证券研究报告

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

DeepSeek：控本筑壁垒，芯模共协同

——大模型商业化系列报告

行业评级：看好（维持）

2026年8月7日

分析师	刘雯蜀	分析师	郑毅
邮箱	liuwenshu03@stocke.com.cn	邮箱	zhengyi@stocke.com.cn
证书编号	S1230523020002	证书编号	S1230524070002

核心结论

■ **DeepSeek商业化加速——DeepSeek 的海外影响力正在由开源传播转化为实际调用和企业付费**

- **DeepSeek一直保持较为克制的产品化速度。**没有专有 Coding 产品，主要通过免费 Web/App、低价 API 和开放权重输出能力，人员扩编重点集中在超算集群、高性能算子、通信与编译器、训练推理框架、分布式存储等底层环节，从人员招聘情况看，有意愿同步完整布局 Agent 全链条研发。The Information报道，DeepSeek ARR 达 4-5 亿美元，V4 模型 API 毛利率 70%-80%。QuestMobile数据显示，国内 App 2026 年 6 月 MAU 1.298 亿。
- **海外商业化有明显提速迹象。**根据 OpenRouter 平台 2026 年 1-6 月统计数据，DeepSeek Token 调用份额由 9% 翻倍升至 18%；根据 Ramp 2026 年 6 月 5 万余家美企账单支付数据，DeepSeek 登顶当月软件趋势榜，Coinbase、Lindy 等美国企业直接付费采购，并非仅采用开源权重本地部署；根据斯坦福《2026 AI Index Report》2026 年 3 月数据，中美头部模型性能差距仅 2.7%，叠加显著价格优势，美企切换至 DeepSeek 后推理成本可下降 30%~95%。

■ **技术降本支撑极致性价比，在提供较大的定价弹性的同时，严控成本形成了一种认知优势。**

- **成本控制的强度可由官方测算佐证：**2025 年 2 月 V3/R1 服务的日租赁成本 8.71 万美元，理论日收入 56.2 万美元，成本利润率 545%（换算为利润÷收入口径约 84.5%）；至 V4，API 业务毛利率仍维持在 70%—80%。具备充足提价盈利空间，2026 年 8 月 6 日官方预告大幅上调 API 定价，并引入峰谷定价，高峰时段调用价格直接翻倍。
- **DeepSeek 的模型研发强调在算力与成本约束下进行系统性优化，这一模式下，降本的第一重意义并不仅在售价，更在于扩大研发搜索空间，**架构与系统效率提升带来单位计算成本下降，成本下降换来更多探索次数，更多探索产生更多可验证结果并更快暴露新瓶颈，反过来又推动下一轮效率提升，本质上是在积累认知优势，卖得便宜只是这一优势外溢的结果。

■ **软硬件 CoDesign 推动技术影响力在国内加速变现，低激活MoE与国产超节点形成较高的架构匹配度**

- DeepSeek 正通过软硬件协同与国产算力形成长期生态绑定。软件层面,适配跨硬件算子语言 TileLang，V3.2 同步开源 CUDA、TileLang 双版本算子，昇腾实现 Day0 同步适配，V4 全线基于 TileLang 开发融合内核，将国产算力适配门槛从 “重写全套 CUDA” 简化为接入统一算子层；数值标准上 V3.1 提前定义 UE8M0 FP8 Scale 适配国产芯片，实现模型定义标准、硬件跟进对齐的双向协同。
- MoE 架构特性与国产超节点系统优势高度契合。MoE 稀疏激活降低单 Token 计算量，但大幅提升专家 All-to-All 通信、统一内存、访存带宽需求，恰好匹配昇腾超节点、灵衢高速互联、CANN 软件栈的系统能力。DeepSeek采用低激活 MoE 架构，V4-Pro 总参 1.6T 仅激活 49B、激活率 3%，搭配 CSA+HCA 混合注意力，百万上下文场景下推理算力、KV 缓存占用仅 V3.2 的 27%、10%。

■ **投资建议：关注国产算力超节点相关标的，以及受益于开源模型的部署的云厂商。建议关注：中芯国际、锐捷网络、紫光股份、阿里巴巴等。**

■ **风险提示：** 技术路线突破不及预期，商业化与盈利不及预期，国产算力适配不及预期，行业竞争加剧与外部政策不确定性。

资料来源：浙商证券研究所整理

01 商业化情况

产品化保持克制

■ **DeepSeek目前公开可见的产品化投入相对克制：**与Kimi、智谱已经形成个人订阅、Coding、Agent和企业服务等较完整的产品矩阵相比，DeepSeek主要通过免费Web/App、低价API和开放权重向外输出能力。

DeepSeek、Kimi、智谱产品体系对比

业务类型	DeepSeek	Kimi	智谱
API	低价 → 分时段定价→将较大幅涨价	K3高价	MaaS / GLM API
Coding	无专有Coding产品	Kimi Code / CLI	ZCode
Agent	模型API供第三方Agent调用	Kimi Claw；Kimi Work	AutoClaw
消费订阅	Web / App免费	￥49—￥699/月个人订阅；Agent Credits	Z.ai订阅 / Coding Plan（￥118-￥1078/月）
企业	开放权重自部署 / 第三方MaaS	Kimi Work / Business：￥2980/席/年	企业/通用LLM/Agent/本地化/MaaS

DeepSeek主要产品为Web/APP及API

资料来源：DeepSeek官网，Kimi官网，智谱官网，Kimi月之暗面公众号，浙商证券研究所整理

商业化以API为主，DeepSeek仍实现4—5亿美元ARR

- 据The Information 报道，DeepSeek 近期 ARR 已达到 **4 亿至 5 亿美元**，主要来自云端 API 调用，其 V4 模型 API 业务毛利率 **70%—80%**。

公司	最近公开ARR	API收入	毛利率	用户规模
DeepSeek	**4亿—5亿美元**（2026年7月）	**API为主要收入来源**	**V4模型API毛利率70%—80%**	**中国App MAU 1.298亿**（2026年6月）
月之暗面	**3亿美元**（2026年6月中）	**API收入占整体收入超过70%**，海外付费用户与API收入均增长400%（截止2026年6月中旬）	**未披露**	**中国App MAU约729万**（2026年6月）
智谱	**10亿美元**（2026年7月）	2025年云端业务收入**1.90亿元人民币**，占总收入26.28%	2025年公司**综合毛利率41.0%**	平台注册企业及用户**超过400万**；全球付费开发者**24.2万**（截至2026年3月）
Anthropic	**741亿美元（**2026年7月**）**	**增长主要来自企业和开发者**，通常从API、Claude Code或Claude Cowork切入	约**44%公司整体毛利率***	Claude移动端约**642万MAU**，（2026年3月）
OpenAI	**413亿美元（**2026年7月**）**	**企业收入超过总收入40%**（2026年4月）	约**42%公司整体毛利率***	ChatGPT **WAU 9亿+**（2026年3月）ChatGPT Work（Codex升级版）用户已超**900万（**2026年7月**）**

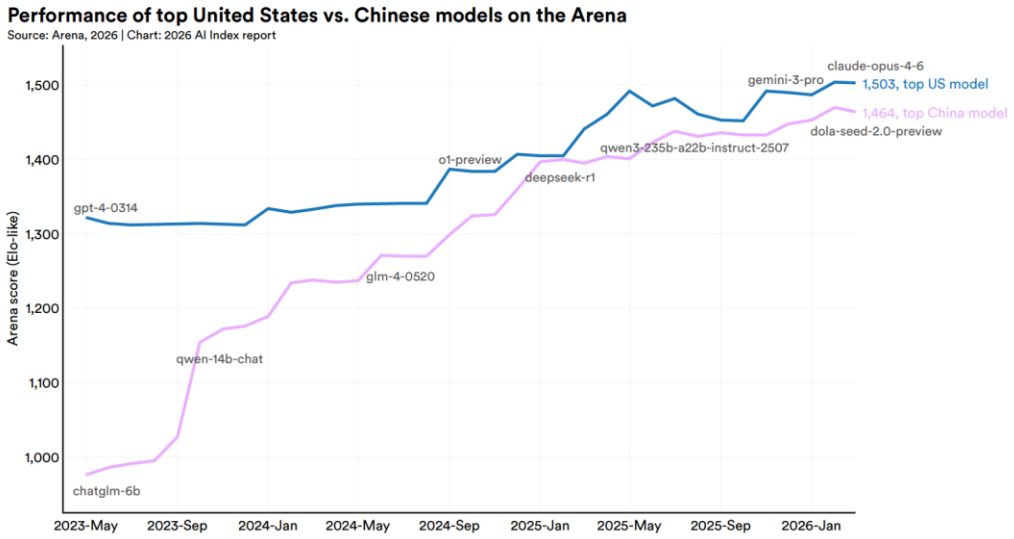
**注：J.P. Morgan 2026年6月基于行业资料的估算*

资料来源：The Information，QuestMobile公众号，上海证券报，36氪，wind，智谱2025年年报，tickertrends，X.com，OpenAI官网， J.P. Morgan等，浙商证券研究所整理

