不可替代性战略

英伟达、月之暗面与 DeepSeek 的三条路径

殷彤 | InsightBridge Global

2026 年 11 月



引言：竞争的终点不是领先，而是不可替代

在人工智能产业里，领先往往只能维持几个月。今天排名第一的模型，下一季度就可能被超越；今天最便宜的接口，下一轮降价就会失去优势。真正决定企业命运的，不是某一时刻的领先，而是客户在离开它时要付出的代价有多大。这种代价越高，企业越接近“不可替代”。

2026 年的全球 AI 格局，为这一命题提供了一组难得的对照样本。英伟达以约 75% 的毛利率和约 5 万亿美元的市值站在产业链顶端[1][3]；月之暗面以 2.8 万亿参数的 Kimi K3 进入全球前沿

模型的第一梯队[8]；DeepSeek 则以近乎免费的价格和持续不断的开源发布，成为全球开发者路由流量中占比最高的单一模型提供方之一[17]。三家企业的路径截然不同，但追求的是同一个目标：成为客户、开发者乃至整个生态体系离不开的那一环。

本文的核心判断是：不可替代性可以通过三条不同的路径获得。第一条是垄断，即通过生态锁定把转换成本做到极高；第二条是技术，即在某个关键维度上做到他人短期无法企及；第三条是普及，即持续提供足够好的技术，以极低的价格先占领市场，让自己成为行业默认的基础设施。英伟达代表第一条，月之暗面代表第二条，DeepSeek 代表第三条。三者都是不可替代性战略，但其来源、代价、可持续性与社会后果截然不同。



图 1　三条通往不可替代的路径

一、不可替代性的本质：转换成本与替代难度

理解不可替代性，需要把它拆成两个变量。其一是替代难度，即市场上是否存在功能相当的选择；其二是转换成本，即即便存在替代品，客户迁移过去需要付出多少时间、资金和风险。一

家企业只要在其中一项上做到极致，就能获得相当程度的定价权；两项同时具备，便形成事实上的垄断。

三条路径的区别，恰恰在于它们分别作用于哪个变量。垄断路径主要抬高转换成本：替代品并非不存在，而是迁移过去太贵、太慢、太冒险。技术路径主要抬高替代难度：在能力的最前沿，暂时没有第二个选择。普及路径则同时改变两个变量的含义：它不阻止客户离开，而是让离开变得没有必要，因为它已经是最便宜、最方便、被最多工具默认支持的选项。

这三种不可替代性的时间属性也不同。垄断型不可替代性来自过去积累的沉没成本，是“向后看”的；技术型不可替代性依赖持续保持领先，是“向前看”的；普及型不可替代性依赖在生态中的位置，是“横向铺开”的。时间属性决定了每条路径最脆弱的地方。

二、英伟达：以垄断实现的不可替代

生态锁定的形成

英伟达今天的地位，与其说来自对人工智能浪潮的精准预判，不如说来自一次历史性的契合。图形处理器为游戏画面渲染而设计，其大规模并行计算的特性，恰好与深度学习的矩阵运算高度吻合。真正把这种偶然契合固化为结构性优势的，是 2006 年推出的 CUDA 软件平台。按英伟达自报的数据，CUDA 如今支撑着约 600 万开发者[5]。主流深度学习框架、数以百计的优化库和大量生产代码都以 CUDA 为基础，迁移到其他硬件意味着重写、重调和重新验证。

这正是垄断型不可替代性的典型结构：硬件本身并非无法替代，替代的代价才是壁垒。第三方机构对英伟达在 AI 加速器市场份额的估计普遍在 80% 至 88% 之间，在训练市场则超过 90%[4]。

定价权与利润分配

垄断型不可替代性最直接的结果是定价权。英伟达 2027 财年第二季度营收 962 亿美元，同比增长 106%，毛利率 75.0%[1]；数据中心业务占总营收约 93%[2]。在一个仍处于早期、下游大量企业尚未实现盈利的新兴市场里，上游供应商长期维持约 75% 的毛利率，意味着产业链的大部分价值被集中在硬件一端。

