

人工智能行业专题（20）：从开源模型崛起，探讨模型公司和云厂商业化格局

行业研究 · 行业专题

互联网 · 互联网 II

投资评级：优于大市（维持）

证券分析师：张伦可

0755-81982651

zhanglunke@guosen.com.cn

S0980521120004

证券分析师：张昊晨

zhanghaochen1@guosen.com.cn

S0980525010001

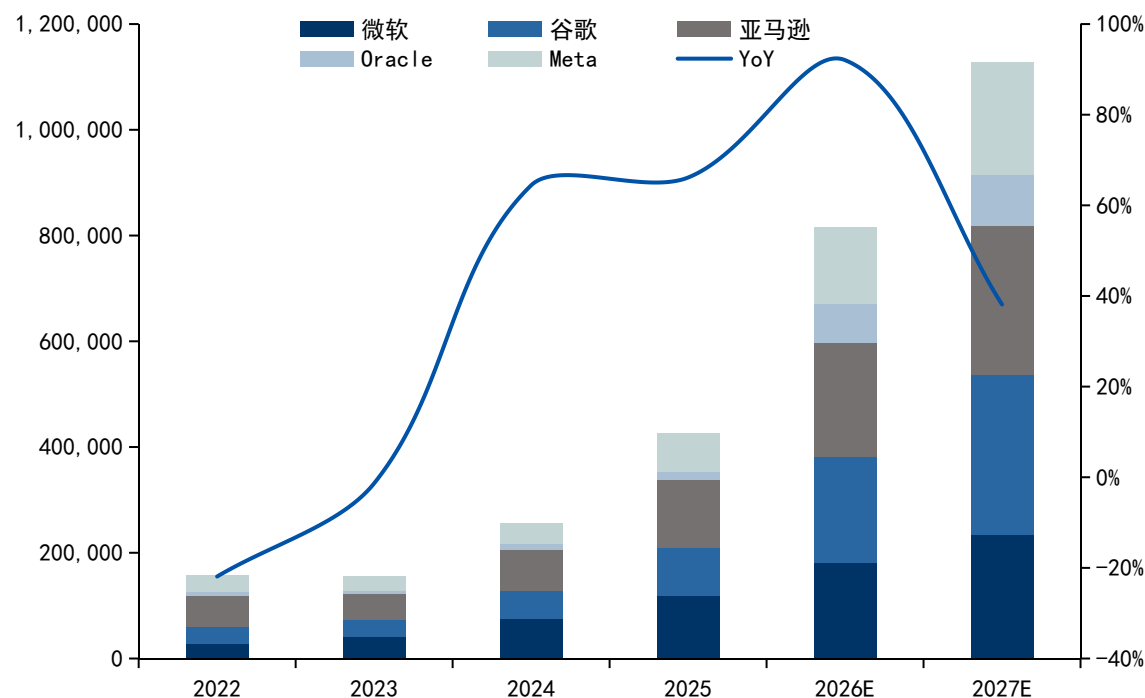
- 美国在大模型能力、基础设施投入和商业化规模方面目前相对领先，头部厂商最新资本开支已超过经营现金流。预计海外巨头2026年资本开支约8000亿美元，2027年超过1万亿美元；资本开支占经营性现金流的比例约为120%，占净利润的比例亦超过100%。国内整体起步较晚，但随着腾讯等厂商加大投入，预计2026年国内头部厂商资本开支合计亦将超过经营性现金流。资本开支持续高企推动外部融资需求上升；以谷歌为例，6月公告拟融资规模超过800亿美元，累计融资规模超过1600亿美元。
- 需求端，国产模型正迈过能力与经济性双重拐点。年初以来，海外模型能力提升推动商业化提速，根据Yipit数据，Anthropic与OpenAI的ARR合计已超过1000亿美元。随着GLM-5.2、Kimi K3、Qwen 3.8等模型相继发布，国产模型逐步具备生产级可用能力；叠加开源模型在Token及单任务成本方面的显著优势，国产模型商业化有望加速，并对海外闭源模型的高定价模式形成挑战。
- 资金端，国内模型厂商融资规模持续扩大，为研发、训练推理基础设施及商业化拓展提供资金保障。根据IT桔子及各公司公告整理，2025年以来国产AI模型创业公司的融资规模已超过1500亿元。基于融资规模静态测算各公司能够获取的算力规模以及对应的ARR水平，按30个月内月度支出线性递增测算，预计到27年6月五家公司单月支出约104亿元，假设推理和训练支出1:1分配，假设推理业务毛利率40%，对应当月推理收入约86亿元，ARR约1036亿元（若模型公司后续获得新增融资、资金使用节奏发生变化，或算力租赁价格、业务成本结构及毛利率水平出现较大波动，实际等效算力规模及ARR可能与测算结果存在偏差）。
- 我们测算1GW数据中心及模型厂商业模式，得到五点结论：①模型参数量、注意力机制的选择最终影响模型的成本和定价；②国产模型毛利率低于海外一方面是因为定价策略，更重要因素是Infra能力差距，包括推理中模型支持的精度，以及国产芯片及超节点整体硬件基础设施的利用率低于海外。③价格战会影响模型厂的利润率，并且如果模型厂为保持毛利率，还会进一步导致云厂收入减少，利润率空间受到压缩。④随着芯片性能的升级，相同模型的毛利率持续提升。⑤考虑到闭源模型的分成成本，开源模型有望帮助海外CSP实现更好的利润率。
- 投资建议：伴随开源模型性能崛起，云厂有望通过开源模型实现更好利润率提升ROI表现，关注亚马逊、微软、谷歌、阿里巴巴等。

- [01] 中美AI投入规模及进展
- [02] 国产模型性价比优势凸显，融资规模持续扩大
- [03] 模型商业化闭环与盈利能力测算
- [04] 国内外云厂收入结构拆分
- [05] 风险提示

AI基础设施需求持续释放，中美巨头资本开支保持增长

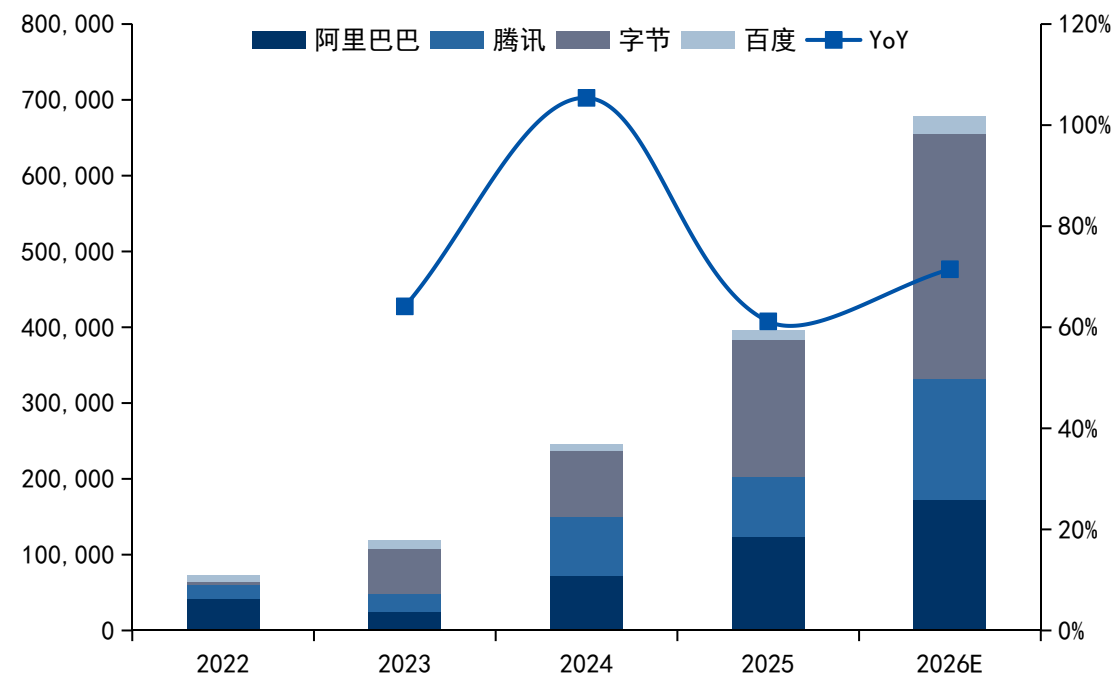
- AI基础设施建设持续提速，中美科技巨头资本开支同步上行，海外投入规模领先，国内厂商增长弹性更高。海外方面，微软、谷歌、亚马逊、Oracle和Meta的资本开支合计预计26年达到接近8000亿美元，27年超过1万亿美元，当前海外厂商仍处于AI基础设施集中建设阶段，预计资本开支将继续维持高位。国内方面，字节跳动投入规模领先，阿里巴巴、腾讯等头部厂商持续加码。