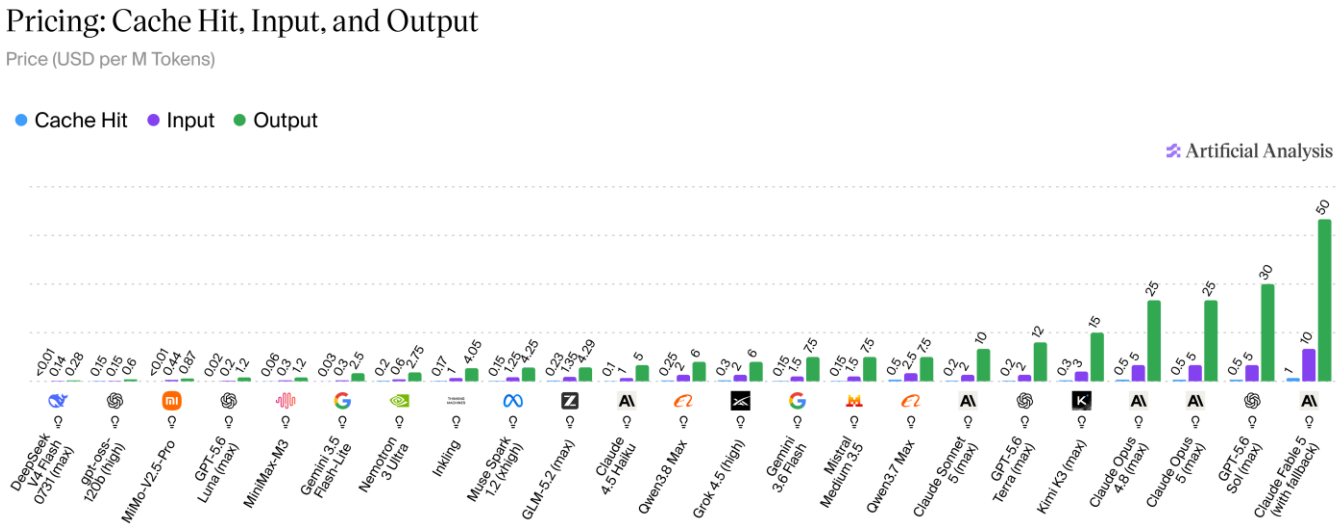
性价比驱动美国企业转向国产模型，DeepSeek低价优势突出

- **性能差距收窄后，价格成为美国企业转向国产模型的重要驱动力。** 斯坦福《2026 AI Index Report》显示，截至2026年3月，中美头部模型性能差距已收窄至约**2.7%**，且中国模型相较美国闭源模型具备较大的价格优势；据每日经济新闻报道，部分受访美国企业在迁移至中国模型后，**推理成本下降30%～95%。**
- **DeepSeek以显著低于海外前沿模型的API价格承接成本敏感型需求。** 主流模型价格对比显示，DeepSeek的输入与输出价格均处于行业低位，进一步放大其在海外市场的性价比优势。

中美头部模型性能差距已收窄至约2.7%



DeepSeek API价格处于全球主流模型低位

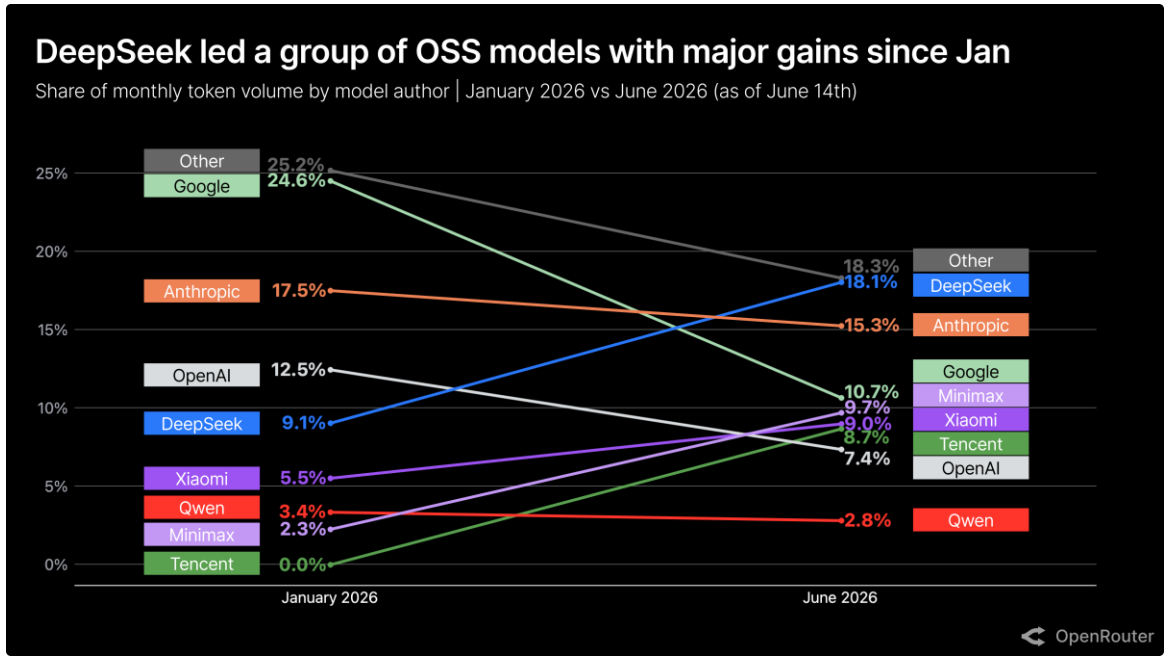


资料来源： 斯坦福《2026 AI Index Report》，Artificial Ananlysis，每日经济新闻公众号，浙商证券研究所整理

从开发者调用到企业付费，DeepSeek海外采用信号增强

- **DeepSeek在OpenRouter的Token份额半年翻倍，开发者调用明显上升。** OpenRouter是聚合多家模型及推理服务商的API路由平台，开发者可通过统一接口调用不同模型；2026年1月至6月，**DeepSeek在该平台的Token份额由约9%升至18%。**
- **DeepSeek登顶Ramp 6月软件趋势榜，开始显现美国企业直接付费采用信号。** Ramp是美国企业支出管理平台，其趋势榜基于5万余家企业的公司卡和账单支付数据，主要识别**当月新获企业采购的软件供应商**；DeepSeek位列榜首，说明其采用方式正在从开源部署延伸至直接付费。

DeepSeek Token份额半年翻倍，平台调用快速增长



DeepSeek登顶Ramp趋势榜，企业付费采用信号增强

Ramp: Top Software Vendors

June 2026

Trending – *breakout growth relative to size*	Fastest growing – *ranked growth in market share*
DeepSeek – *Foundational LLMs*	Anthropic – *Foundational LLMs*
PheedLoop – *Event Management*	Granola – *AI Notetaker*
Fireworks AI – *Model Serving & Inference*	Vercel – *Web Hosting & Site Builders*
Paper – *Software Design*	Netlify – *Web Hosting & Site Builders*
Marpipe – *Ad Creation*	Canva – *Content Creation*
fal AI – *Model Serving & Inference*	Grammarly – *Productivity AI*
DeepInfra – *Model Serving & Inference*	Figma – *Software Design*
Vast.ai – *GPU Cloud*	Perplexity AI – *Search & Research AI*
Sitebulb – *SEO*	Tailscale – *VPN/Networking*
DataForSEO – *SEO*	Higgsfield AI – *AI Video*

资料来源：openrouter，RampAI，浙商证券研究所整理

需求增长与低价基数形成提价基础，DeepSeek开启API涨价

- **以中国模型为代表的开源token价格在上涨，闭源模型token价格在下降。**
- **DeepSeek已启动API整体提价，初步兑现模型采用增长带来的定价空间。** 8月6日，DeepSeek发布公告称，计划近期整体上调API服务定价，**预计涨幅较大**，具体方案以正式通知为准。
- **产品升级、调用需求增长及资源压力共同构成此次涨价的基础。** 6月底，DeepSeek表示V4正式版将进一步优化功能与性能，并通过调整API价格、引入峰谷定价改善资源配置与服务稳定性，其中高峰时段调用价格将翻倍。

DeepSeek预告较大幅度涨价

用量信息

计划近期整体上调 DeepSeek API 服务的定价，预计涨幅较大，请合理安排您的使用。具体方案以正式通知为准。 知道了

DeepSeek曾在6月预告高峰时段API价格翻倍

deepseek-v4-pro

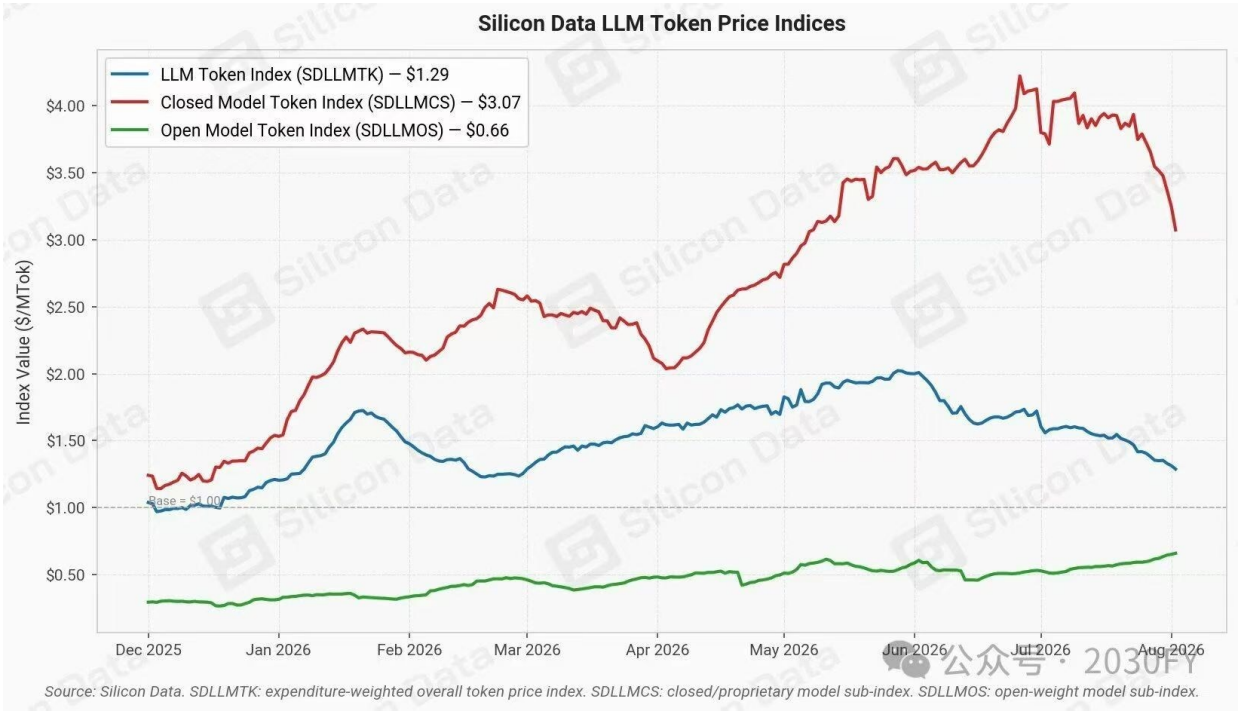
计费项	平时价格	高峰时段价格
百万tokens输入（缓存命中）	0.025 元	0.05 元
百万tokens输入（缓存未命中）	3 元	6 元
百万tokens输出	6 元	12 元

deepseek-v4-flash

计费项	平时价格	高峰时段价格
百万tokens输入（缓存命中）	0.02 元	0.04 元
百万tokens输入（缓存未命中）	1 元	2 元
百万tokens输出	2 元	4 元

** 高峰时段定义：每日 9:00～12:00 和 14:00～18:00（北京时间）

以中国模型为代表的开源token价格在上涨，闭源模型token价格在下降



Source: Silicon Data. SDLLMTK: expenditure-weighted overall token price index. SDLLMCS: closed/proprietary model sub-index. SDLLMOS: open-weight model sub-index.