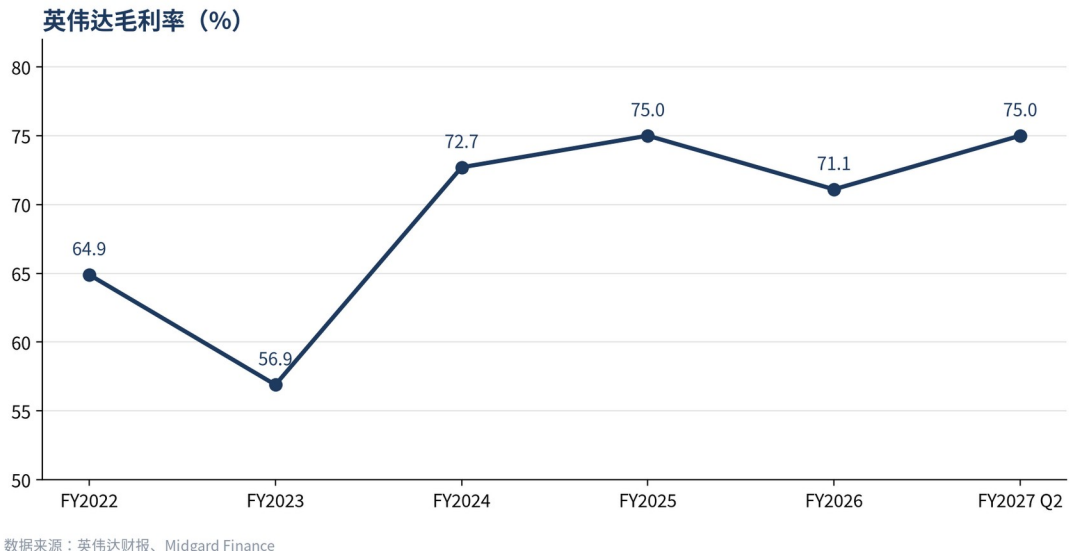


图 2　英伟达毛利率走势

从战略角度看，这一利润结构有两层含义。其一，它证明了生态锁定的威力：客户在没有足够替代选择的情况下，只能接受高价。其二，它也埋下了反向激励。下游云服务商、模型公司和应用开发者承受的成本越高，寻找替代方案的动力就越强。

垄断的代价：质量纪律与替代反扑

垄断型不可替代性有一个内在风险：当客户无法离开时，企业改进产品的外部压力会减弱。英伟达面向开发者的桌面 AI 系统 DGX Spark 提供了一个具体案例。其官方开发者论坛上，用户持续报告在长时间推理和训练负载下出现冻屏、无响应、需要物理断电才能恢复等问题；对于内存溢出导致整机挂起的问题，论坛版主确认这是“已知问题，正在积极修复”[7]，英伟达也在

2026 年通过多轮固件更新处理电源与稳定性问题[7]。单一产品的问题不能概括一家公司的全部技术能力，但它揭示了一种结构性倾向：当利润主要来自锁定而非持续的产品优势时，产品打磨的优先级容易让位于产能扩张与资本叙事。

更重要的是，高额利润正在推动替代力量加速成长。谷歌 TPU、亚马逊 Trainium、Meta MTIA 等自研芯片在 AI 算力负载中的占比，预计将从 2025 年的 20.9% 升至 2026 年的 27.8%[6]。在中国市场，英伟达对 2027 财年第三季度的业绩指引已假设来自中国的数据中心计算收入为零[23]；DeepSeek 则在 2026 年 9 月开源了全套华为昇腾平台基础组件，与其此前为英伟达平台开源的组件一一对应[22]。这类工作针对的正是 CUDA 锁定最核心的部分：迁移成本。