图：美国巨头CAPEX变化趋势（单位：百万美元）



资料来源：Bloomberg、国信证券经济研究所整理（截至2026/8/2）

图：中国巨头CAPEX变化趋势（单位：百万人民币）

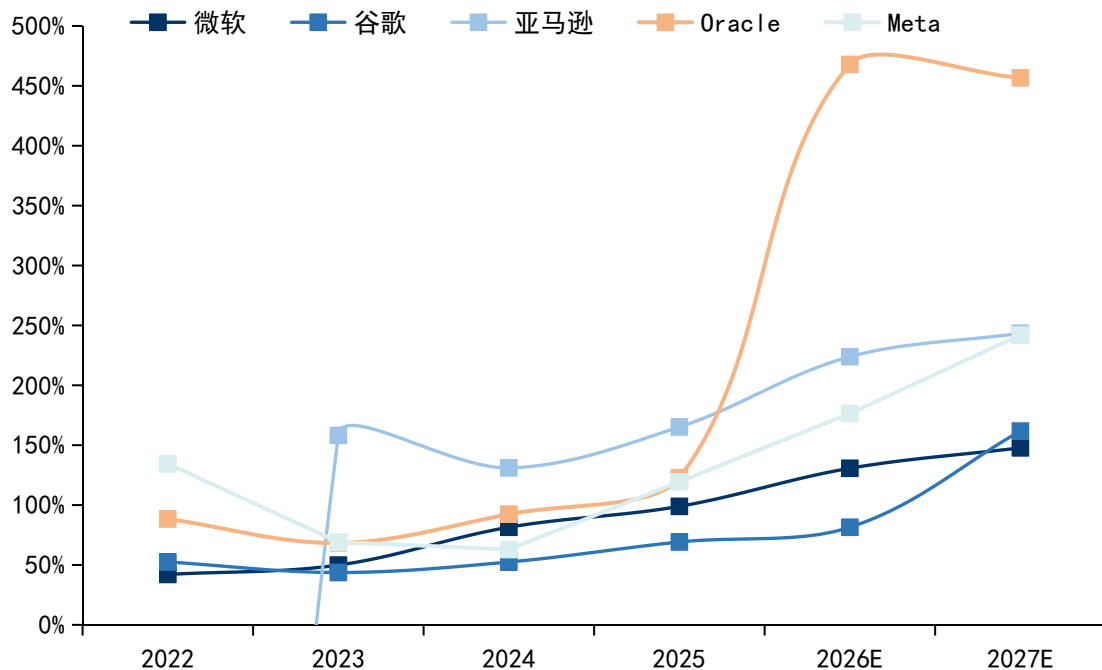


资料来源：Bloomberg、国信证券经济研究所整理预测

海外厂商维持高强度资本投入，国内头部企业投入力度快速抬升

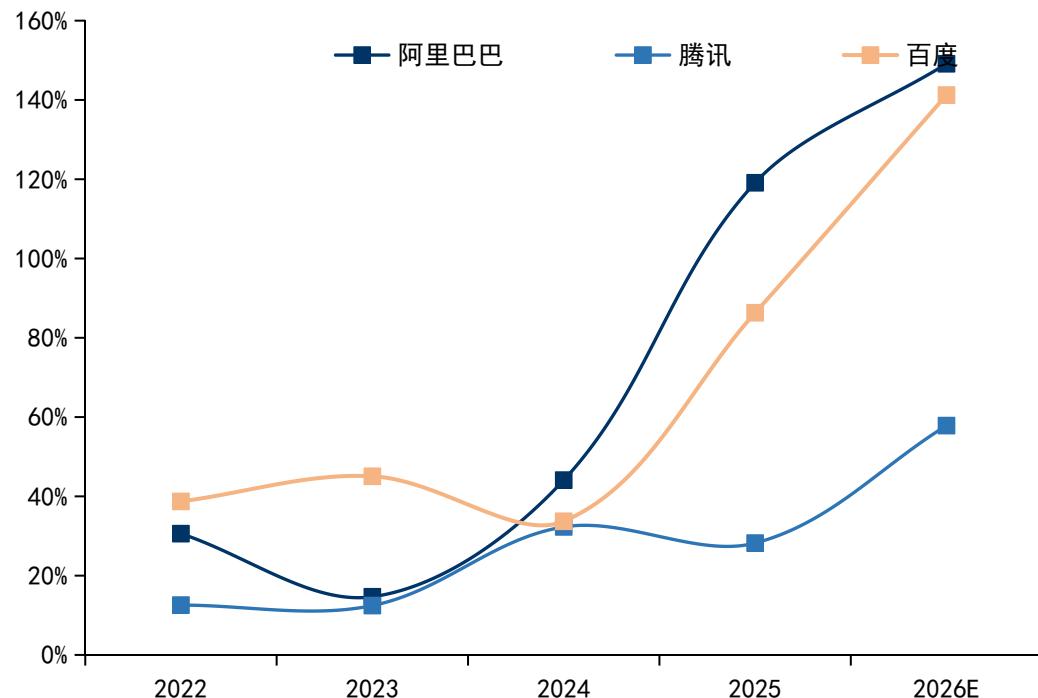
- 过去三年，海外科技厂商的资本开支呈现持续提升趋势，且目前各家资本开支增速指引依然显著高于业务收入/利润增速，我们预计2026年四家资本开支约为经营性现金流的1.2倍，占净利润比例也均超过了100%。
- 国内巨头中，三家巨头（腾讯、阿里、百度）的资本开支规模占净利润和经营性现金流的比例同样加速增长，预计26年阿里巴巴、百度也将超过100%，腾讯也将同比翻倍达到60%。

图：美国巨头CAPEX/利润比率变化趋势



资料来源：Bloomberg、国信证券经济研究所整理（截至2026/8/2）

图：中国巨头CAPEX/利润比率变化趋势

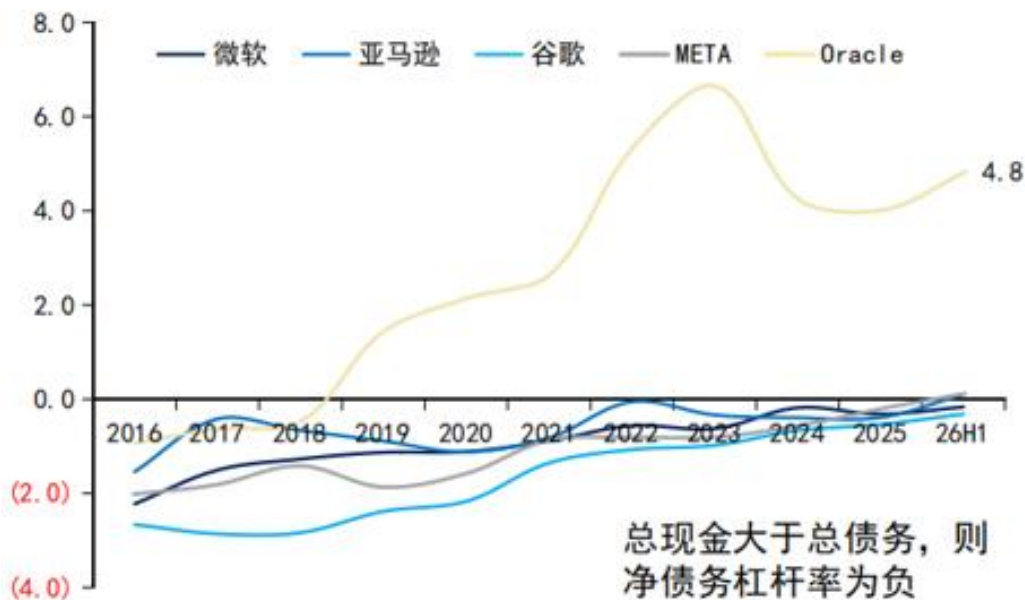


资料来源：Bloomberg、国信证券经济研究所整理（截至2026/8/2）

国内外巨头净债务杠杆率变化

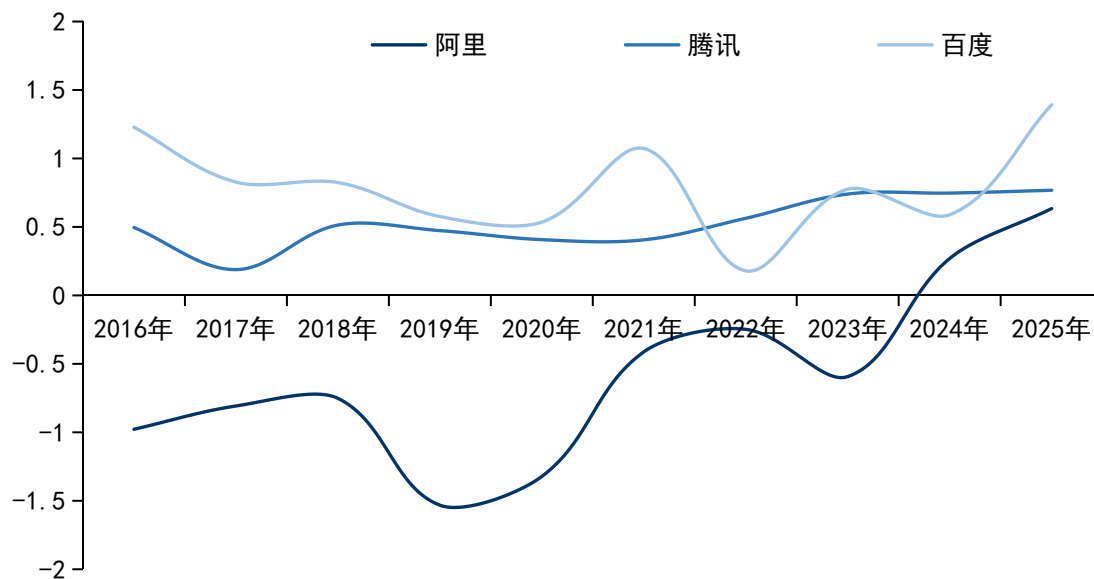
- 公司层面“净债务/EBITDA”是反映现金流覆盖与信用承受度的核心指标。
- 当前五大Hyperscaler平均总债务/EBITDA约1.3x，26H1净杠杆约0.3x（不考虑Oracle约0附近），Oracle为个例，净杠杆约4.8x。国内方面，截至25年平均值为0.93，其中百度已达到1.4
- 对标IG基准：北美IG整体中位净债务/EBITDA，A级：1.24x、BBB级：2.38x；投资级中位利息保障倍数(EBITA/当期利息支出)5.8x。