资料来源：2030FY公众号，浙商证券研究所整理

资本支持：长期自有资金完成技术验证

- **DeepSeek长期自有资金完成技术验证，2026年首次引入外部资本，一轮融资即达高估值**
  - 7 月 16 日，据上市公司开润股份披露的投资进展公告，按公告中的投资金额和持股比例推算，DeepSeek 上轮融资完成后估值约为 **￥3510 亿**；
  - DeepSeek重启第二轮融资，拟融资约 **500 亿**人民币的，投前估值约 **5000 亿元**人民币；

DeepSeek　**长期自有资金完成技术验证，2026年首次引入外部资本**

已完成外部融资：1轮

2023—2025	2026.06 | 首轮完成	2026.08 | 重启第二轮融资
幻方内部孵化	**融资金额约500—510亿元**	**拟融资≥500亿元**
无外部融资 算力与研发由幻方支持	投后估值约3510 亿元 梁文锋200亿；腾讯100亿；宁德时代体系50亿	投前估值约5,000亿元

Kimi（月之暗面）　**连续融资推动模型、产品与市场扩张**

主要公开融资≥6轮

2023	2024.02	2024.08	2025.12	2026.02	2026.05	2026.06	2026.07
> 2亿美元	**> 10亿美元**	**> 3亿美元**	**5亿美元**	**> 7亿美元**	**约20亿美元**	**约20亿美元**	**暂未公开**
估值约3亿美元 红杉中国、真格	投后估值约25亿美元 阿里领投；红杉、美团	估值约33亿美元 腾讯、高榕、阿里	投后估值约43亿美元 IDG领投；阿里、腾讯	估值约100亿美元 阿里、腾讯、五源	投后估值约200亿美元 美团龙珠、中国移动、CPE	投前估值约315亿美元	Pro-IPO提前开启，估值500亿美元

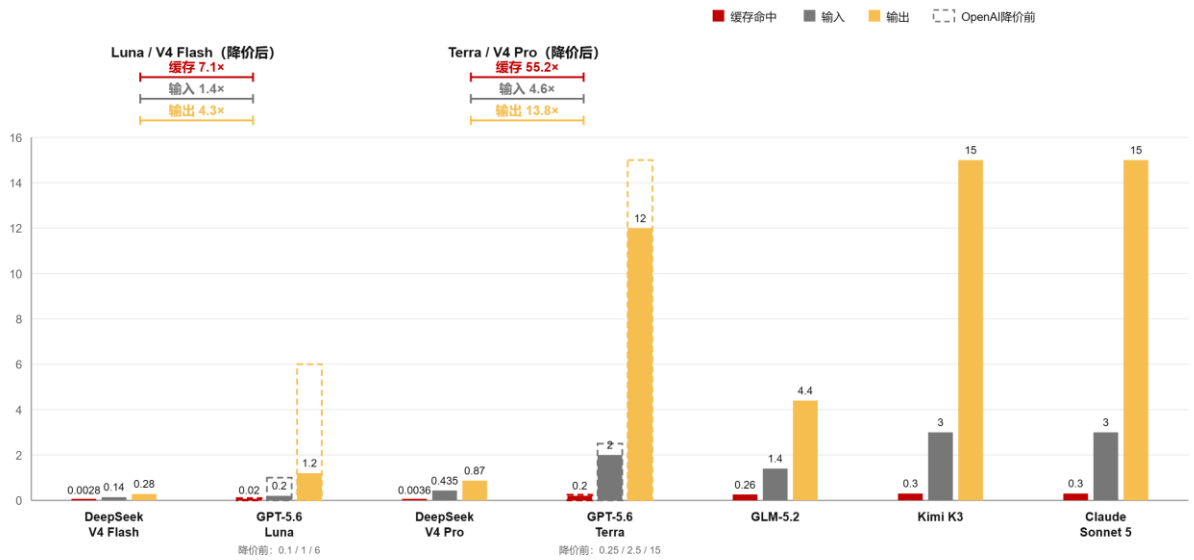
资料来源：第一财经，财经杂志公众号，36Kr，Bloomberg，晚点LatePost，科创板日报，每日经济新闻，浙商证券研究所整理

02 成本优势→认知优势

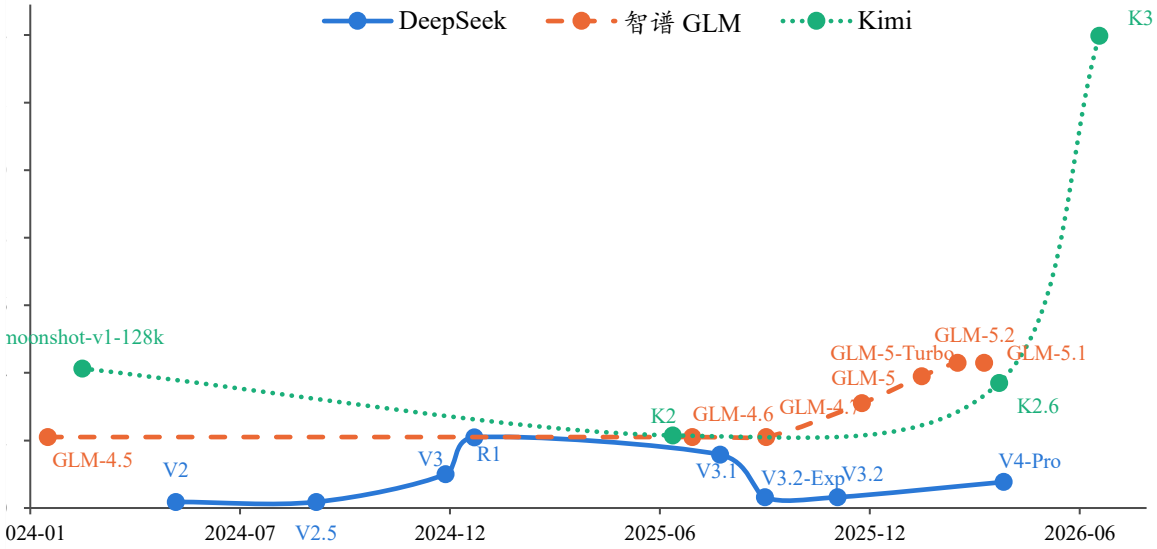
极致成本控制构筑定价底气

■ **DeepSeek 维持较低价格但宣布计划“近期整体上调DeepSeek API服务的定价，预计涨幅较大”；Kimi 与智谱均呈现一定涨价趋势**

当前API价格｜美元/百万token



DeepSeek价格始终维持低位（每百万token输出价格）

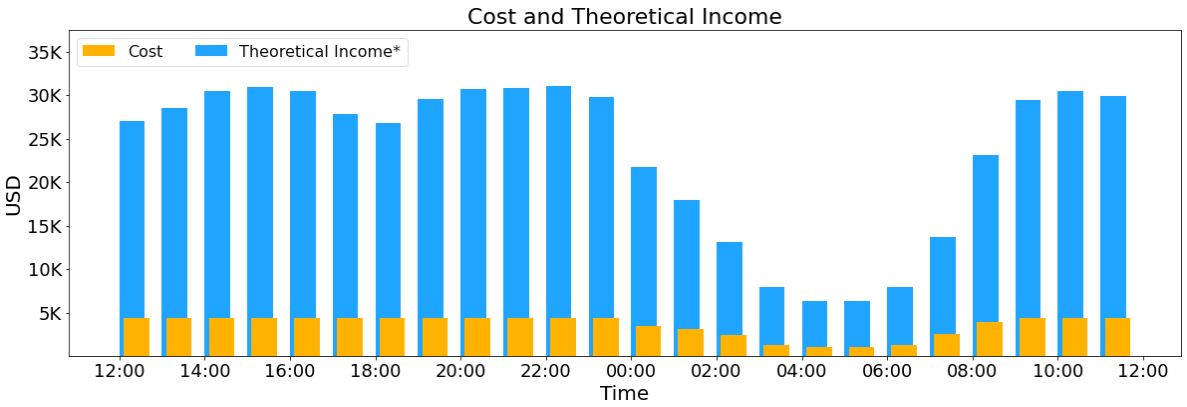


注：DeepSeek V4 将引入峰谷定价机制，高峰时段价格翻倍;官方宣布API价格将“大幅上涨”

■ **DeepSeek 维持低定价的底气来自于其极致的成本控制：**

- V3技术报告披露，**模型完整训练共消耗278.8万H800 GPU小时，按2美元/GPU小时计算仅约557.6万美元**，并在官方评测中达到与GPT-4o、Claude 3.5 Sonnet相近的能力水平。
- 据官方测算，2025年2月27日至28日，V3/R1服务的H800租赁成本为8.71万美元/日；假设全部6080亿输入Token和1680亿输出Token均按R1价格收费，每日总收入将达到 562,027 美元，成本利润率为 **545%**（*按“利润÷GPU租赁成本”计算；若换算成通常的“利润÷收入”口径约为**84.5%***）。

DeepSeek V3/R1在线推理的每小时成本与理论收入



* The theoretical income is calculated based on R1's standard API pricing, taking into account all tokens across web, APP, and API. It is not our actual income.