垄断型不可替代性的脆弱之处由此显现：它“向后看”，依赖历史积累的锁定。一旦替代生态的迁移成本被系统性降低，过去的沉没成本就不再是护城河，而是客户急于摆脱的负担。

三、月之暗面：以极致技术实现的不可替代

在能力前沿占据位置

月之暗面选择了另一条路。2026 年 7 月 16 日发布的 Kimi K3，总参数 2.8 万亿，是迄今规模最大的开放权重模型，具备原生视觉能力和 100 万 token 的上下文窗口，完整权重于 7 月 27 日公开[8]。该模型在 Artificial Analysis 综合智能指数上首次亮相即位列全球第三，仅次于 Anthropic 与 OpenAI 的旗舰模型[8]；在前端代码竞技场上，它成为首个登顶的中国模型[8]。10 月发布的 K3.1 Pro 在同一指数上得分 55，较前代提高 11 分[13]。

技术型不可替代性的逻辑很直接：在长上下文、长链路推理和复杂编程这些能力维度上，如果客户需要最好的结果，可选项就只有少数几家。月之暗面并不追求最低价格。K3 的接口定价为

每百万输入 token 3 美元、输出 15 美元[10]，约为美国同级闭源旗舰的一半[9]，却明显高于 DeepSeek。它卖的是能力，而不是便宜。

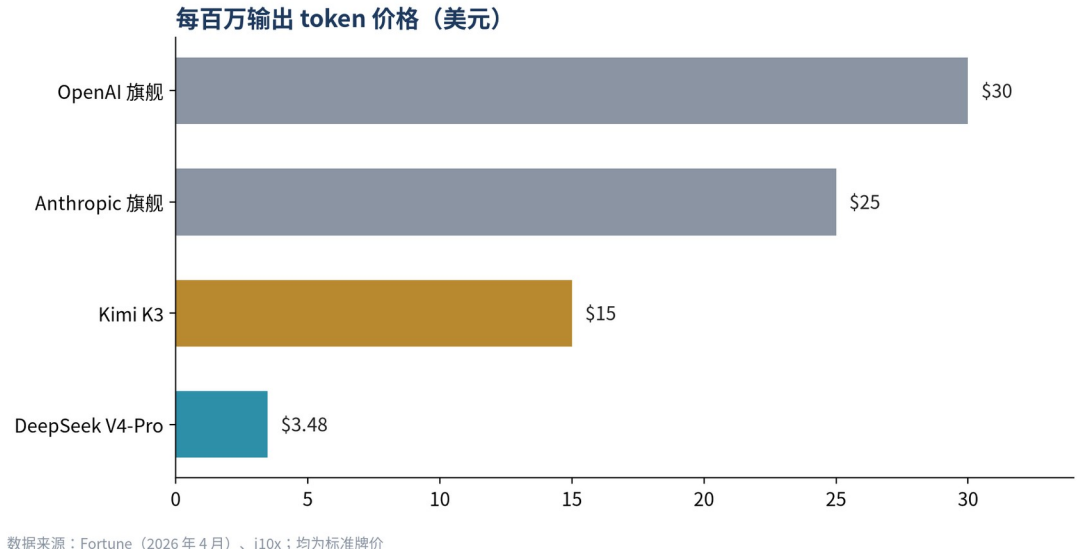


图 3　前沿模型每百万输出 token 价格对比

开放与收费的精巧边界

月之暗面的独到之处，在于它没有因为追求技术高度而走向封闭。K3 以开放权重发布，绝大多数开发者和企业可以免费下载、修改、微调和商用。但其许可证设置了两道门槛：第一，若被许可方以“模型即服务”方式向第三方提供推理或微调，且其与关联方在任意连续 12 个月内的总收入超过 2000 万美元，须先与月之暗面另行签订协议；第二，月活跃用户超过 1 亿或月收入超过 2000 万美元的商业产品，须在界面显著标注“Kimi K3”[11]。

这一设计把开放的扩散效应与前沿技术的商业价值结合在一起：对小开发者，它是免费的前沿模型；对大型云平台和模型服务商，它是需要谈判的商业资产。亚马逊 Bedrock 自 2026 年 9 月起正式提供 K3，正是这一边界在国际市场上的第一次检验[17]。

资本对稀缺性的定价

资本市场对技术型不可替代性给出了明确定价。月之暗面在 2026 年 7 月以 350 亿美元估值完成 35 亿美元融资[25]，据报道其后又以约 500 亿美元估值完成新一轮私募融资，并计划于 2027 年第一季度在香港上市[12]。据财新报道，其年化经常性收入已达约 3 亿美元[8]。估值与收入之间的巨大倍数，本质上是资本对“稀缺能力”的溢价。