图：海外巨头净债务杠杆率变化



资料来源：Bloomberg、国信证券经济研究所整理

图：国内互联网净债务杠杆率变化



资料来源：公司财报、国信证券经济研究所整理

资本开支持续攀升，中美巨头增加外部融资



- 随着现金流压力逐渐加大，各公司开始扩大融资规模。2025年微软、谷歌、亚马逊、Oracle和Meta合计发售1210亿美元公司债，为2020年至2024年间年均280亿美元的四倍。

图：中美巨头2025-2026上半年融资情况

	微软	谷歌	亚马逊	Oracle	Meta	阿里巴巴	腾讯	百度
2025-2026年6月累计融资规模	约420亿美元	1650亿美元以上	820亿美元以上	约480亿美元	约550亿美元	约47亿美元	约60亿美元	约40亿美元
2025年Q3	-	-	-	9月发行180亿美元六档公司债，另有180亿美元相关数据中心项目贷款	-	7月发行约120亿港元（约合15.3亿美元零息可交换债），期限至2032年 9月再发行约32亿美元零息可转换优先票据，期限至2032年	9月发行90亿元离岸人民币债，约合12.7亿美元	发行44亿元人民币点心债，票息1.9%，2029年到期
2025年Q4	两笔合计145亿美元，增设超长30年期品种	发行约250亿美元债券，包括175亿美元美元债和65亿欧元欧元债	11月发行150亿美元公司债（5-30年期，票面利率4.6%-4.9%）	-	通过与BlueOwl合资SPV发行270亿美元债务（表外负债，Meta仅持股20%，债务不并入母公司报表）	-	-	-
2026年Q1	单笔70亿美元美元优先无抵押债	2月完成315.1亿美元全球债券发行，包括200亿美元美元债、55亿英镑债和30.55亿瑞郎债，其中10亿英镑为100年期债券，票息6.125%	370亿美元美元债和145亿欧元欧元债	计划全年融资450亿-500亿美元，2月已通过250亿美元债券及50亿美元强制可转优先股融资300亿美元，另设不超过200亿美元ATM增发计划	-	-	6月完成约46.6亿美元双币种融资	-
2026年Q2	美元债90亿及欧元等值20亿，合计110亿美元	5月发行90亿欧元债券和85亿加元债券，合计约170亿美元；随后发行5765亿日元，约36亿美元债券 6月计划股权融资规模800亿美元，其中公开发行700亿美元，向伯克希尔哈撒韦配售100亿美元	6月发行140亿加元债券（约100亿美元），创加拿大公司债市场记录，同步发行欧元中长期债券	-	追加发行250亿美元中长期债券，配套数十亿美元私募信贷额度	-	-	-
整体融资特征	侧重内部现金流管理与资本开支平衡，较少大规模直接进入债券市场	大额单次发行，推出百年超长期限债券；同时启动股权融资	大额集中式发行，单笔数亿级融资为主，多币种布局（美元、加元、欧元）	外部融资依赖度较高，采用公司债、股权和项目融资	母公司直发+SPV表外融资双模式，配套公开债券+私募信贷多元融资工具	零息股权挂钩债券	利用较低人民币融资成本，同时重新进入美元债市场	低息人民币债+零息可交换债

资料来源：Reuters、公司年报、公司公告、新浪财经、国信证券经济研究所整理

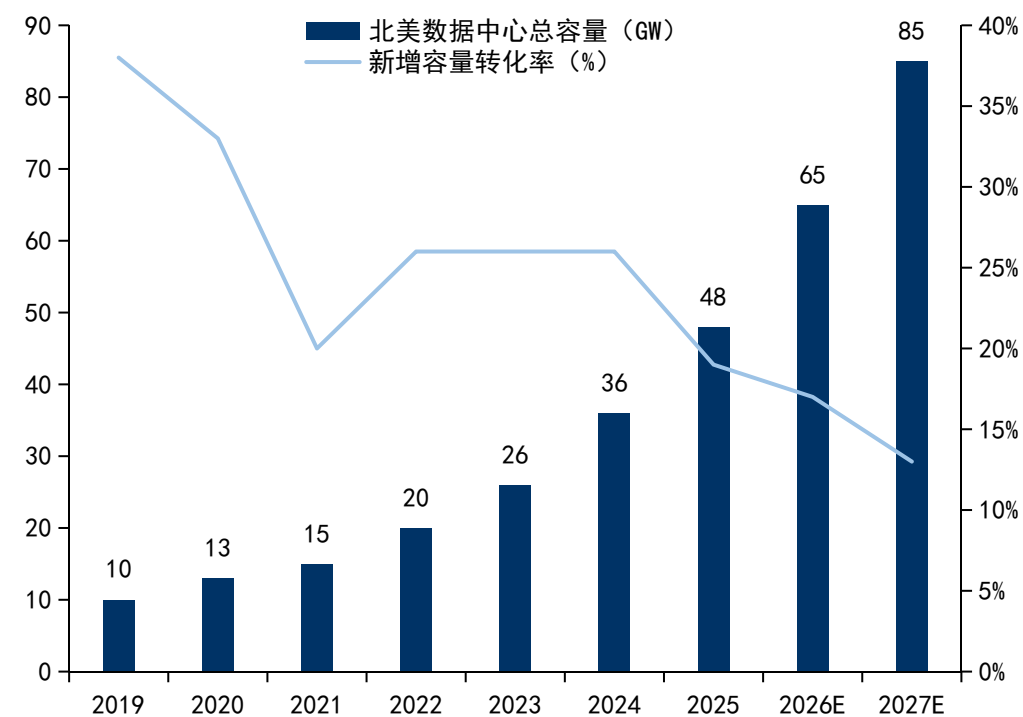
请务必阅读正文之后的免责声明及其项下所有内容

中美数据中心容量扩张加速，中国需求稳步释放



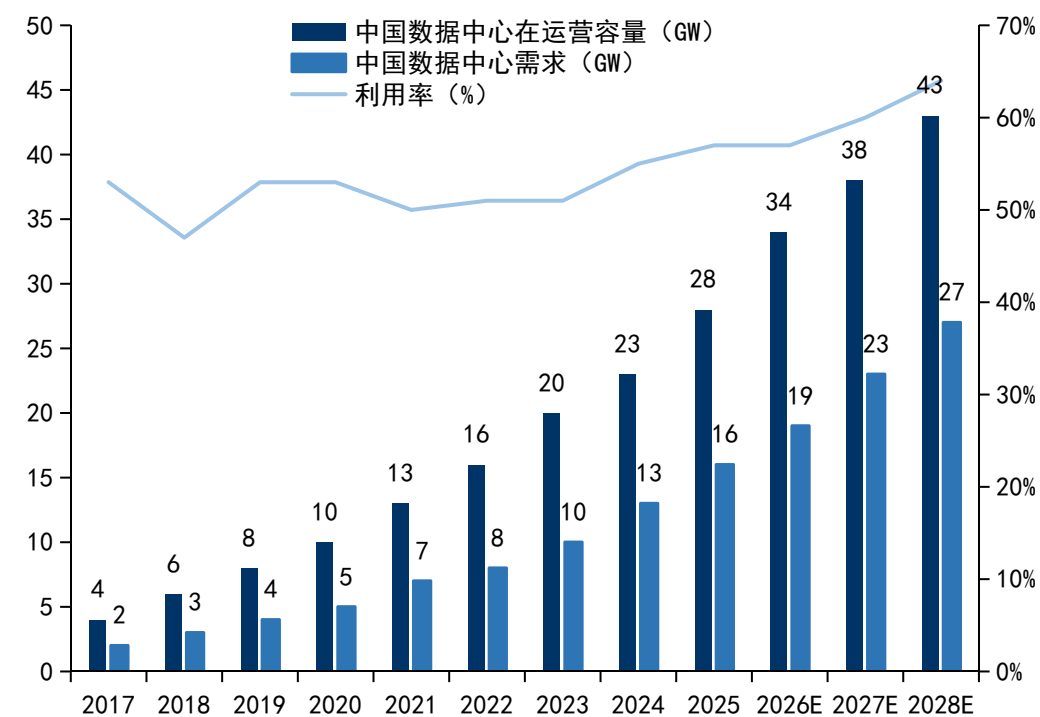
- 北美数据中心进入集中扩容阶段，新增容量明显提速。北美数据中心容量由2019年的10GW增至2025年的48GW，预计2027年进一步达到85GW。2024—2027年累计新增容量约49GW，超过2024年的全部存量规模，显示AI基础设施建设正在推动北美数据中心进入大规模交付周期。
- 中国数据中心容量保持稳步增长，未来仍有较大扩建空间。中国在运营数据中心容量由2017年的4GW增至2025年的28GW，预计2028年达到43GW；2025—2028年预计新增约15GW，数据中心基础设施建设持续推进。