资料来源：Kimi官网，智谱官网，OpenAI官网，Anthropic官网，DeepSeek官网，浙商证券研究所整理

成本优势的来源：Infra“巨鲸”，持续推动MoE技术边界

■ **激活比例持续降低**

维度	DeepSeek V4-Pro	Kimi K3	DeepSeek V4-Flash	GLM-5.2	DeepSeek V3
发布时间	2026-04-24 (预览)	2026-07-16	2026-07-31 (正式版)	2026-06	2024-12-26
总参数	1.6T	2.8T	284B	744B	671B
激活参数	49B	104B	13B	40B	37B
激活率	**3.0%**	3.7%	**4.6%**	5.4%	**5.5%**

资料来源：DeepSeek官网，Kimi技术报告，Github，X.com，浙商证券研究所整理

成本优势的来源：Infra“巨鲸”，持续推动MoE技术边界

- **DeepSeek的模型进化围绕控制成本进行全栈优化，并将成本优势转化为价格优势**
  - **主要架构创新**：V3 的无辅助损失负载均衡 + 多 token 预测、V3.2-Exp 的 DSA 稀疏注意力、V4 的CSA+HCA混合注意力等；
  - **两次“效率创新 → 立刻降价”**：V3.2-Exp 因 DSA 提效而价格腰斩，V4 因混合注意力把 1M 上下文做便宜、随后 V4-Pro 折扣转永久。

- **DeepSeek将成本优化贯穿模型架构、训练框架与算力集群：围绕稀疏计算、长上下文、强化学习和基础设施等进行全栈优化。**

研究线	技术	旗舰集成	关键增量
稀疏模型	DeepSeekMoE	V2 / V3 / V4	从细粒度专家到无辅助损失均衡、专家波次
长上下文	MLA / DSA / CSA-HCA	V2 / V3.2 / V4	压KV → 稀疏选择 → 混合注意力；
强化学习	DeepSeekMath、GRPO	R1 / V4 specialists	去 critic → 规则奖励 → 多领域专家与蒸馏
系统基础设施	DualPipe、DeepEP、3FS	跨代开放项目	通信、算子、流水线和存储共同演进

2023

DeepSeek LLM 11月
67B Base 在推理、代码、数学、中文理解上超越 Llama2 70B Base

2024

DeepSeek-V2 5月
MoE 架构实现经济高效，性能反超此前的 DeepSeek 67B

DeepSeek-V3 12月
MLA + DeepSeekMoE，引入无辅助损失负载均衡与多 token 预测，实现高效推理与低成本训练

2025

V3-0324 3月
推理、Web 前端、中文写作与搜索、Function Calling 五项能力改进

V3.1 8月
混合推理架构，单模型支持思考 / 非思考双模式，被定位为迈向 Agent 时代第一步

V3.1-Terminus 9月
Code 与 Search Agent 增强，语言一致性修复；HLE 由 15.9% 升至 21.7%

V3.2-Exp 9月
引入 DSA 稀疏注意力优化长上下文训练与推理效率；性能略降，价格降过半

V3.2 / Speciale 12月
沿用 V3 架构达到 GPT-5 水平；Speciale 斩获 IMO、CMO、ICPC、IOI 四项金牌

2026

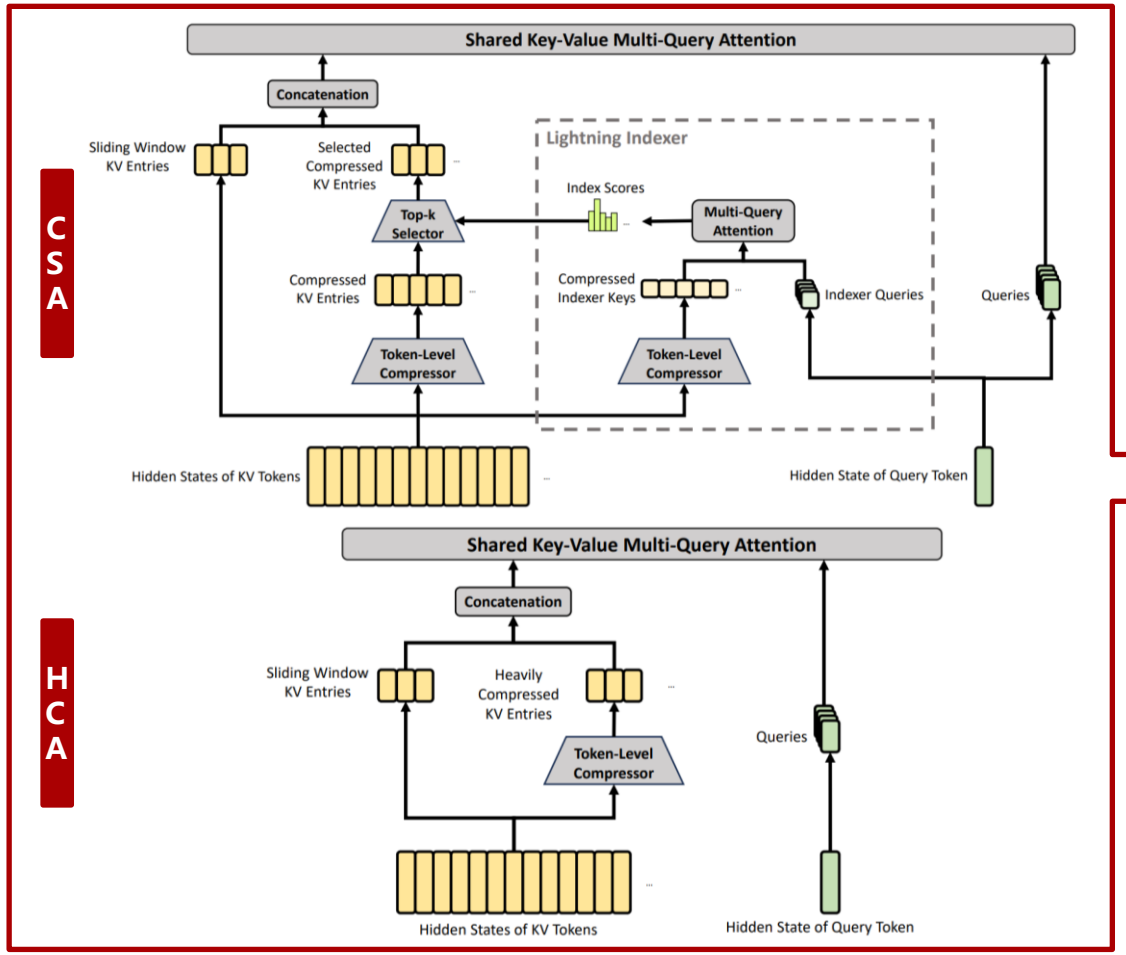
DeepSeek-V4 4月
混合注意力架构，1M 上下文下长文本推理效率大幅提升

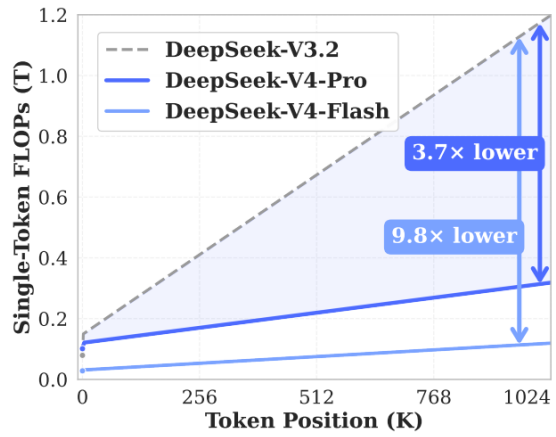
V4-Pro 降价 5月
折扣转永久价，约为 GPT-5.5 输出价 1/34，压低前沿模型 API 价格地板

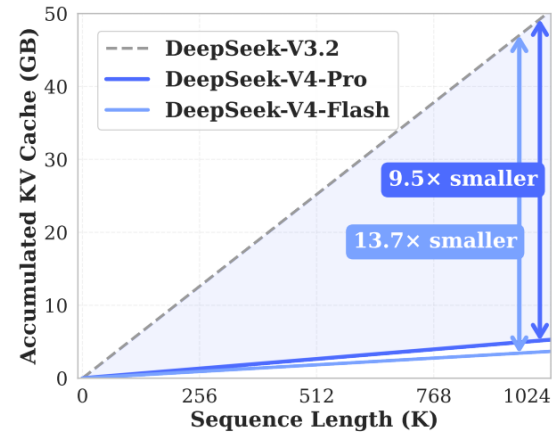
资料来源：FisherSama，DeepSeek技术论文，浙商证券研究所整理

成本优势的来源：Infra“巨鲸”，持续推动MoE技术边界

- **V4最重要的创新：CSA与HCA组成的混合注意力**
  - CSA先将每4个Token的KV压缩成一个，再通过稀疏索引选择最重要的历史信息；HCA进一步将每128个Token压缩为一个全局表示；同时保留滑动窗口注意力处理近期细节。
  - 配合KV Cache的BF16/FP8混合存储，以及索引器和专家权重的FP4量化，**V4显著降低了长上下文的计算量、显存占用和访存压力。1M 上下文场景下，与V3.2相比， V4-Pro 只需要 27% 的单 token 推理 FLOPs 和 10% 的 KV cache。**







在100万Token的上下文条件下，V3.2相比，V4-Pro的单Token推理计算量仅为其**27%**，KV Cache占用仅为其**10%**

资料来源：DeepSeek技术论文，浙商证券研究所整理

模型硬件Co-Design，扩编聚焦底层系统与Agent能力闭环

■ **人才布局覆盖全栈，重视Infra能力**

- **招聘需求贯穿算力底座、模型研发和产品应用全栈。** 招聘范围覆盖数据中心、IT基础设施、超算集群、算子与通信、训推框架、分布式存储、数据工程、模型研究及产品应用。
- **Infra能力是其工程招聘的重点。** 相关岗位从传统运维延伸至RDMA网络、异构算力调度、KV Cache存储、GPU/NPU算子、集合通信、编译器和Agent沙箱，并普遍要求计算机体系结构、操作系统、网络、分布式系统和性能优化能力。

