政策产生影响。从服务条款来看，这种数据提取行为已经属于违规。未来，一种类似“了解你的客户”（KYC）的机制可能会出现，以杜绝此类数据提取行为。

多头潜在注意力机制（MLA）

MLA 是深度求索大幅降低推理成本的关键创新。它可将每次查询所需的 KV 缓存减少约 90%（相较于标准注意力机制）。KV 缓存是 Transformer 模型中的一种内存机制，用于存储对话上下文数据，减少不必要的计算。

正如我们在规模定律文章中所讨论的，随着对话上下文的增加，KV 缓存也会增大，从而带来显著的内存限制问题。大幅减少每次查询所需的 KV 缓存，意味着每次查询所需的硬件资源减少，进而降低成本。不过，我们认为深度求索以成本价提供推理服务是为了获取市场份额，实际上并未盈利。谷歌的 Gemini Flash 2.0 Thinking 价格更低，而且谷歌不太可能以成本价提供服务。MLA 尤其引起了美国许多领先实验室的关注，它于 2024 年 5 月随深度求索 V2 发布。由于 H20 相较于 H100 具有更高的内存带宽和容量，深度求索在使用 H20 进行推理工作负载时效率更高。他们还宣布与华为建立合作关系，但目前在昇腾计算方面的合作成果尚不明显。

我们认为，MLA 对利润率的影响最为值得关注，这对整个生态系统意义重大。以下是我们对未来人工智能行业定价结构的展望，同时详细阐述了为何认为深度求索在补贴价格，以及杰文斯悖论初现端倪的原因。此外，我们还将探讨出口管制的影响、中国政府可能对深度求索日益增长的主导地位做出的反应等问题。

对利润率的广泛影响

在利润率方面，有一个关键发现：R1 并非从技术层面削弱了 o1 的进展，而是以更低的价格实现了相当的能力。这在本质上是合理的，现在我们引入一个关于未来定价机制的框架。

提升能力能够带来更高的利润率。这与半导体制造行业的发展极为相似，台积电率先进入新节点（实现新能力）时，由于创造出了前所未有的产品，从而获得了显著的定价权。

其他落后的竞争对手（如三星、英特尔）为了在性价比上达到平衡，会以低于领先者的价格提供产品。对芯片制造商（在此类比为人工智能实验室）而言，幸运的是他们可以调整产能。如果在新模型上能够实现更高的性价比，他们就可以将产能转移到新模型的生产上。旧型号仍会得到支持，但供应量会减少。这与当前人工智能实验室的实际情况以及半导体制造行业的规律高度吻合。

能力的商品化与对更强能力的不懈追求

这或许就是能力竞争的未来走向。率先达到新的能力层级，将获得可观的定价溢价；而那些迅速跟上的参与者，只能获得微薄利润。处于能力层级下游的产品，如果能满足特定用例的需求，仍会继续存在。每一代能够追赶上领先能力的参与者将越来越少。

我们见证的是，R1 达到了领先的能力水平，却以零利润率定价。这种巨大的价格差异引发了一个问题：为什么 OpenAI 的产品如此昂贵？这是因为他们基于最前沿的技术定价，并享受着前沿技术带来的溢价。

我们认为，未来的发展将比领先的芯片制造动态更快。追逐最新的能力意味着持续的定价权（例如 ChatGPT Pro），而落后的能力则意味着更低的定价，此时利润主要来源于为令牌服务的基础设施。

鉴于我们正处于快速的技术周期中，为追求领先的能力，产品更新换代的速度也会加快。只要你能不断拓展能力，开发出创造价值的新功能，就理应获得定价权；否则，在开放模型市场中，你很快就会面临产品同质化的问题。

我们认为，在这种背景下，人们对当前发生的事情存在根本性的误解。我们所描述的情况类似于超高速发展的芯片制造行业，这是世界上资本密集度最高的行业。全球没有哪个行业在研发上的投入比芯片制造行业更多，但与之最相似的现实情况却被认为对支持模型公司的芯片产业不利。