图：北美数据中心容量情况



资料来源：Datacenter Hawk 、国信证券经济研究所预测

图：中国数据中心容量情况



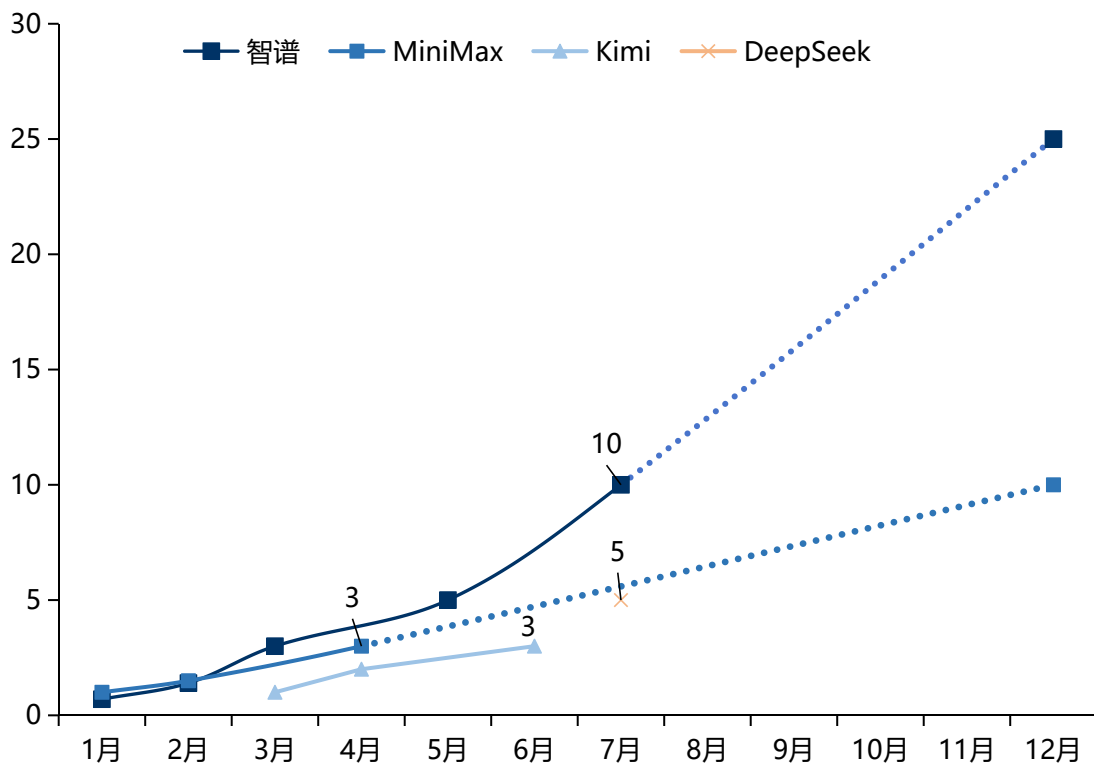
资料来源：MIIT、国信证券经济研究所预测

- [01] 中美AI投入规模及进展
- [02] 国产模型性价比优势凸显，融资规模持续扩大
- [03] 模型商业化闭环与盈利能力测算
- [04] 国内外云厂收入结构拆分
- [05] 风险提示

海外大模型商业化领先，国产模型进入加速放量期

- 美国头部大模型公司ARR规模远超中国，中国大模型厂商商业化进程提速。1) 美国大模型厂商已经形成了结构清晰的盈利金字塔，商业化成熟度更高：根据Yipit数据，Anthropic2026年6月底ARR达690亿美元，OpenAI ARR也达400亿美元，模型能力向真实收入转化；2) 中国大模型厂商进入小基数高速增长阶段，商业化处于爬坡阶段：根据36氪，智谱ARR由年初的不到1亿美元提升至7月的10亿美元，根据公司官方，MiniMax4月ARR突破3亿美元，根据彭博，Kimi ARR在6月达到3亿美元，商业化进程提速，增长弹性高。

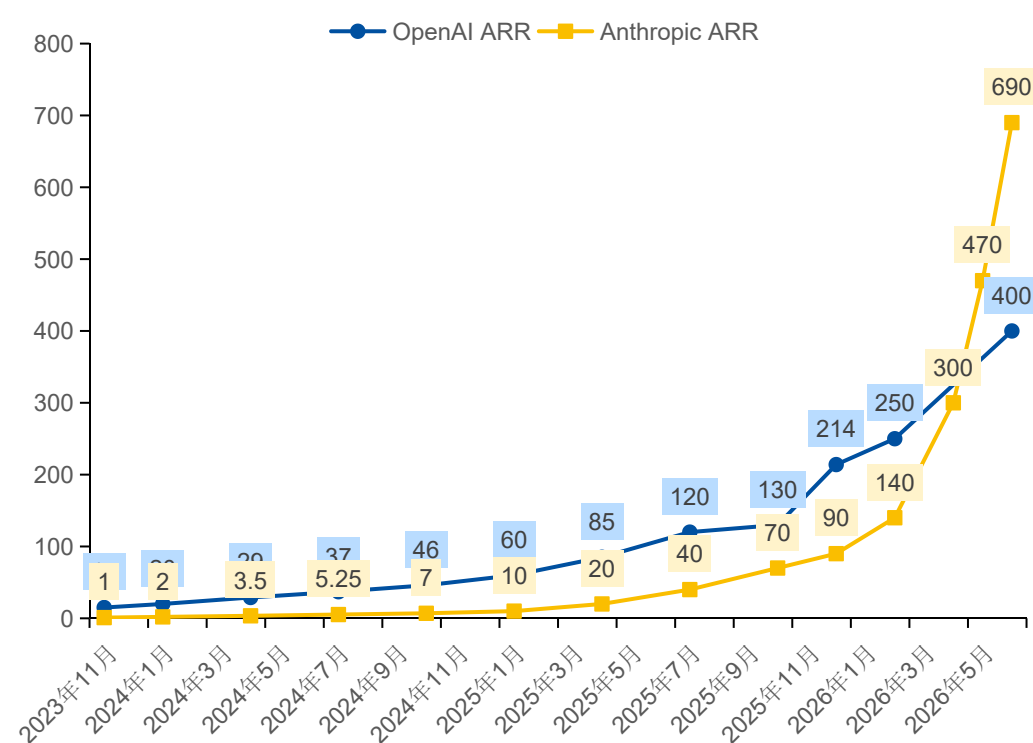
图：智谱、MiniMax、Kimi、DeepSeek ARR跟踪及预测（亿美元）



资料来源：智谱、Kimi、MiniMax、Reuters、Bloomberg、国信证券经济研究所整理
预测

请务必阅读正文之后的免责声明及其项下所有内容

图：OpenAI及Anthropic ARR变化趋势（亿美元）

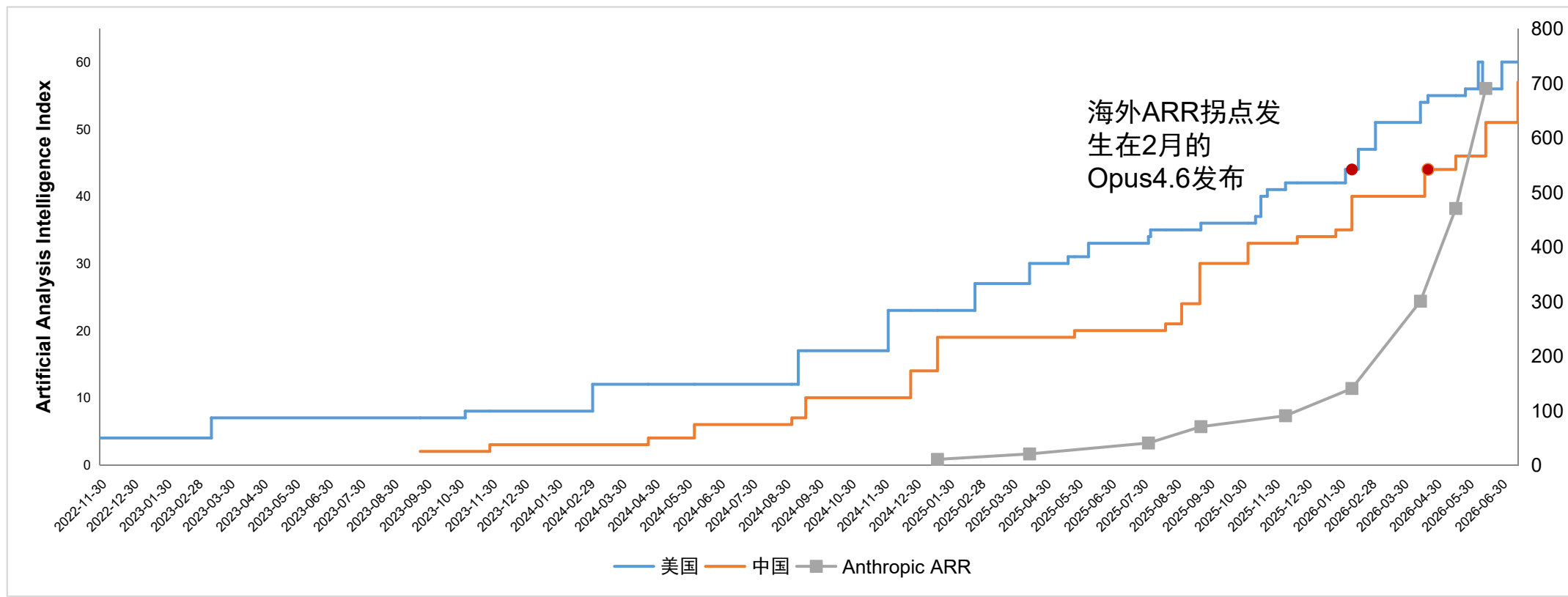


资料来源：Yipit、openai、Anthropic、国信证券经济研究所整理

国产模型迅速拉近中美差距，已达到生产级可用拐点

- 26年2月Opus 4.6的发布后模型真正达到生产级可用，Anthropic ARR由此开始进入爆发拐点，而国产模型大约在4月末左右开始达到Opus 4.6的综合水平，并且与海外前沿模型的差距快速缩小，7月Kimi K3推出后在前端设计表现上达到了当前全世界SOTA，超越Fable 5，国产模型正跨越生产级可用，开始依靠进入ARR的加速期。