■ **多类岗位聚焦Agent，形成跨部门人才需求**

- **Agent已成为横跨研究、工程、数据和产品部门的招聘主线。** 相关岗位包括Agent Harness研究与研发、Agent Infra、Agent后端、Code Agent数据工程、通用Agent数据产品经理和Agent产品经理等。
- **人才需求覆盖Agent从运行环境到训练评测和产品落地的完整链条。** 具体能力涉及上下文管理、长期记忆、工具调用、MCP、Agent沙箱和强化学习环境建设等，部分岗位还需利用真实任务、执行轨迹和产品反馈构建训练及评测数据。

DeepSeek招聘岗位（标亮职位为偏基础设施岗位）

岗位大类	招聘职位
全栈开发 / 算法	**服务端开发工程师** **预训练数据工程师** **AI搜索算法 / 架构工程师** **Agent Harness团队** **Agent Infra研发工程师** 前端 / 客户端开发工程师 测试开发工程师 AI跨界技术人才
AI核心系统研发	**超算集群研发工程师** **高性能算子 / 通信 / 编译器工程师** **大模型训练 / 推理框架工程师** **高性能分布式存储工程师**
运维	**AI算力集群性能与可靠性工程师** **AI平台运维工程师** **IT基础设施团队** **IDC数据中心团队**
产品部门	AI产品经理 AI产品运营（体验与服务方向）
模型数据策略产品经理 / 工程师	**Code Agent数据工程师** **通用Agent数据产品经理（办公 / 生活 / 搜索）** 专业领域数据产品经理（小语种、医学、法律等学科） AI创作数据产品经理 情感智能数据产品经理
深度学习研究员	Frontier研究员（持续学习 / 自进化 / 新范式） 预训练（数据 / 算法）研究员 后训练（数据 / 算法）研究员 多模态理解（数据 / 算法）研究员
职能部门	……

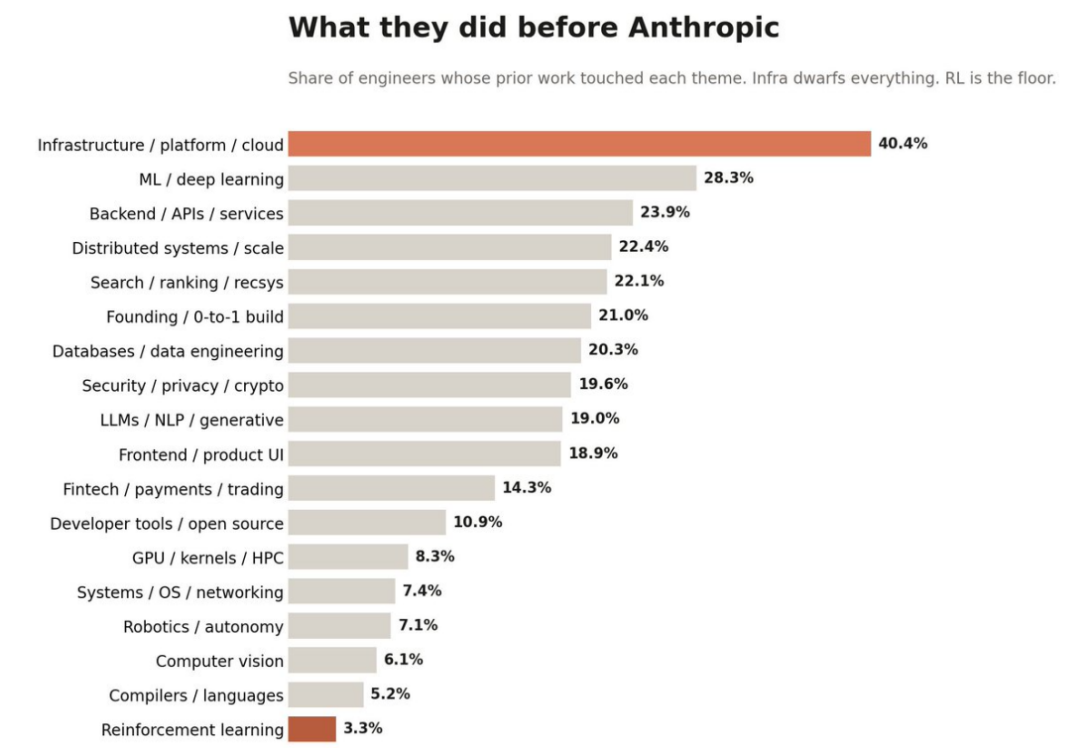
资料来源：DeepSeek公众号，浙商证券研究所整理

海外验证：Anthropic 40%的工程师背景与基础设施相关

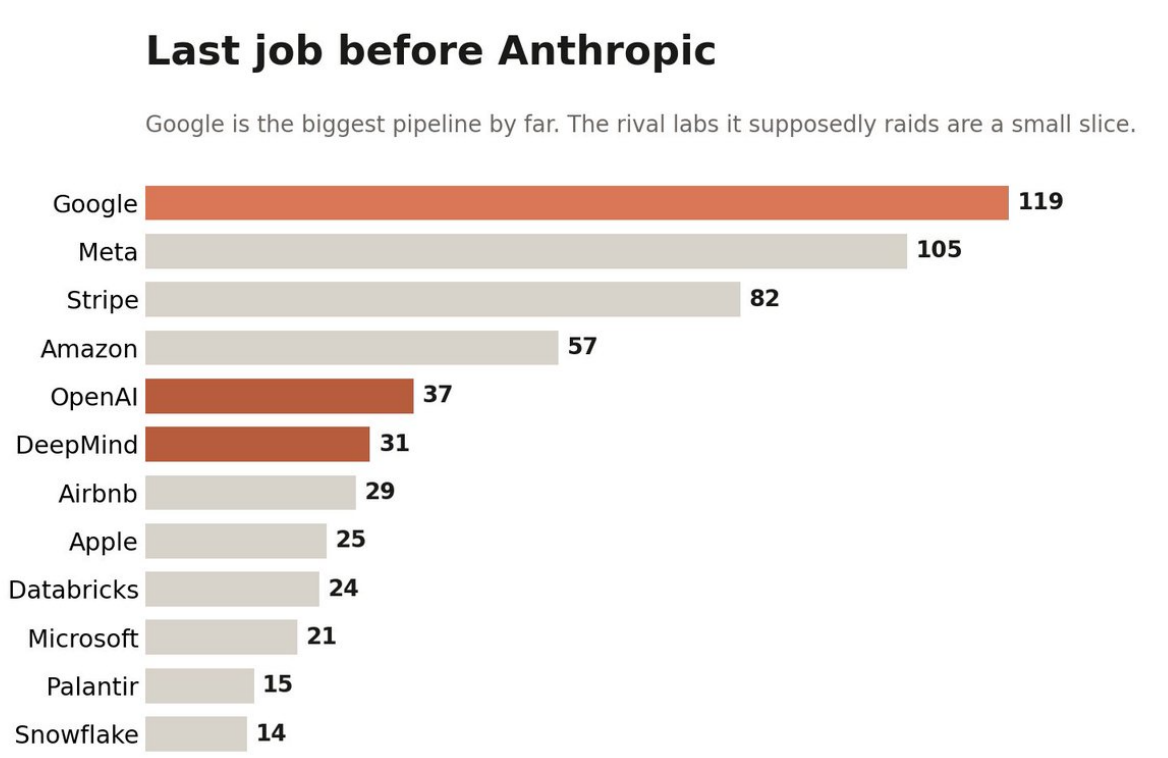
■ **Anthropic 当前工程团队的画像不是研究员密集型 AI 实验室，而更接近一家高速扩张的基础设施公司。**2026年6月，猎头 Sebastian Cuadros 爬取了所有在 LinkedIn 上将 Anthropic 列为当前雇主的人共 5,306 人，从中筛出 1,680 名真正从事工程岗位的员工，再分析这些人加入前的 7,986 段过往岗位描述。结论是：

- **40% 的工程师背景与基础设施相关：**后端开发、分布式系统、数据库、安全各占约 20%；强化学习只出现在 3.3% 的工程师履历中。典型的 Anthropic 工程师画像，过去十年是在**超大规模平台厂商**或**基础设施导向的创业公司**里构建大规模生产系统；
- **大量招聘来自以工程严谨性著称的公司的人：**Google、Stripe、Databricks、Snowflake、Palantir、Airbnb，**只有约 94 人是直接从前沿 AI 实验室跳槽而来。**

入职前具有基础设施经验的工程师，远多于具有强化学习经验的工程师



入职前具有基础设施经验的工程师，远多于具有强化学习经验的工程师



资料来源：Linkedin，浙商证券研究所整理

海外验证：Microsoft 模型、硬件 Co-design

- **构建让模型负载参与芯片定义、芯片再反过来定义模型架构的双向闭环**
  - **芯片未流片，模型负载已介入设计：**从 Maia 200 架构设计的最初阶段起，微软即通过成熟的流片前仿真环境，对大语言模型的计算与通信模式做高保真建模，并以此指导架构决策。其核心原则是：在芯片实际可用之前，尽可能提前完成端到端系统验证。
  - **芯片反向定义模型，能力验证从推理场延伸到训练场**
    - **推理：**GPT-5.2 以 FP4 在 Maia 200 上取得良好性能，证明微软具备在模型、芯片、内存、网络与整机之间做端到端优化的能力。（FY26Q2）
    - **训练：**Superintelligence 团队将用Maia 200**开展合成数据生成与强化学习，直接进入下一代自研模型的训练环节**；其架构设计有助于加速高质量领域数据的生成与筛选。（FY26Q2）
    - **未来：**自研模型团队与 Maia 团队反复协同，未来 MAI 模型均针对 Maia 优化，并统一以"面向企业场景的高性价比推理"为设计目标。（FY26Q4）
  - **可量化的效率：Copilot 中使用最广泛的模型推理吞吐提升 40%，**微软明确归因于数据中心、芯片、系统软件与模型架构的联合优化；（FY26Q3）MAI-Thinking-1 已针对 Maia 200 优化，端到端运行时在 Maia 硬件自身性能优势之上，再获 1.4 倍每瓦性能提升。（FY26Q4）