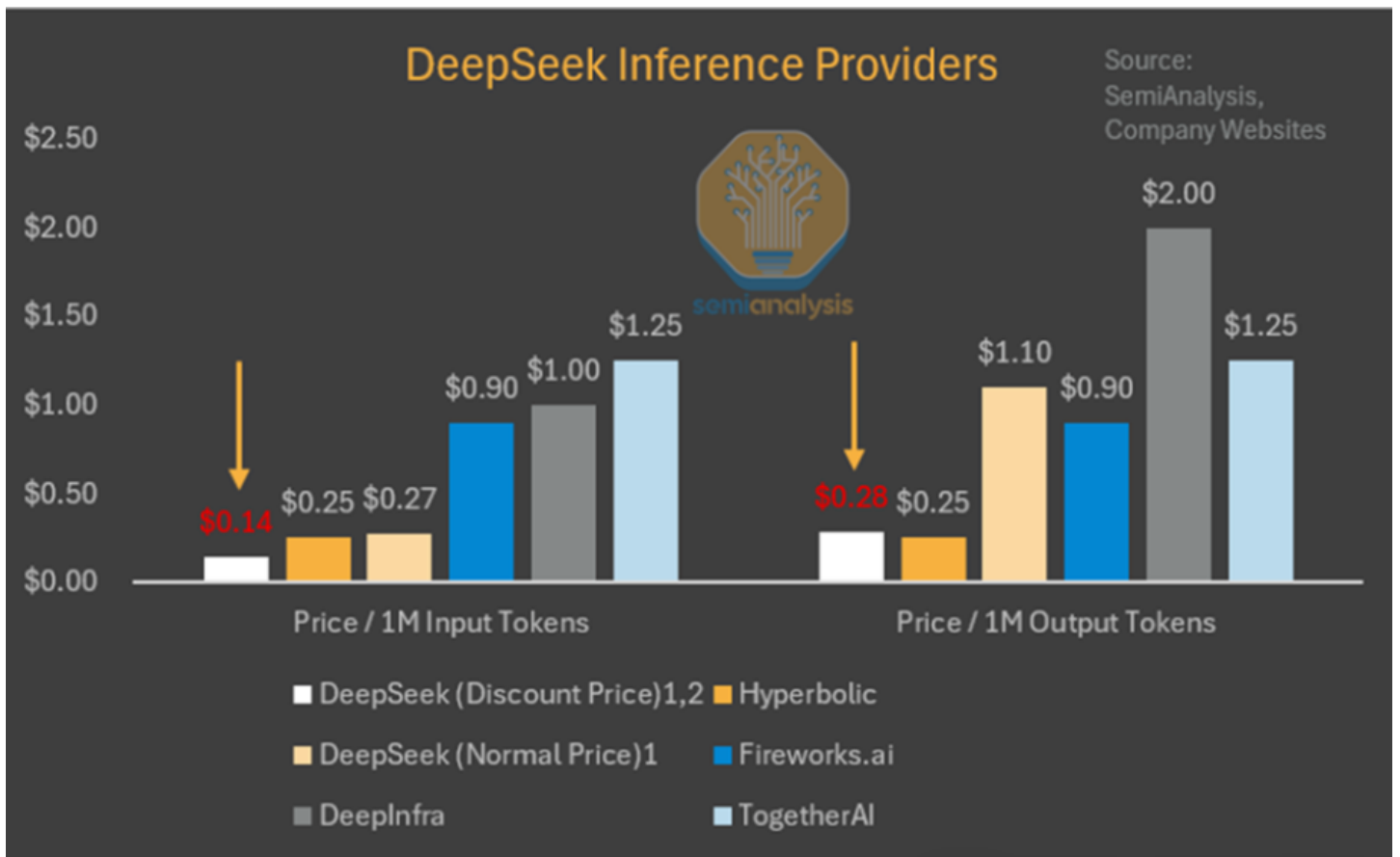
将人工智能令牌与杰文斯悖论相比较，会发现二者有着深刻的历史相似性。起初，人们并不确定晶体管是否能够不断缩小尺寸；而当这一趋势明确后，整个行业便致力于将互补金属氧化物半导体

（CMOS）技术的尺寸缩小到极致，并在此基础上构建出各种重要功能。我们目前正处于整合多种思维链（CoT）模型和能力的初期阶段，就像最初对晶体管进行规模化发展一样。虽然从技术进步的角度来看，这可能是一个动荡时期，但对英伟达来说却是有利的。

深度求索补贴推理利润率

实际情况是，市场在寻找一个理由，而他们选择了这一点。如果深度求索愿意接受零利润率甚至负利润率，那么他们的产品价格可能会如此之低，但显然，提供前沿令牌服务的价格弹性点要高得多。考虑到深度求索正在进行新一轮融资，他们有动机这样做。

深度求索在推理领域的关键切入点上，打破了 OpenAI 的领先利润率。这种领先地位会持续下去吗？我们认为不会——毕竟一个开放实验室展示出了封闭实验室的能力。尽管这一点至关重要，但我们仍需注意，深度求索是一个快速追随者。



我们确实认为，一个更强大的开放实验室（深度求索目前是其中的佼佼者）对新兴云服务提供商和服务供应商来说是非常有利的。无论是开放模型还是封闭模型，计算资源的集中化仍然很重要，但如果基于计算资源构建的上层服务免费提供产品，那么计算资源的价值就有可能提升。更多的资金会流向计算资源领域，而非封闭模型供应商，这意味着支出更多地流向了硬件领域。软件企业也能从中受益匪浅。