图：中美前沿模型智能水平以及Anthropic ARR（亿美元）

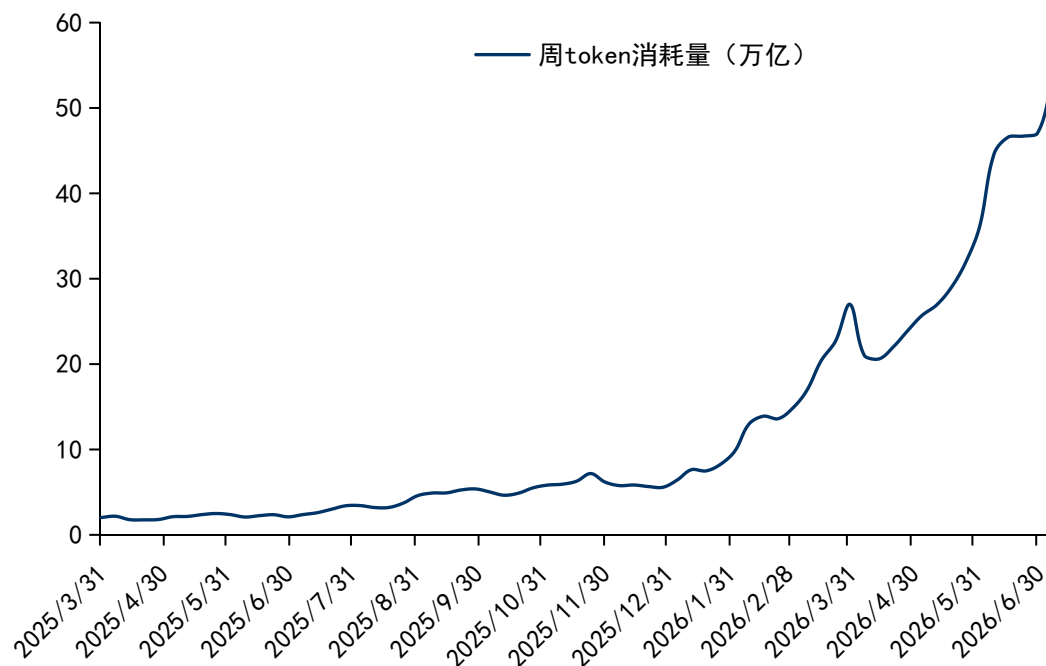


资料来源：Artificial Analysis、国信证券经济研究所整理

Token调用量加速扩张，模型生态走向多元化

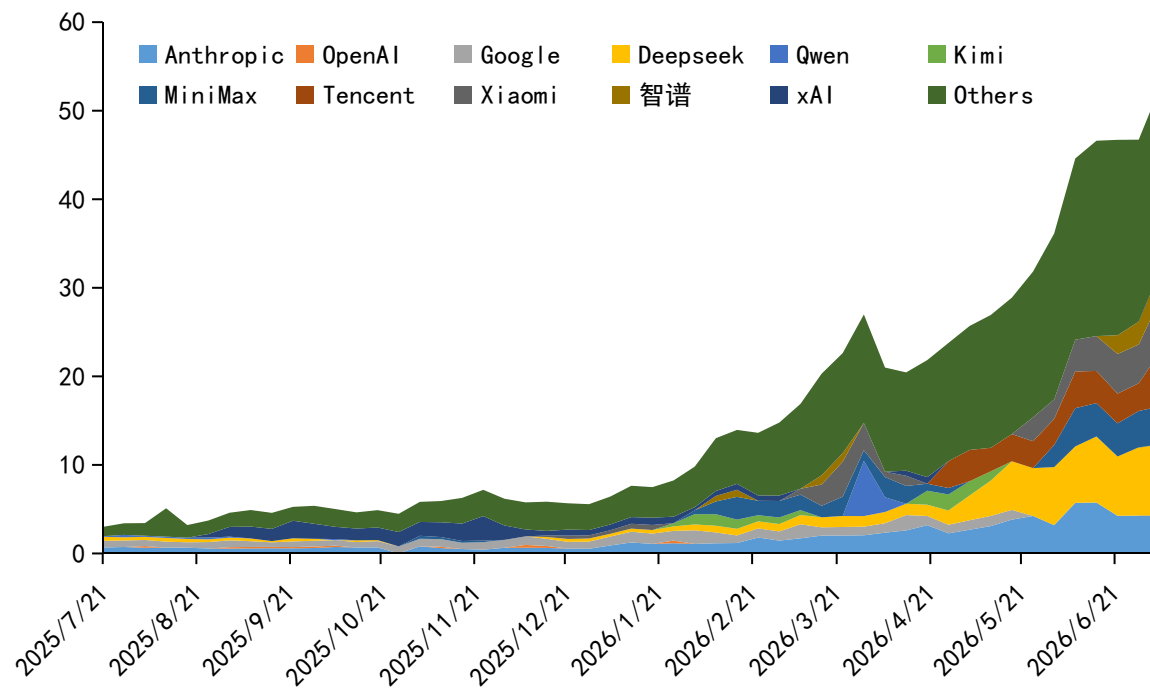
- **OpenRouter平台周度Token调用量实现数量级增长。**2025年6月周度Token消耗均值约2.24T，2026年6月周度Token消耗均值达44.2T，约为去年同期的20倍，大模型调用需求持续快速释放。
- **调用结构走向多元化，国产模型份额持续提升。**从OpenRouter平台各厂商模型调用量看，调用结构由早期的集中于Anthropic、谷歌等海外头部厂商逐步向多元化发展，DeepSeek、Kimi、MiniMax等国产模型调用量快速增长，成为平台增量的重要来源。

图：OpenRouter周度Token调用量



资料来源：Openrouter、国信证券经济研究所整理

图：OpenRouter周度Token调用量（按厂商分类，单位：万亿Token）



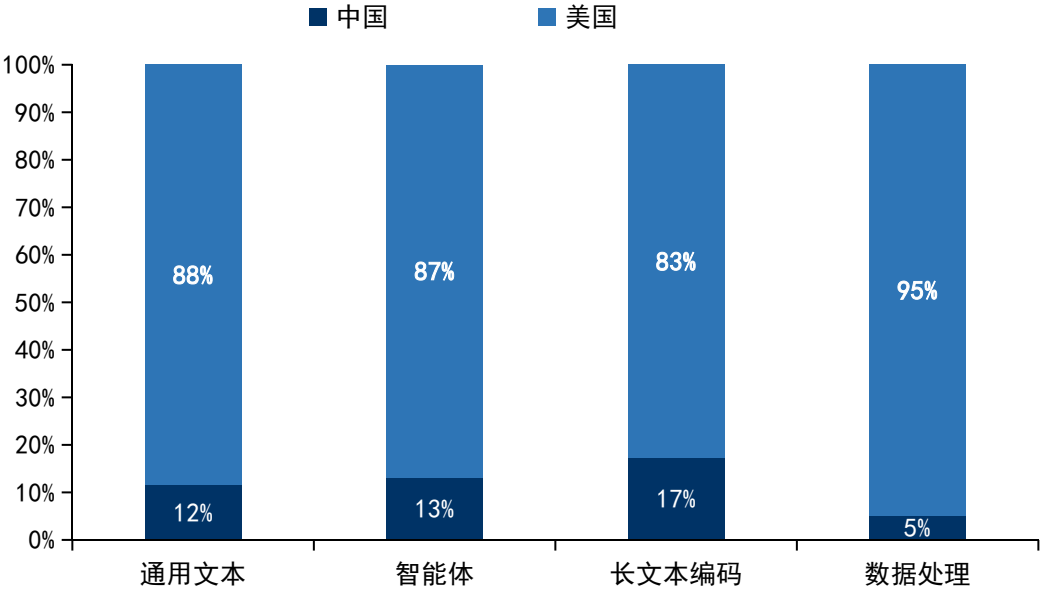
资料来源：Openrouter、国信证券经济研究所整理

国产模型任务级成本优势驱动Token份额领先，规模化应用加速



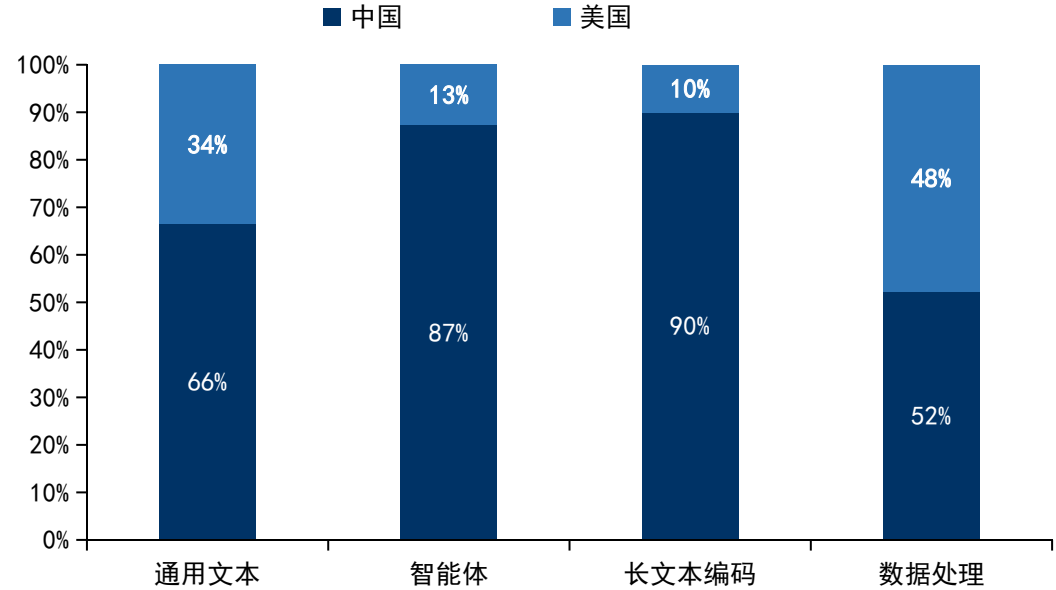
- 国产模型凭借价格竞争力实现Token调用份额领先。OpenRouter数据显示，海外市场对中国模型的采用率快速上升，在通用文本、智能体、长文本编码以及数据处理任务中，国产模型Token使用量占比分别达约66%、87%、90%和52%，其中在长文本编码和Agent这类高Token消耗场景任务中，中国模型已经占据了多数Token使用量，与此同时，中国模型在上述任务中的支出仅占12%、13%、17%和5%，单位支出对应的调用规模显著更高，验证了单任务成本优势对模型采用率的推动。较低的任务成本相比之下更适合高频调用、长上下文及复杂Agent任务，后续有望通过调用规模的扩张以及高价值应用形成持续商业化增长。