在Maia 200上运行MAI模型时，进一步实现了 1.4 倍每瓦性能提升

MAI-Thinking-1（MoE 架构, 1T/35B ）推理开销小于体量大得多的模型，却在 SWE-Bench Pro 取得与 Opus 4.6 相近结果

Category	Benchmark	MAI-Thinking-1	Sonnet 4.6	Opus 4.6	GPT 5.4	Kimi K2.6	Deep Seek V3.2	Deep Seek V4	GLM-5.1
STEM	AIME 2025	97.0	95.6	99.8	–	–	93.1	–	–
	AIME 2026	94.5	–	–	–	96.4	–	–	95.3
	HMMT Feb 2026	84.9	–	–	–	92.7	–	95.1	82.6
	GPQA Diamond	84.2	89.9	91.3	92.8	90.5	82.4	90.1	86.2
	LCB v6	87.7	–	–	–	89.6	83.3	93.5	–
Agentic Coding	Terminal Bench 2.0	46.0	59.1	65.4	75.1	66.7	46.4	67.9	69.0
	SWE-Bench Verified	73.5	79.6	80.8	–	80.2	73.1	80.6	-
	SWE-Bench Pro	52.8	–	53.4	57.7	58.6	–	55.4	58.4

资料来源：Microsoft官网，浙商证券研究所整理

低成本首先是认知优势，其次才是价格优势

- **当前模型规模是从资源倒推的，不是从需求正推的。**
- **低成本首先是认知优势，其次才是价格优势。**成本越低，同样算力就能做更多消融实验、生成更多Agent轨迹、运行更长的推理、容纳更多教师模型。所以DeepSeek降成本，V3的软硬件协同和V4的百万上下文效率，不只是为了卖得便宜，而是在提高自己的认知优势。

资料来源：公开资料，浙商证券研究所整理

03

代表中国
国产替代力量

供给约束 + 负载变化 → 把竞争推向系统层

- **供给侧**——须在可获得的芯片工艺上创造系统增量

供给侧：算力可持续必须建立在实际可获得的芯片工艺上

> *“中国半导体制造工艺将在相当长时间处于落后状态；**可持续的算力只能基于实际可获得的芯片制造工艺。**”*
>
> — 徐直军，华为全联接大会2025

现实起点	可获得的制造工艺 约束单芯片演进速度
↓	
行业催化	AI需求仍持续增长 单颗芯片无法独立承载
↓	
竞争单位	算力、内存、互联、软件 必须联合优化
↓	
架构答案	**“超节点 + 集群” 把增量转向系统层**

英伟达和昇腾芯片指标对比

时间	品牌	芯片型号	FP8算力	FP4算力	FP16算力	内存容量	内存带宽
2025 Q1	Ascend	910C	—	—	800 TFLOPS	128GB	3.2TB/s
2024 Q4—2025	NVIDIA	B200	4.5 PFLOPS	9 PFLOPS	4.5 PFLOPS	192GB HBM3e	8TB/s
2026 Q1	Ascend	950PR	1 PFLOPS	2 PFLOPS	—	128GB	1.6TB/s
2025 H2—2026	NVIDIA	B300	7 PFLOPS	15 PFLOPS	4.5 PFLOPS	288GB HBM3e	8TB/s
2026 Q4	Ascend	950DT	1 PFLOPS	2 PFLOPS	—	144GB	4TB/s
2026 H2	NVIDIA	Rubin R100	16 PFLOPS	50 PFLOPS	8 PFLOPS	288GB HBM4	22TB/s
2027 Q4	Ascend	960	2 PFLOPS	4 PFLOPS	—	288GB	9.6TB/s
2027 H2	NVIDIA	Rubin Ultra	约32 PFLOPS	100 PFLOPS	约16 PFLOPS	384GB HBM4e	约32TB/s
2028 Q4	Ascend	970	4 PFLOPS	8 PFLOPS	—	288GB	14.4TB/s

资料来源：华为全联接大会2025徐直军主题演讲，英伟达官网，浙商证券研究所整理

供给约束 + 负载变化 → 把竞争推向系统层

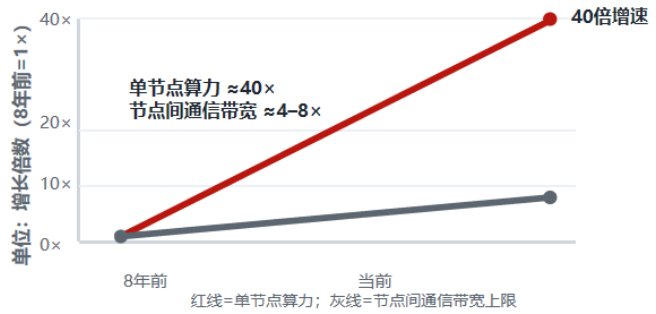
- **负载侧**——让通信、访存和资源编排成为Token产出的共同瓶颈

中兴训练结果：同样约2000卡，单卡训练性能指数随超节点规模提升

■ 约2000卡仿真中，256卡超节点单卡训练性能较8卡服务器高15%。

■ 64/128/256卡收益趋平，说明不是无限堆卡，而是寻找模型并行、互联成本和规模效率的最优Scale-Up域。

8年变化：算力增长远快于通信，系统瓶颈从单芯片转向多芯片协同



■ 近8年单节点算力约增长40倍，而节点间通信带宽仅增长约4-8倍。

■ 当模型被切到多卡后，All-Reduce、All-to-All、KV Cache搬运会把等待时间放大。

■ 传统Scale-Out更适合松耦合任务；大模型并行训练/推理要求更接近“单机内部总线”的协同效率。

负载侧：MoE、长上下文与P/D分化改变资源需求

MoE **总参数 → 单Token激活参数**

稀疏激活降低计算量，却增加专家路由、同步与All-to-All通信

长上下文 **KV Cache容量 ∝ 并发用户数**

上下文扩展同时抬升Attention计算量与KV缓存容量，**数据放置和访问方式成为系统约束**

P / D **同一模型的不同阶段，资源瓶颈方向相反**

- **Prefill** 需并行处理输入序列，核心约束更多来自**计算能力**；
- **Decode** 则以自回归方式逐Token生成，对**内存带宽和互联能力**更为敏感。

若仍沿用CPU、NPU、内存与网络资源固定绑定的节点架构，则计算密集阶段可能受制于算力不足，而生成阶段则可能受制于访存与通信。

超节点的价值正在于扩大资源协同边界，使计算、内存与互联能够围绕实时负载进行更高效的统一调度。

有效Token产出并非由单一算力指标决定，而是**算力、内存容量与带宽、互联、资源编排及可靠性**共同作用的**系统工程问题**

资料来源：中兴通讯超节点白皮书，浙商证券研究所整理

DeepSeek明确释放国产算力相关人才需求

■ **招聘需求涉及国产算力适配，相关人才需求覆盖集群、算子、框架与运维等环节**

- **官网现有4个岗位明确提出国产算力相关要求。** 相关岗位覆盖超算集群、算子与通信、训推框架及基础设施运维。
- **国产算力人才需求已覆盖硬件适配、软件开发和集群运维环节。** 具体要求包括NPU资源抽象与调度、国产芯片体系结构分析、Ascend C算子开发、HCCL通信优化以及Ascend集群运维和压测。

国产算力人才需求覆盖集群、算子、框架与运维全链条

招聘岗位	核心职责	国产算力相关要求
超算集群研发工程师	建设**异构算力集群**的调度、训练基础设施和高性能网络，推动集群架构与模型训推需求协同优化	负责CPU/GPU/**NPU**资源的抽象、池化和拓扑感知；熟悉**国产AI加速器**或**国产CPU**的体系结构和性能特性者加分
高性能算子/通信/编译器工程师	开发并优化GPU/**NPU**计算算子、集合通信和深度学习DSL编译器，支持**新模型与不同硬件架构的协同设计**	明确涉及**Ascend NPU、Ascend C、Ascend C API和HCCL**；同时要求TileLang等具有Ascend**适配生态的DSL及编译技术**
大模型训练/推理框架工程师	建设分布式训练、推理、强化学习和多模态系统，并优化KV Cache、负载均衡和训练效率	将**Ascend C**与CUDA、**Triton、TileLang**并列列为算子开发加分项；TileLang和Triton均已有面向Ascend NPU的适配项目
IT基础设施团队	负责GPU服务器、网络和硬件配置管理、基准压测、故障处置及基础设施自动化	具备大规模GPU集群（NVIDIA/**Ascend**）运维或压测经验者加分

资料来源：公司官网，浙商证券研究所整理

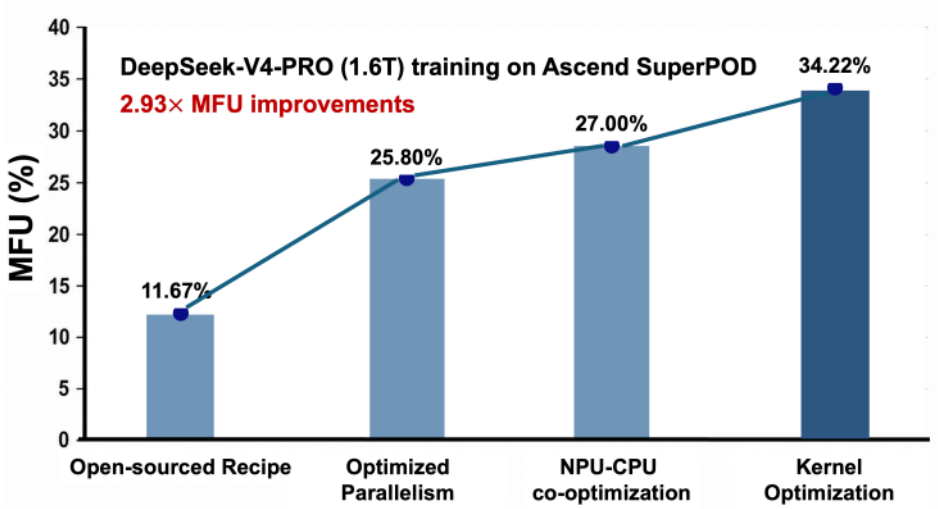
国产算力对DeepSeek的适配已延伸至全参数后训练

- **DeepSeek的架构创新使算力竞争从单卡峰值转向系统协同。** MoE、压缩注意力及FP8/FP4显著降低了单Token计算量与KV Cache占用，同时对专家All-to-All通信、统一内存、访存带宽、低时延互联和定制算子提出更高要求。
- **昇腾超节点与CANN已围绕DeepSeek V4形成软硬件协同适配。** 昇腾通过超节点将大规模NPU组织为统一计算系统，并围绕并行策略、计算通信重叠、内存与缓存管理以及低精度算子，适配V4的万亿参数、百万上下文和稀疏专家架构。
- **千卡级国产集群已验证DeepSeek V4-Pro的全参数后训练能力。** 深圳河套学院联合团队基于约1000颗昇腾910C，完成V4-Pro全参数后训练；系统优化后MFU达到34.22%，较开源基线提升2.93倍，表明国产算力正由推理部署进一步延伸至万亿模型训练。