技术路径的脆弱之处

技术型不可替代性“向前看”，因此它最大的风险是前沿本身在移动。今天的第三名，可能在下一轮模型发布后跌出前五；任何一次代际落后，都会直接侵蚀其不可替代的基础。此外，前沿技术依赖巨量算力。K3 发布后需求激增，月之暗面一度不得不暂停新的接口订阅[9]，这说明技术领先若没有足够的服务能力支撑，就难以转化为稳定的市场地位。技术路径要求企业持续不断地投入研发，任何松懈都会被迅速惩罚。

四、DeepSeek：以持续供给和市场普及实现的不可替代

先占领市场，同步推进技术

DeepSeek 走的是第三条路。它的核心不是在每一个基准上做到世界第一，而是以惊人的工程效率持续提供足够好的技术，并以极低的价格让它无处不在。2026 年 4 月发布的 V4 系列中，V4-Pro 每百万输出 token 定价 3.48 美元，而 OpenAI 和 Anthropic 同等工作量的收费分别为 30 美元和 25 美元[14]。此后 DeepSeek 又推出分时定价，非高峰时段价格减半；其网页和移动应用完全免费且不设付费层级，模型权重以 MIT 许可证开源，可自由自托管[15]。

更关键的是供给的持续性。从 4 月的 V4 预览版，到 8 月 V4-Pro 正式版，再到 9 月的 V4.1-Flash 以及面向智能体的搜索与网页抓取接口（每千次请求 0.75 美元，远低于同类服务）[15]

[16]，DeepSeek 用一连串快速迭代证明：它不是一次性的价格冲击，而是一个可以长期依赖的供应源。

普及带来的不可替代性

普及路径的成果体现在使用数据上。在全球最大的中立模型路由平台 OpenRouter 上，截至 2026 年 11 月 1 日当周，中国开发者的模型承载了 72% 的 token 流量，而 2025 年初这一比例仅为 7%[17]；在 2026 年 10 月 5 日当周的文本请求中，DeepSeek 一家占 24.5%，高于 OpenAI 的 18.6% 和谷歌的 17.8%[17]。在 Hugging Face 上，中国开放模型的累计下载量到 2026 年 3 月已达 11.5 亿次，超过美国的 7.23 亿次[18]。