图：中国AI模型在各类任务中的支出份额更低



资料来源：Openrouter、国信证券经济研究所整理

图：中国AI模型的Token使用量份额中占据多数

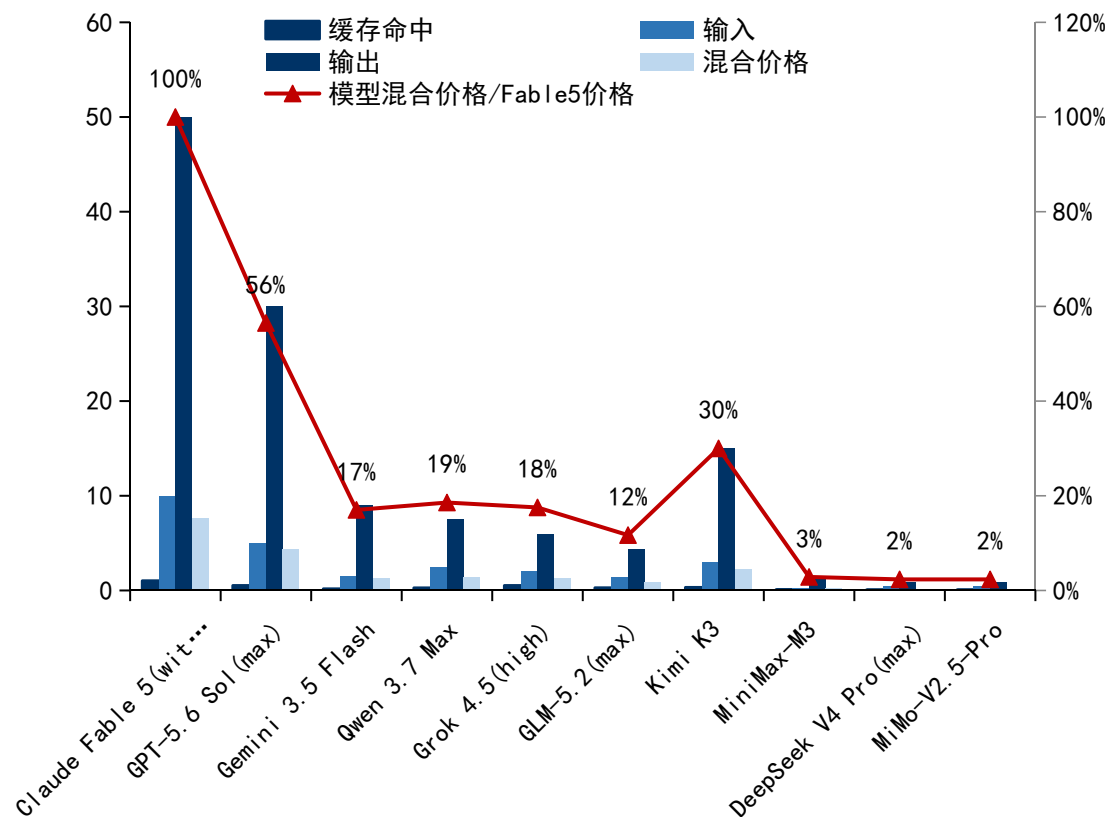


资料来源：Openrouter、国信证券经济研究所整理

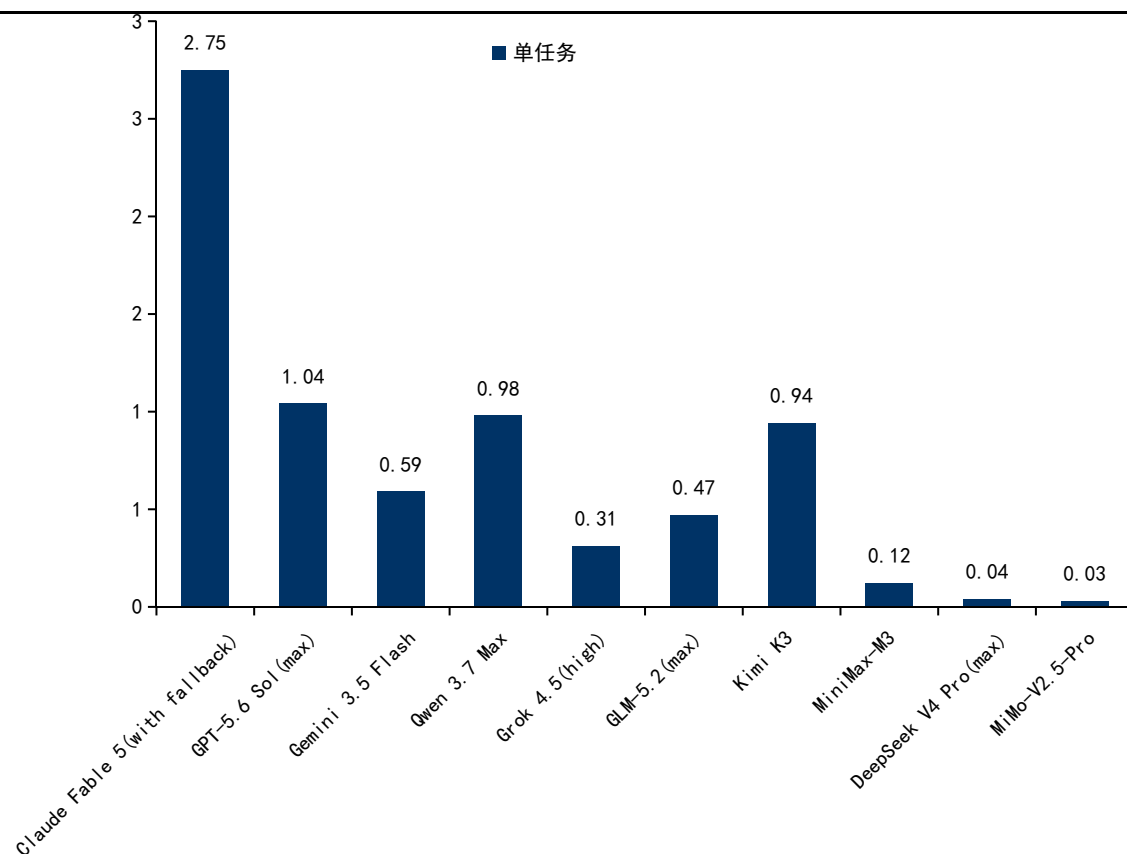
国产模型成本优势显著

- 国产模型优势已由“单Token价格更低”进一步延伸至“完成同一任务的总成本更低”。1) Token价格方面，国产模型在输入、输出、缓存命中以及混合价格上整体仍处于较低水平。2) 单任务成本（Cost per Task）为商业维度的业务价值单元，核心关注点为任务总ROI以及业务收益能否覆盖单次调用。Artificial Analysis口径下，DeepSeek V4 Flash、MiniMax-M3和GLM-5.2的单任务成本分别为0.02、0.12和0.47美元，显著低于美国模型成本。

图：中美模型Token价格成本对比（单位：USD/百万Tokens）



图：中美模型单任务成本对比（单位：USD/Artificial Analysis Intelligence Index task）



资料来源：Artificialanalysis、国信证券经济研究所整理

资料来源：Artificialanalysis、国信证券经济研究所整理

- **融资情况：**月之暗面自2023年启动外部融资以来，已完成天使轮至D轮多轮融资，据公司公告和IT桔子，累计披露融资金额约300亿元。公司融资规模持续扩大，投后估值由天使轮的20亿元提升至2026年D轮的1356亿元。投资方覆盖红杉中国、真格基金、腾讯、阿里巴巴、五源资本、美团龙珠等头部产业与财务资本，股东资源较为多元。
- **模型最新进展：**7月16日，公司发布新一代旗舰模型Kimi K3，总参数规模达2.8T，支持原生图文多模态及100万Token上下文，并提供低、高、最高三档推理强度。K3面向复杂工程、长程推理及Agent任务，模型定位由AI编程进一步扩展至通用Agent与知识工作。K3的Artificial Analysis Intelligence Index达到57分，较K2.6提升13分，综合能力已接近Claude Opus 4.8、GPT-5.5等全球前沿闭源模型；GDPval-AA v2 Elo达到1668，AutomationBench-AA得分53%并位居首位，真实知识工作及SaaS workflow执行能力显著提升。

表：Kimi月之暗面融资情况

时间	轮次	融资金额（人民币，亿元）	投后估值（人民币，亿元）	主要投资方
2023/6/12	天使轮	4.07	20.34	红杉中国
				真格基金
				Monolith Management 砺思资本
				今日资本
2023/7/31	Pre-A轮	数亿美元	33.90	趣加FunPlus
				美团龙珠（领投）
				红杉中国
				Monolith Management 砺思资本
2024/2/19	A轮	67.80	152.56	阿里巴巴（领投）
				蚂蚁集团（领投）
				红杉中国
				小红书
				美团
				Monolith Management 砺思资本
				九安医疗
				蓝驰创投
2024/5/21	B轮	20.34	223.75	襄禾资本
				宿华
				腾讯投资
				高榕创投
				源码资本
2025/12/31	C轮	33.90	291.55	五源资本
				云九资本
				IDG资本（领投）
				王慧文
				腾讯投资
2026/2/17	C+轮	47.46	678.03	阿里巴巴
				Lollapalooza Capital
				九安医疗（领投）
				五源资本（领投）
				腾讯投资（领投）
				阿里巴巴（领投）
2026/5/6	D轮	135.61	1356.05	凯辉基金
				高榕创投
				美团龙珠（领投）
				中移创新产业基金（中国移动）
				CPE源峰
				水木资本
合计		309.18		华峰资本（财务顾问）