模型架构创新将算力瓶颈推向通信、存储与算子协同

模型架构变化	系统影响	关键算力需求	昇腾适配情况
DeepSeekMoE：V4-Pro参数为1.6T/49B激活	每个token的实际计算量	专家权重的总内存容量；跨卡专家并行；高频All-to-All；小包通信低时延	384/950超节点统一内存编址；灵衢高速互联；扩大单一高速互联域
MLA、CSA、HCA及1M上下文	KV Cache和长上下文FLOPs	高内存带宽；KV Cache池化；稀疏Top-K与随机访存；上下文并行	48TB/256TB统一内存空间；V4专用缓存管理；TopK、SWA、CFA等融合算子
V4专家FP4＋其他FP8	权重、激活、通信量和计算成本	精确的低精度累加；细粒度缩放；在线量化；原生FP4/FP8算力	昇腾950支持FP8、MXFP8、MXFP4；CANN提供算子与编译支持
mHC超连接	提升深层模型信号传播稳定性	增加激活显存、流水线通信和特殊矩阵算子	超节点内存池、通信重叠、定制融合算子；昇腾上已有V4后训练实践
Agent与长思维链推理	提升复杂任务能力	超长Prefill、高频KV读写、持续Decode、高并发和低时延	昇腾950PR面向Prefill，950DT面向Decode/训练
万亿模型长时间训练	获得更大模型能力	集群可靠性、故障恢复、有效算力、存储与检查点	超节点减少网络层级，灵衢提供故障检测、切换及统一调度

全栈优化使V4-Pro训练MFU提升2.93倍至34.22%



注：效率提升主要来自并行策略、NPU-CPU协同及Kernel优化

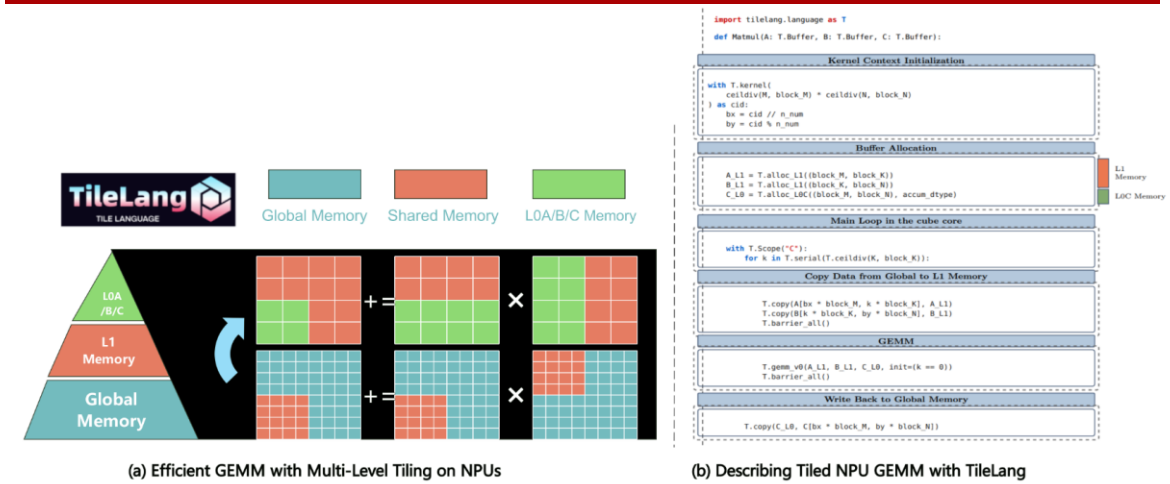
资料来源：深圳河套学院，深圳发布公众号，技术论文，浙商证券研究所整理

开放全栈基础设施，逐层降低CUDA生态锁定

- **DeepSeek 在软件层逐步与CUDA生态解耦：**
  - **数值格式上：**V3.1 率先采用 UE8M0 FP8 Scale，官方说明其**为“即将发布的下一代国产芯片” 设计**，同时保持与 NVIDIA **MXFP8** 兼容 。 模型不再被动等待硬件，而是提前定义标准让硬件对齐；
  - **算子表达上：**V3.2-Exp 将 DSA 算子以 **TileLang 与 CUDA 双版本同时开源**，并明确推荐研究场景优先使用 TileLang，华为昇腾同日实现 Day0 支持并开源全部推理代码与算子实现；**V4采用 TileLang 开发大量融合内核。**华为CANN已明确支持Triton、**TileLang**、PyTorch、vLLM、verl等业界开源社区和开源项目。
- **国产算力的门槛从“重写一套 CUDA” 变成了“接入一层统一算子” ，生态壁垒被逐渐拉平 。**

	CUDA	Triton	TileLang
归属	NVIDIA 专有	OpenAI 开源	北大杨智团队 + 微软亚研，2025-01 开源
硬件覆盖	仅 NVIDIA	为NVIDIA / AMD	NVIDIA CUDA 、 昇腾、摩尔线程等
开发门槛	高	中	类 Python，三类接口降低编程门槛

TileLang 支持Huawei AscendC和Ascend NPU IR后端

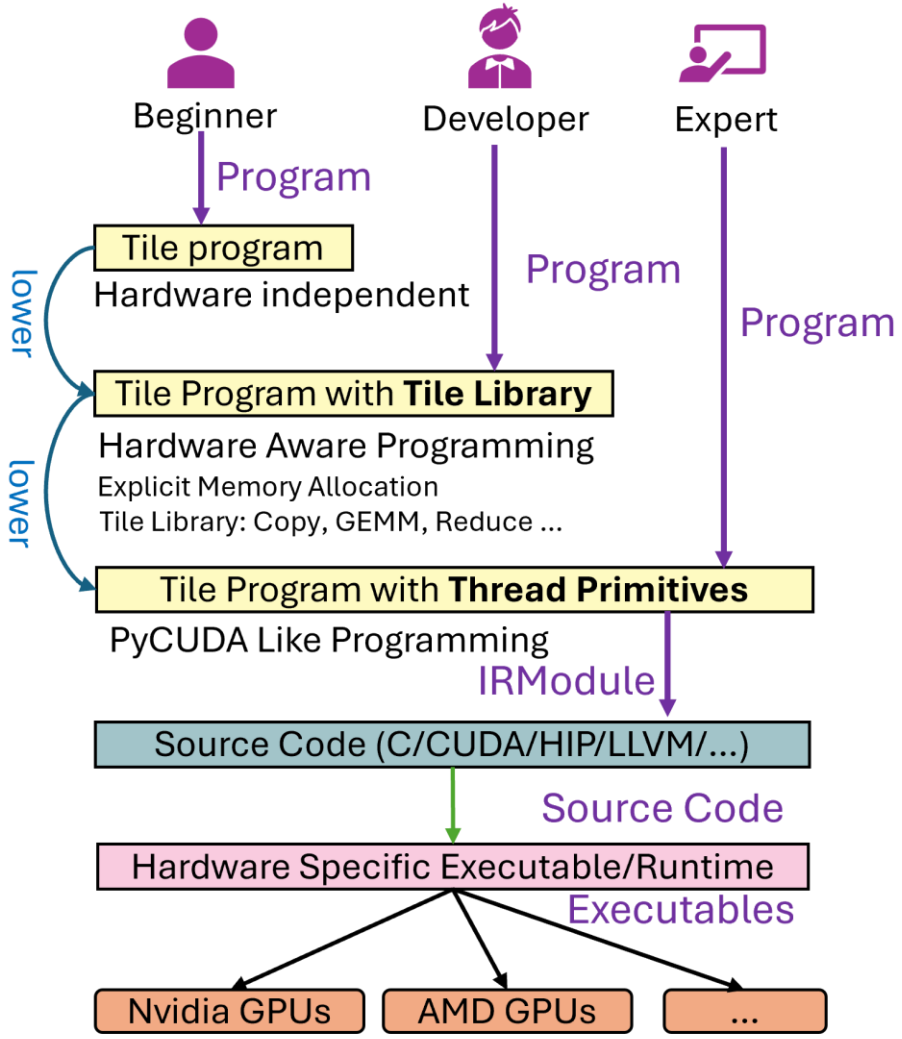


(a) Efficient GEMM with Multi-Level Tiling on NPUs　(b) Describing Tiled NPU GEMM with TileLang