H100 价格飙升 —— 杰文斯悖论的体现

我们已经看到了这一理论的早期迹象。自 V3 和 R1 发布以来，AWS 多个地区的 H100 GPU 价格上涨，H200 也更难获取。

V3 发布后，H100 价格大幅上涨，因为 GPU 开始以更高的费率实现货币化。更低的成本实现更强的智能意味着更多的需求。这与前几个月 H100 现货价格的低迷形成了鲜明对比。

出口管制的影响、深度求索与中国政府

从地缘政治的角度来看，深度求索与西方实验室在能力方面的对比，以及出口管制的影响，都值得深入思考。目前已经实施的人工智能扩散管制措施，我们认为不会取消。有消息称，出口管制因深度求索的发展而失败，但这是对出口管制机制的误解。最初，H100 被禁止出口，而计算能力相近（但带宽受限）的 H800 被允许出口；随后，H800 也被禁止，现在仅允许 H20 出口。我们在《加速器模型》中提到，尽管需求巨大，但英伟达在 1 月份取消了大量 H20 订单，这可能预示着美国即将出台新的禁令。

在这些法律的实施过程中存在宽限期，深度求索很可能在这段时间内大量囤积所需芯片。需要注意的是，H100 自发布以来就被禁止出口。从这个角度来看，出口管制未能完全限制高性能芯片的供应。出

口管制的目的并非完全切断中国获取芯片的渠道，而是对整个生态系统进行严格限制，意味着限制数十万甚至数百万芯片的供应，而不仅仅是数万个。

然而，我们预计未来 H20 也将被禁止出口，这将进一步限制深度求索获取芯片的能力。

而他们对芯片的需求十分迫切。

深度求索的产能限制

深度求索难以满足急剧增长的需求。尽管他们拥有世界上最出色的推理技术之一，但进行架构研发、训练模型，与为数千万用户提供可靠服务是截然不同的挑战。深度求索的注册服务时常关闭，即便开放注册时，R1 的响应速度也极慢（不过巧妙的用户体验设计掩盖了这一问题）。

我们本月看到的模型受之前出口管制的影响，存在一定滞后性。随着时间推移，深度求索在扩展模型和服务能力方面将面临越来越大的困难。扩展能力迫在眉睫，中国也深知这一点。

在与深度求索的首席执行官兼创始人会面后的第二天，中国银行宣布未来 5 年将为人工智能产业链提供 1400 亿美元（1 万亿元人民币）的补贴。该补贴的明确目标是助力中国在科技领域实现完全自主，涵盖基础研究、产业应用和开发等方面。人工智能与机器人、生物技术和新材料是重点关注领域。此外，补贴还包括计算基础设施和数据中心建设，以及为第一代技术设备提供保险和风险管理支持。

我们认为，未来出口管制的影响将更加显著：算法和硬件都将不断进步，美国的实验室能够利用这些创新成果进行扩展，达到中国难以企及的高度。虽然中国可能仍会推出与美国实验室相媲美的模型，但将继续处于追赶地位。

我们也认为，从长期来看，深度求索有可能不再开源模型，尤其是在中国政府对其工作给予更多关注，并致力于保护算法创新的情况下。

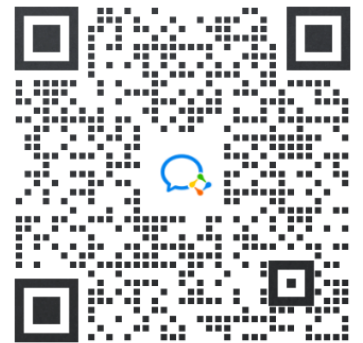
免责声明：

1. 本资料来源于网络公开渠道，版权归属版权方；
2. 本资料仅限会员学习使用，如他用请联系版权方；
3. 会员费用作为信息收集整理及运营之必须费用；
4. 如侵犯您的合法权益，请联系客服微信将及时删除



行业报告资源群

1. 进群福利：进群即领万份行业研究、管理方案及其他学习资源，直接打包下载
2. 每日分享：6份行研精选报告、3个行业主题
3. 报告查找：群里直接咨询，免费协助查找
4. 严禁广告：仅限行业报告交流，禁止一切无关信息



微信扫码，长期有效

更多AI工具可直接访问：<https://www.faxianai.com/>

Deepseek满血版入口：<https://www.faxianai.com/ai/6039.html>

知识星球 行业与管理资源

专业知识社群：每月分享8000+份行业研究报告、商业计划、市场研究、企业运营及咨询管理方案等，涵盖科技、金融、教育、互联网、房地产、生物制药、医疗健康等；已成为投资、产业研究、企业运营、价值传播等工作助手。



微信扫码，行研无忧