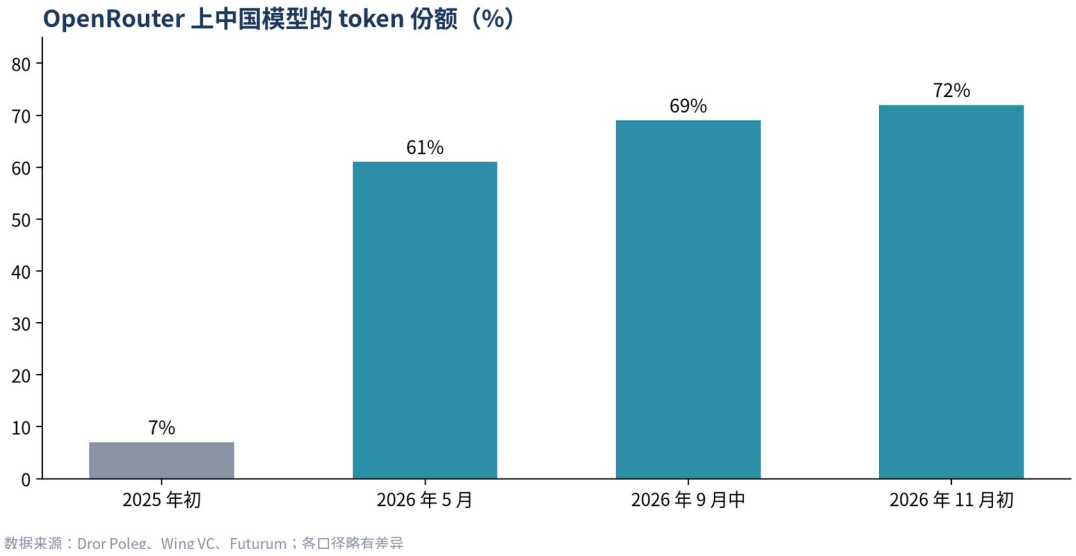


图 4　OpenRouter 上中国模型的 token 份额

这种不可替代性与英伟达式锁定截然不同。DeepSeek 并不阻止客户离开，它的权重任何人都可以下载、部署和修改。它的不可替代性来自另一种机制：当一个模型成为大量智能体框架、开发工具和下游应用的默认选项时，它就成了生态的公共底座。替代它在技术上可行，在经济上却没有理由。Vercel AI 网关 2026 年 6 月的数据显示，开放权重模型承载了近三分之一的

token，却只占不到 4% 的支出[24]。这组数字准确描述了普及路径的商业特征：以极低的单位价格换取极大的使用规模。

资本与上市

普及路径同样获得了资本的确认。据彭博报道，DeepSeek 正在完成至少 800 亿元人民币的第二轮融资，腾讯和宁德时代为主要投资方，估值至少 5000 亿元人民币，约合 750 亿美元，并计划于 2027 年初上市[19]；多方报道显示其目标上市地为上海科创板[19]。

普及路径的脆弱之处

普及路径“横向铺开”，它最脆弱的地方不在技术或价格，而在两个外部约束。其一是服务能力。DeepSeek 自己承认，由于芯片不足，目前尚无法向大多数客户提供 V4-Pro 服务[21]。低价换来的巨大需求，必须由足够的算力承接。其二是政治风险。2026 年 10 月公开的美国国土安全部命令，要求联邦民事机构从信息系统中移除 DeepSeek 产品及基于其技术开发的第三方产品，并禁止采购[20]。这项命令不适用于私营企业和个人用户[20]，但它表明：当一家企业的普及程度足以影响产业格局时，它就不再只被视为商业竞争者。

五、三条路径的比较

维度	英伟达：垄断路径	月之暗面：技术路径	DeepSeek：普及路径
不可替代性的来源	生态锁定，转换成本极高	能力前沿，短期无替代	默认选项，替代无必要
时间属性	向后看，依赖历史沉没成本	向前看，依赖持续领先	横向铺开，依赖生态位置
定价逻辑	高价，毛利率约 75%	中高价，约为美国闭源旗舰一半	极低价，应用免费、权重开源
价值获取方式	硬件销售与系统溢价	接口收费、大客户另行授权	规模化低价接口、资本市场
对生态的影响	价值向上游集中	前沿能力有条件开放	价值向下游扩散
主要风险	替代生态降低迁移成本	前沿移动、算力不足	服务能力不足、政治限制

这张表揭示了一个关键差异：三条路径对产业生态的影响方向相反。垄断路径把价值从下游吸向上游，越成功，下游越有动力寻找替代；普及路径把价值从上游推向下游，越成功，下游越依赖它。技术路径介于两者之间，用有条件的开放兼顾扩散与收费。

由此可以得出第一个结论：不可替代性的可持续性，取决于它是否与客户利益一致。垄断型不可替代性与客户利益存在结构性冲突，客户留下来是因为离开太贵，而不是因为留下最好。这样的地位很稳固，但它每一天都在为替代者积累动力。普及型和技术型不可替代性则与客户利益基本一致：客户选择它们，是因为它们确实更便宜或更强。

第二个结论是：后发者恰恰最适合走后两条路。先发者需要承担开荒、试错和商业化的全部成本，并背负投资人对高利润的期待，很难主动降价或开源。后发者可以在已被验证的技术方向上集中资源，以效率或单点突破切入；它们没有既有利润需要保护，可以用开放和低价改写行业的定价规则。月之暗面与 DeepSeek 在不同维度上把后发优势推向了极致：一个以质量镇守能力前沿，一个以效率渗透生态的每一个角落。