资料来源：Github，IT之家，DeepSeek官网，界面新闻，DeepSeek技术论文，华为官网，浙商证券研究所整理

开放全栈基础设施，逐层降低CUDA生态锁定

- **TileLang是一种专门用于开发GPU内核领域的专用语言，以“让AI算子开发更容易”为设计初衷**：主要设计理念为Tile级语言抽象，通过将数据分块处理（Tiling）实现高效内存管理和计算调度。
- **TileLang可大幅降低GPU编程的技术门槛**：提供三类不同层级的编程接口，分别面向初学者、开发者和专家用户。这些接口位于编译降低流程的不同层级。Tile 语言还允许在同一个 Kernel 中混合使用不同层级的接口，用户可以根据实际需求选择。
- **TileLang性能上可对标英伟达CUDA**：**DeepSeek V3.2** 技术论文曾明确推荐使用此版本做实验，在方便调试和快速迭代上有优势；**DeepSeek V4** 采用TileLang开发了一组融合内核以替代其中绝大多数运算符，将碎片化的小kernel融成大块，调用开销从百微秒压到1微秒以内。
- **TileLang与国产算力生态适配**：
  - **摩尔线程**：2026 年 1 月 30 日宣布并开源 TileLang-MUSA 项目，实现对其全功能 GPU（MTT S5000/S4000）的完整支持。
  - **沐曦**：2026 年 4 月 27 日，TileLang-Metax 正式成为 TileAI 社区官方主线版本，代码开源并托管在 TileAI 组织之下；**建立了与社区主线的周级持续同步机制**。模力方舟平台已基于沐曦曦云 C 系列 GPU 提供预置 TileLang 镜像的在线环境。
  - **华为昇腾**：TileLang 已实现将高级数据流描述自动转换并优化为 Ascend C 代码（昇腾的算子编程语言），并在 2025 年 9 月华为全联接大会上作为鲲鹏昇腾生态创新成果展出。在 DeepSeek 采用 TileLang 后，昇腾第一时间公告了对 TileLang 的支持。



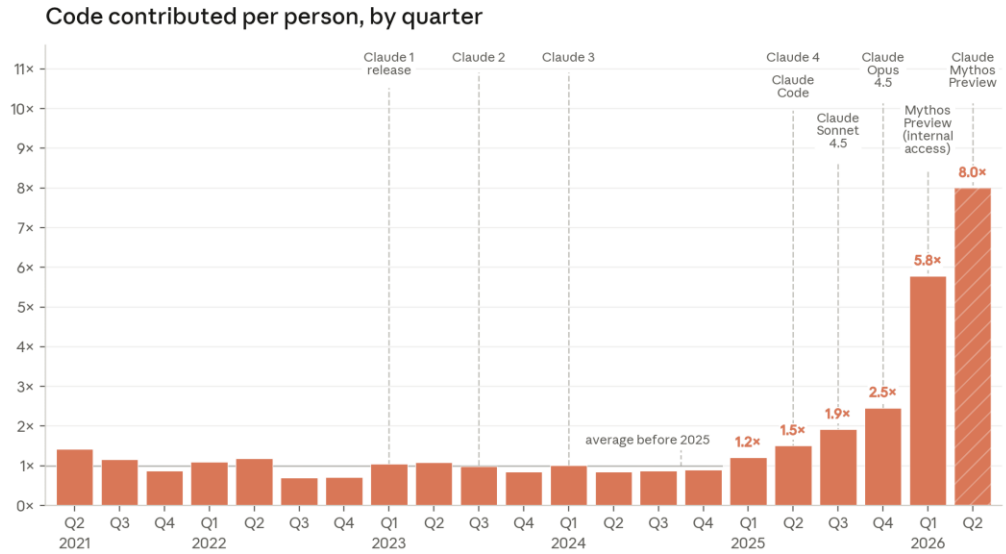
TileLang可大幅降低了GPU编程的技术门槛

资料来源：DeepSeek官网，华为官网，DeepSeek技术论文，昇腾CANN公众号，Tilelang官网，摩尔线程开发者公众号，新浪新闻，浙商证券研究所整理

AI正从辅助研发走向参与自身优化，持续学习闭环初步显现

- **Anthropic：《When AI Builds Itself》**
  - **AI加速AI研发已经成为可量化的事实。** 截至2026年5月，Anthropic合入生产代码中超过80%由Claude编写，人均季度代码产出较2024年提升约8倍；Claude完成开放式工程任务的成功率半年内由约26%升至76%。
  - **模型参与训练系统优化的能力也在快速提升。** 在固定的训练代码提速实验中，模型实现的加速幅度一年内由约3倍提升至约52倍，明显超过熟练工程师在相近任务中的优化效率。
- **OpenAI：模型开始参与优化自身推理系统**
  - **AI参与研发已从代码生产延伸至模型自身的运行效率优化。** 在人工设定目标和验证边界的流程中，GPT-5.6 Sol自主重写生产GPU Kernel、设计和执行数百项实验，并参与推测解码模型的训练监控。
  - **模型自优化已带来可量化的成本与价格优势。** 相关优化使OpenAI端到端模型服务成本下降20%、Token生成效率提升15%以上；随后Luna和Terra的API价格分别下调80%和20%。
- **持续学习的可能路径持续学习的核心，是让模型在真实研发环境中形成“执行—反馈—优化”的长期闭环。** 随着上下文、工具调用和结果验证机制逐步完善，模型有望从完成单次任务，转向持续积累经验、改进运行方法并参与后续模型迭代；现阶段这一过程仍依赖人类设定目标、提供基础设施并验证结果。

Claude承担超80%合入代码，AI研发效率进入加速阶段



GPT-5.6 Sol参与优化自身推理系统

利用 GPT-5.6 Sol 加速推理

在算力受限、模型需求增长快于容量扩充的环境中，效率是每项系统设计的核心。对于运行已训练模型来生成响应的推理技术栈而言，尤其如此。我们的首要目标是在相同硬件上处理更多 Token，同时保持用户期望的智能、延迟、可用性和可靠性。

要实现这一目标，必须优化整个系统。模型本身可以非常高效，但如果请求分配不当、硬件闲置或数据移动拖慢计算，服务成本仍然可能很高。各层改进会产生叠加效应，收益来自路由（将请求发送到何处）、调度（何时发送请求）、内核（在 GPU 上运行的软件）、缓存（保存并复用的工作）和模型实现（GPU 代码的执行顺序）等方面的优化。

Codex 中的 GPT-5.6 Sol 在所有这些优化中都发挥了关键作用。

资料来源：Anthropic官网，OpenAI官网，浙商证券研究所整理

总结：以技术降本为核心，构建国产算力与全球商业化闭环

- **海外调用与企业采购共同验证商业化需求**
  - **DeepSeek的海外影响力正在由开源传播转化为实际调用和企业付费。** OpenRouter上的Token份额快速提升，Ramp企业软件趋势榜亦反映出美国企业直接采购信号；开发者调用与企业支付两类数据相互印证，其海外商业化进度领先多数国产开源模型。
- **技术降本支撑极致性价比，并形成较大的定价弹性**
  - **DeepSeek的低价优势建立在架构与工程降本之上，并非单纯依靠亏损获客。** 极低激活率MoE、低精度计算及持续工程优化压缩了训练和推理成本；即使OpenAI大幅降价，DeepSeek仍保持显著价格优势。
  - **当前较低的价格基数也为峰谷定价、模型分层及高端模型差异化收费预留了空间。**
- **软硬件CoDesign推动技术影响力在国内加速变现**
  - **DeepSeek正通过软硬件协同与国产算力形成更长期的生态绑定。** 公司将能力延伸至数据中心和基础设施环节，并围绕国产芯片推进集群调度、集合通信、编译器及算子适配。相较通用硬件适配，这一路径更有利于其技术影响力在中国转化为算力需求和商业价值。
- **低激活MoE与国产超节点形成较高的架构匹配度**
  - **模型规模扩张将进一步放大国产超节点的部署价值。** MoE架构在降低单Token实际计算压力的同时，提高了对多卡通信、分布式调度和统一内存的要求，与国产芯片及多卡超节点的系统优势较为契合，也提升了国产算力承载超大模型的可行性。
- **投资建议：关注国产算力超节点相关标的，以及受益于开源模型的部署的云厂商。建议关注：中芯国际、锐捷网络、紫光股份、阿里巴巴等。**
- **风险提示：** 技术路线突破不及预期，商业化与盈利不及预期，国产算力适配不及预期，行业竞争加剧与外部政策不确定性。

风险提示

1. **技术路线突破不及预期。** 持续学习等关键能力仍处于探索阶段，公司自身亦表示尚未做通；同时稀疏注意力、KV 压缩等架构方向已被行业广泛跟进，现有效率优势存在被追平的可能。

2. **商业化与盈利不及预期。** 收入与毛利率数据多源自第三方报道或理论测算，与实际经营口径存在差异；产品化保持克制，在订阅与企业服务市场的拓展节奏存在不确定性。

3. **国产算力适配不及预期。** 当前深度适配集中于推理侧，训练侧仍依赖海外算力；国产芯片在单卡性能、精度支持与产能爬坡上均存约束，规模化复制进度有待观察。

4. **行业竞争加剧与外部政策不确定性。** 头部厂商价格与能力竞争持续升级，可能压缩性价比优势；海外拓展面临数据合规、出口管制等政策变化风险。

行业评级与免责声明

行业的投资评级

以报告日后的6个月内，行业指数相对于沪深300指数的涨跌幅为标准，定义如下：

1、看好 ：行业指数相对于沪深300指数表现＋10%以上；

2、中性 ：行业指数相对于沪深300指数表现－10%～＋10%以上；

3、看淡 ：行业指数相对于沪深300指数表现－10%以下。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重。

建议：投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者不应仅仅依靠投资评级来推断结论

行业评级与免责声明

法律声明及风险提示

本报告由浙商证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，经营许可证编号为：Z39833000）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但浙商证券股份有限公司及其关联机构（以下统称“本公司”）对这些信息的真实性、准确性及完整性不作任何保证，也不保证所包含的信息和建议不发生任何变更。本公司没有将变更的信息和建议向报告所有接收者进行更新的义务。

本报告仅供本公司的客户作参考之用。本公司不会因接收人收到本报告而视其为本公司的当然客户。

本报告仅反映报告作者的出具日的观点和判断，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。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本公司的交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理公司、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

联系方式

浙商证券研究所

上海总部地址：杨高南路729号陆家嘴世纪金融广场1号楼25层

北京地址：北京市东城区朝阳门北大街8号富华大厦E座4层

深圳地址：广东省深圳市福田区广电金融中心33层

邮政编码：200127

电话：(8621)80108518

传真：(8621)80106010

浙商证券研究所：http://research.stocke.com.cn