资料来源：公司公告、IT桔子，国信证券经济研究所整理

Minimax：从M2.7到M3迭代升级，Agent生态加速完善

- **融资情况：**MiniMax于2021年启动外部融资，现已完成多轮融资，累计披露融资金额超过300亿元。7月中旬，公司完成95亿港币配售+65亿港币可转债发行，直接配售稀释约10%股权，若可转债完全转换稀释约15%股权。
- **模型最新进展：**6月，公司发布新一代旗舰模型 MiniMax M3。相较此前M2.7，M3将编程、Agent、长上下文及原生多模态能力集成于统一模型框架，最高支持100万Token上下文，并具备图像和视频内容理解能力。同时，模型引入自研稀疏注意力架构MSA，在扩大上下文长度和提升模型能力的同时，兼顾推理效率与部署成本。

表：Minimax上市后融资情况

对比指标	单位	MiniMax (0100.HK)
时间与结构		
协议签署/公告日	日期	2026-07-10
股权配售完成日	日期	2026-07-14
可转债完成日	日期	2026-07-16
融资结构	文本	35.6m新A类股 + HK\$6.5bn零息可转债
融资规模与定价		
股权配售毛额	HK\$m	9,540.80
可转债本金	HK\$m	6,500.00
总融资净额（完成口径）	HK\$m	15,877.01
股权稀释与流通盘供给		
配售新股数	股	35,600,000
立即稀释：配售股/配售后总股本	%	10.19%
可转债全部转股新增股数	股	19,402,985
完全摊薄：新增股/转股后总股本	%	14.92%

资料来源：公司公告、国信证券经济研究所整理
请务必阅读正文之后的免责声明及其项下所有内容

表：Minimax融资情况

时间	轮次	融资金额（人民币，亿元）	投后估值（人民币，亿元）	主要投资方
2021年12月	天使轮	2.10	13.56	投资方未透露
2022年3月	Pre-A轮	3.39	37.29	投资方未透露
2023年5月	A轮	17.43	78.45	投资方未透露
2023年7月	A+轮	3.39	108.48	投资方未透露
2024年3月	Pre-B轮	44.34	173.17	投资方未透露
2024年12月	Pre-B+轮	8.37	211.78	投资方未透露
2025年8月	Pre-B++轮	26.47	287.51	投资方未透露
2025/12/22	战略投资	未透露	284.77	视觉中国
2025/12/31	基石轮	23.55	402.24	阿布扎比投资局ADIA
				Aspex Management
				博裕资本
				Janchor Partners建岐
				阿里巴巴
				汇添富基金
				IDG资本
				易方达基金
				韩国未来资产Mirae Asset Venture Investment
				泰康保险
				毅达资本
				Perseverance Asset Management
2026/1/9	IPO上市	36.24	435.98	公共股东
2026/7/10	IPO上市后	138.40	未透露	投资方未透露
合计		303.68		

资料来源：公司公告、IT桔子、国信证券经济研究所整理

智谱：GLM-5.2重磅开源，Coding Plan企业化加速

- **融资情况：**智谱自2021年启动外部融资以来，已完成多轮融资，累计融资金额超过400亿元。7月公司完成314亿元配售，预计配售后占总股本的4.25%。
- **模型最新进展：**6月16日，公司发布并开源新一代旗舰模型GLM-5.2，采用753B总参数、40B激活参数的MoE架构，并以MIT协议开放权重。相较GLM-5.1，新模型重点强化稳定的100万Token上下文、复杂Coding及长程任务执行能力，并提供多档推理强度，在模型效果与响应效率之间灵活权衡。

表：智谱融资情况

对比指标	单位	智谱 (2513.HK)
时间与结构		
上市日期	日期	2026-01-08
协议签署/公告日	日期	2026-07-09
股权配售完成日	日期	2026-07-13
融资规模与定价		
股权配售毛额	HK\$m	31,410.64
可转债本金	HK\$m	0.00
总融资毛额	HK\$m	31,410.64
总融资净额（完成口径）	HK\$m	31,374.95
配售价	HK\$/股	1,588.00
前一交易日收盘价	HK\$/股	1,825.00
配售价较前收折价	%	12.99%
股权稀释与流通盘供给		
配售新股数	股	19,780,000
立即稀释：配售股/配售后总股本	%	4.25%
配售股/融资前可交易股池代理	%	8.94%
配售股/配后可交易股池代理	%	8.20%

资料来源：公司公告、IT桔子，国信证券经济研究所整理

表：智谱融资情况

时间	轮次	融资金额（人民币，亿元）	投后估值（人民币，亿元）	主要投资方
2021年4月	天使轮	0.20	4.07	未列示具体天使轮投资方
2022年5月	A轮	1.52	8.52	通智投资、达晨创鸿、北京华控、荣品投资、将门创业投资、凌云光、南京图灵财智创赢
2023年2月	B1轮	2.08	21.08	君联相道、启明融乾、启明融凯
2023年8月	B2轮	4.10	32.10	天津三快（美团）、君联相道、星连肇基 青岛华控、启明融乾、启明融凯 达晨创鸿、财智创赢
2024年1月	B3轮	6.09	45.72	全德美嘉、上海云珩、君联相道 杭州光合、天津合远、星连肇基 青岛华控
2024年8月	B4轮	12.29	72.29	腾讯投资、北京顺赢、欣欣相融 小锋科技、厦门雅恒、宁波明恒 星连肇基、社保中关村创新基金等
2024年11月	B5轮	13.59	133.59	上海飞珩、无锡云晖、君联相道 招商数科、社保中关村创新基金等
2025年5月	B6轮	43.77	243.77	中关村科学城、大兴产业基金、星连肇基 君联相道、社保中关村创新基金 海河富新优达等
2025/12/30	基石轮	25.78	442.02	WT Asset Management 广发证券、泰康保险、3W Fund Management、方圆资本Optimas Capital 凌云光、高毅资产 JSC International Investment Fund 雾凇资本、JinYi Capital
2026/1/8	IPO上市	37.61	442.50	公共股东
2026/7/9	IPO上市后	271.71	未披露	
合计		418.74		

资料来源：公司公告、IT桔子，国信证券经济研究所整理

- **融资情况：**阶跃星辰自2023年启动外部融资以来，已完成A轮、B轮、B+轮及多笔Pre-IP0融资，融资规模持续扩大。投资方覆盖红杉中国、启明创投、五源资本、腾讯、上海国投等财务及产业资本，Pre-IP0阶段进一步引入中兴通讯、华勤技术、龙旗科技、豪威集团等产业链企业及港投公司，股东结构趋于多元。
- **模型最新进展：**阶跃星辰于2026年5月末发布并开源Step 3.7 Flash，模型定位由此前的高速推理进一步升级至生产级多模态Agent。新模型采用稀疏MoE架构，总参数约198B、单次激活参数约11B，支持256K上下文及三档推理强度，最高生成速度超过400 Tokens/s。在Step 3.5 Flash基础上，Step 3.7 Flash新增原生图片及视频理解能力，并重点强化联网搜索、工具调用、代码执行及多Agent协同等能力。**产品化方面**，公司同步将Step 3.7 Flash接入开放平台及Step Plan，并适配Claude Code、OpenClaw等主流Agent及AI编程框架，进一步降低开发者调用和部署门槛。
- **根据公司官方，**7月13日阶跃星辰发布AI终端品牌STEPX、智能体操作系统Step AOS、个人智能体阶跃Amoo及首款智能体手机STEPX Neo，Amoo可根据用户意图跨应用调用多类服务，推动终端交互由手动操作App向智能体自主完成任务升级。

表：阶跃星辰融资情况

时间	轮次	融资金额（人民币，亿元）	投后估值（人民币，亿元）	ARR	主要投资方
2023/11/1	A轮	数千万美元	67.80		红杉中国
					启明创投
					五源资本
2024/12/23	B轮	数亿美元	135.61		上海国投
					腾讯投资
					五源资本
					启明创投
					上海国投
2026/1/26	B+轮	50.00	200.00		国寿股权
					浦东创投
					厦门国贸控股
					华勤技术
					腾讯投资
					启明创投
					五源资本
					博华产投
					徐汇科技
					光源资本（财务顾问）
2026/5/8	Pre-IP0	169.51	372.91		中兴通讯
					华勤技术
					龙旗科技
					港投公司
					腾讯投资
2026/5/25	Pre-IP0	1.50	390.00		豪威集团
2026/5/27	Pre-IP0	3.00	400.00		米奥兰特
合计		224.01			莲花控股

资料来源：公司公告、IT桔子，国信证券经济研究所整理

Deepseek：首轮融资510亿元，V4双模领跑Agentic Coding赛道



- 融资情况：**2026年6月，DeepSeek完成成立以来的首轮外部融资，融资规模约510亿元，投后估值接近4,000亿元，成为国内大模型领域规模较大的单轮融资之一。公司此前主要依托创始人梁文锋及幻方量化体系支持模型研发，本轮融资标志着其资金来源由内部投入逐步转向外部资本协同。据《金融时报》及路透社7月报道，DeepSeek在完成首轮外部融资后，正考虑启动新一轮融资，拟议投前估值约710亿美元，折合人民币约4,800亿元，较上一轮估值进一步提升。
- 投资方阵容兼具产业资本、财务资本及国家级基金属性，**其中梁文锋仍为最大单一出资方，腾讯、宁德时代体系、京东、网易、IDG资本及国家人工智能产业投资基金等参与投资。多元股东背景有望在算力基础设施、国产芯片适配、云服务、应用入口及产业资源方面形成协同。
- 模型最新进展：**DeepSeek于2026年4月末发布并开源DeepSeek-V4 Preview，包括V4-Pro和V4-Flash两个版本。其中，V4-Pro总参数约1.6万亿、激活参数约490亿，定位是性能旗舰；V4-Flash总参数约2,840亿、激活参数约130亿，主打高效与性价比。两款模型均支持100万Token上下文及思考、非思考双模式，并重点强化Agentic Coding、工具调用及长上下文任务处理能力，模型迭代方向由通用问答进一步转向复杂工程任务和智能体执行。