六、最大的变数：国家力量

三条路径之外，还有一个不属于商业范畴、却可能改变一切的变量：国家力量。美国对先进芯片的出口管制，限制了中国模型公司获取训练前沿模型所需的算力；美国国土安全部对 DeepSeek 的联邦采购禁令，则把竞争从市场延伸到合规与安全审查[20]。资本市场同样可能受到影响，相关限制措施可能左右国际机构投资者参与相关融资和上市的意愿。

然而，开放权重路线在这一压力下表现出相当的韧性。其一，已经被全球开发者下载并部署的权重无法被收回，任何行政命令都难以从分散的私人服务器中删除它们。其二，下游生态的依赖会产生反向压力：如果强制切断与最具性价比的模型的联系，首先受损的是本国开发者和企业的成本结构。其三，在资本层面，全球资本的配置早已高度分散，中东主权基金、亚洲家族办公室和国际机构投资者并不只看单一国家的政策信号，数百亿美元规模的融资并非只有一个来源。

国家力量能够提高不可替代性战略的成本，却很难完全消解一种已经嵌入全球开发者日常工作的技术存在。真正的约束在算力：只要先进芯片受到限制，技术路径和普及路径都必须在算法效率和国产硬件适配上持续投入。DeepSeek 把 V4 系列的推理全面适配到华为昇腾平台，并开源相应的基础软件栈[21][22]，正是对这一约束的直接回应。

七、对企业战略的启示

不可替代性战略并不只属于人工智能巨头。任何行业的企业，都需要回答同一个问题：如果客户明天离开我，他们会失去什么，又要付出什么？

第一，企业必须明确自己选择哪一种不可替代性。垄断式锁定可以带来最高的短期利润，但它与客户利益存在冲突，迟早会招致替代。技术领先可以带来稀缺性溢价，但它要求持续不断的

研发投入，没有终点。市场普及可以带来最稳固的生态地位，但它要求极高的成本效率，以及承接规模需求的服务能力。

第二，三条路径可以组合，但不能混淆。月之暗面的许可证设计说明，技术领先者可以借用开放的扩散效应；DeepSeek 的持续迭代说明，普及者也必须保持足够的技术水准。但一家企业若同时想要最高价格、最大份额和最强锁定，往往会在三个方向上都不彻底。

第三，最持久的不可替代性来自客户的主动选择，而不是客户的无路可走。这是英伟达与两家中国模型公司之间最根本的差别。前者的地位建立在“离开太贵”之上，后两者的地位建立在“留下更好”之上。在一个技术迭代以月计算的产业里，后一种不可替代性更可能穿越周期。

结语

英伟达、月之暗面和 DeepSeek 构成了不可替代性战略的三种范式。英伟达用二十年的生态积累把转换成本做到极致，获得了产业链上最丰厚的利润，也培养了最多想要替代它的对手。月之暗面用极致技术占据能力前沿，以有条件的开放兼顾扩散与收费。DeepSeek 用持续的技术供给和极低的价格，让自己成为全球开发者的默认底座。

三条路径都通向不可替代，但通向的是不同性质的不可替代。一种依靠客户无法离开，另外两种依靠客户不想离开。人工智能产业的下一个阶段，很可能就由这两种逻辑之间的竞争决定。

参考资料

[1] NVIDIA, “NVIDIA Announces Financial Results for Second Quarter Fiscal 2027.” https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2027

[2] Midgard Finance, “NVIDIA Revenue, Data Center Growth and Financials.” https://www.midgardfinance.com/research/nvidia/

[3] Yahoo Finance, “Nvidia Is Getting Dangerously Close to Getting Surpassed.” https://finance.yahoo.com/markets/stocks/articles/nvidia-getting-dangerously-close-getting-165800696.html

[4] Axis Intelligence, “Nvidia GPU Market Share 2026.” https://axis-intelligence.com/nvidia-gpu-market-share/

[5] The Dynamics, “NVIDIA CUDA, Explained: The Software Behind NVIDIA’s AI Moat.” https://thedynamics.ai/articles/nvidia-cuda

[6] Axis Intelligence, “Nvidia Blackwell Statistics 2026.” https://axis-intelligence.com/nvidia-blackwell-statistics/

[7] NVIDIA Developer Forums, “Spark hangs – requires a hard-reset.” https://forums.developer.nvidia.com/t/spark-hangs-requires-a-hard-reset-physically-unplugging/358951；GPUsmith, “NVIDIA DGX Spark Specs, Price & Benchmarks.” https://gpusmith.com/articles/en/nvidia-dgx-spark-review-specs-performance

[8] Kimi, “What Is Kimi Agent?” https://www.kimi.com/en/help/agent/agent-overview；South China Morning Post, “Moonshot AI’s Kimi chatbot.” https://www.scmp.com/topics/moonshot-ais-kimi-chatbot；Caixin Global, “Inside the $50 Billion Rise of Moonshot AI.” https://www.caixinglobal.com/2026-09-11/in-depth-inside-the-50-billion-rise-of-moonshot-ai-102484028.html

[9] ThursdAI, “Moonshot AI Releases: Timeline of Every Launch.” https://thursdai.news/companies/moonshot-ai

[10] i10x, “Moonshot AI.” https://i10x.ai/news/company/moonshot

[11] Moonshot AI, “Kimi K3 License,” Hugging Face. https://huggingface.co/moonshotai/Kimi-K3/raw/main/LICENSE

[12] Sacra, “Moonshot AI revenue, valuation & funding.” https://sacra.com/c/moonshot-ai/

[13] China Daily via NCSTI, “Moonshot AI launches upgraded Kimi model.” https://en.ncsti.gov.cn/Latest/news/202610/t20261029_260063.html

[14] Fortune, “DeepSeek unveils V4 model, with rock-bottom prices.” https://fortune.com/2026/04/24/deepseek-v4-ai-model-price-performance-china-open-source/

[15] The AI Rankings, “DeepSeek V4: Benchmarks, Pricing & Specs.” https://theairankings.com/deepseek/deepseek-v4/

[16] Yahoo Finance, “DeepSeek Monetizes the Agent Retrieval Layer at $0.75 per 1,000 Calls.” https://finance.yahoo.com/technology/ai/articles/deepseek-monetizes-agent-retrieval-layer-185236272.html

[17] Dror Poleg, “Chinese and Open-Weight Models’ Share of OpenRouter Tokens.” https://www.drorpoleg.com/charts/openrouter-china-models/；Asia Daily, “Chinese Open AI Models Win Overseas Users.” https://asiadaily.org/news/15338/

[18] “The ATOM Report: Measuring the Open Language Model Ecosystem,” arXiv. https://arxiv.org/html/2604.07190

[19] Oninvest (Bloomberg), “DeepSeek set to raise $12 bln in new funding.” https://en.oninvest.com/article/bloomberg-deepseek-set-to-raise-12-bln-in-new-funding-that-bumps-valuation-by-50；Big Hat Group, “China AI Weekly.” https://www.bighatgroup.com/blog/china-ai-weekly-2026-08-30/

[20] CNBC, “U.S. bans DeepSeek from government agency systems.” https://www.cnbc.com/2026/10/20/us-deepseek-federal-agency-ban.html

[21] Council on Foreign Relations, “DeepSeek V4 Signals a New Phase in the U.S.-China AI Rivalry.” https://www.cfr.org/articles/deepseek-v4-signals-a-new-phase-in-the-u-s-china-ai-rivalry

[22] Frontier Now, “DeepSeek opens Ascend stack.” https://frontiernow.dev/en/news/deepseek-ascend-open-components/

[23] Morningstar, “NVIDIA Corp earnings.” https://morningstar.com/stocks/xmex/nvda/earnings

[24] NewsCord, “Chinese Open-Weight Models Surpass U.S. Models on Hugging Face and OpenRouter.” https://newscord.org/article/chinese-open-weight-models-surpass-us-models-on-hugging-face-and-openrouter–Story_20260626_ItsnotaboutAnthropice54000ee

[25] Stock Analysis, “Moonshot AI Valuation.” https://stockanalysis.com/private/moonshot-ai/

[26] Futurum Group, “Enterprises Embrace Open Weights but Infrastructure Sellers Profit Most.” https://futurumgroup.com/insights/enterprises-embrace-open-weights-but-infrastructure-sellers-profit-most/