表：Deepseek融资情况

时间	轮次	融资金额（人民币，亿元）	投后估值（人民币，亿元）	主要投资方
2026年6月	A轮	510.00	4000.00	梁文锋 腾讯投资 宁德时代体系（宁德时代、溥泉资本） 京东 网易 IDG资本 国家人工智能产业投资基金 正心谷资本 九安医疗 拾象科技

资料来源：公司公告、IT桔子，国信证券经济研究所整理

主要国产模型2025以来累计融资规模及等效算力和ARR测算



- **假设一：资金使用周期。** 五家主要国产AI模型公司2025年至今累计融资约1,555亿元。假设上述资金能够支持相关公司运营至2027年6月，即对应约30个月的资金使用周期。
- **假设二：支出及商业化参数。** 假设月度资金支出由低位起步并随时间线性增长，融资资金主要用于训练及推理算力投入；训练与推理成本各占50%，推理业务毛利率为40%。B200单卡月租采用2026年7月市场租赁价格，约为2.93万元/月。
- **测算方法及结论：**按30个月内月度支出线性递增测算，2027年6月五家公司单月支出约104亿元，对应等效算力规模约35万张B200。其中，单月推理成本约52亿元；按40%的推理业务毛利率倒算，当月推理收入约86亿元，对应ARR约1,036亿元。
- 注：若模型公司后续获得新增融资、资金使用节奏发生变化，或算力租赁价格、业务成本结构及毛利率水平出现较大波动，实际等效算力规模及ARR可能与测算结果存在偏差。

表：主要国产模型2025-2026年累计融资及对应等效算力和ARR测算

公司	2025年至今累计融资 (亿元人民币)	线性递增下27年6月支出 (亿元人民币)	B200单卡月租 (人民币 元)	等效B200卡数 (张)	27年6月推理成本 (亿元人民币)	月收入 (亿元人民币)	ARR (亿元人民币)
DeepSeek	510.00	34.00	29,290.68	116,078	17.00	28.33	340
阶跃星辰	224.01	14.93	29,290.68	50,986	7.47	12.45	149.34
智谱	378.87	25.26	29,290.68	86,232	12.63	21.05	252.58
MiniMax	224.66	14.98	29,290.68	51,133	7.49	12.48	149.77
月之暗面	216.97	14.46	29,290.68	49,383	7.23	12.05	144.65
合计	1,554.51	103.63	29,290.68	353,812	51.82	86.36	1,036.34

资料来源：公司公告、IT桔子，国信证券经济研究所整理测算

- [01] 中美AI投入规模及进展
- [02] 国产模型性价比优势凸显，融资规模持续扩大
- [03] 模型商业化闭环与盈利能力测算
- [04] 国内外云厂收入结构拆分
- [05] 风险提示

吞吐与总时延

$$TPS_{copy} = \frac{B(N_{in} + N_{out})}{T_{total}}$$

$$TPS_{card} = \frac{TPS_{copy}}{G}$$

$$T_{total} = (T_{decode} + T_{prefill})(1 + r_{wait})$$

PREFILL

Prefill 总时延

$$T_{prefill} = \max(T_{pf,comp}, T_{pf,mem})$$

计算时延

$$T_{pf,comp} = \left(\frac{2BN_{(in)}P_{act}}{CG} \right) \times \left(\frac{1000}{\eta_c} \right)$$

访存时延

$$T_{pf,mem} = \frac{b_p P_{model}}{BW_{(HBM)} G \eta_{HBM}}$$

B: Batch

N_{in} / N_{out} : 单请求输入 / 输出 Token 数

Pact / Pmodel: 激活 / 模型参数量 G: 卡数

C: 单卡计算能力

b_p : 每参数字节数

η : 效率

数据来源: 英伟达技术文档, 国信证券经济研究所整理

DECODE

单 Token Decode 时延累计

$$T_{decode} = N_{(out)}(T_{mem1} + T_{mem2} + T_{comm})$$

权重访存

$$T_{mem1} = \frac{b_p P_{act}}{BW_{(HBM)} \eta_{HBM}}$$

KV Cache 访存

$$T_{mem2} = \frac{BL_{(ctx)} L_{(MoE)} (2D_{KV}) b_p}{BW_{(HBM)} G \eta_{HBM}}$$

EP 通信

$$T_{comm} = \frac{2L_{(MoE)} E_{EP}}{BW_{(link)} \eta_{topo}}$$

$$E_{EP} = B \cdot TopK \cdot H \cdot b_p$$

BW_HBM: 单卡 HBM 带宽 BW_link: 互联带宽 L_ctx: 上下文长度 L_MoE: MoE 层数
D_KV: KV 维度 H: Hidden Size

模型、基础设施、推理精度等参数及假设

表：模型、基础设施、推理精度等参数及假设

模型	说明	GLM5.2	GLM5.2	GLM5.2	GLM5.2	Opus5	Opus5	Opus5	GLM5.2
芯片		910C	昇腾950	H200	GB200	H100	GB200	Rubin	GB200 (在美国)
其他假设									
存储精度		2	2	2	2	1	1	1	2
部署所需最小计算单元数量		19	17	18	14	117	52	32	14
假设模型参数占比		60%	60%	60%	60%	60%	60%	60%	60%
实际部署芯片数量		19	17	18	14	117	52	32	14
每个参数占字节数（推理精度）									
		2	2	2	2	1	1	1	2
缓存命中		14000	14000	14000	14000	14000	14000	14000	14000
输入Token数量		4000	4000	4000	4000	4000	4000	4000	4000
输出Token数量		2000	2000	2000	2000	2000	2000	2000	2000
模型参数									
激活参数量（B）		40	40	40	40	500	500	500	40
模型参数量（B）		744	744	744	744	5600	5600	5600	744
TopK		8	8	8	8	32	32	32	8
HiddenSize		6144	6144	6144	6144	6144	6144	6144	6144
Moe层数		75	75	75	75	75	75	75	75
KV维度		576	576	576	576	576	576	576	576
硬件参数									
计算能力（TFLOPS）		800	500	989	2500	1979	5000	17500	2500
		FP16	FP16	FP16	FP16	FP8	FP8	FP8	FP16
显存带宽（GB/s）		3200	4000	4800	8000	3350	8000	22000	4800
网络带宽（GB/s）		784	2000	900	1800	900	1800	3600	900
内存大小(GB)		128	144	141	180	80	180	288	180
计算效率/MFU	英伟达有成熟软件栈，国产卡适配难度大	30%	40%	45%	50%	48%	50%	55%	50%
HBM带宽效率		50%	55%	60%	65%	65%	70%	75%	70%
拓扑效率	是否超节点/跨节点通信能力/卡数量等都会产生影响	50%	60%	70%	70%	80%	85%	90%	85%

资料来源：各模型、芯片公司官网，国信证券经济研究所整理测算

模型推理毛利率和TPS的测算

- 模型参数假设：假设Opus5是一款总参数5.6T，激活参数500B的模型，存储和计算精度都是FP8。GLM5.2则采用FP16。
- 副本：副本是一套可以独立处理请求的完整模型实例。假设副本显存按60%/30%/10%分配给模型参数/KV Cache/其他用途，由此计算出一个副本所需芯片数量。
- 输入/输出假设：假设缓存命中1.4万Token，输入4000 Token，输出2000 Token，比例为7:2:1。
- Batch假设：Batch指同一个模型副本在一次计算中，同时处理多少条请求或多少个Token。通过30%显存用于KV Cache计算Batch。
- 利用率假设：国内芯片利用率普遍低于海外，我们对计算效率、HBM带宽效率、拓扑效率进行假设，计算得到不同模型在不同芯片上的利用率。
- 吞吐口径：先求Decode输出TPS，再倒算请求量与计费Token；η sched统一吸收Prefill插入、等待与碎片化损耗，避免高估吞吐。
- 结论一：模型参数量越大，单卡TPS越小，芯片性能越强，单卡TPS越大。
- 结论二：影响模型成本和定价的原因：模型层面核心来自参数量，以及注意力机制的选择（如何让相同上下文占用更小的KV Cache）
- 结论三：国产模型毛利率低于海外不全是因为定价便宜，核心是因为Infra能力差距 ①推理中模型支持的精度不够低；②国产芯片及超节点整体硬件基础设施的利用率低于海外。

表：模型推理毛利率影响因素测算

模型 芯片	GLM5.2 910C	GLM5.2 昇腾950	GLM5.2 H200	GLM5.2 GB200	Opus5 H100	Opus5 GB200	Opus5 Rubin	GLM5.2 GB200 (在美国)
Prefill时间	4.48	3.92	3.92	3.78	6.78	6.67	6.50	6.67
Prefill计算时间	2.80	2.80	2.80	2.80	5.42	5.42	5.42	5.42
Prefill内存时间	0.0480	0.0393	0.0294	0.0208	0.0220	0.0193	0.0105	0.0321
Decode时间	0.1300	0.0809	0.0658	0.0376	0.3874	0.1615	0.0609	0.0534
Decode内存时间	0.0500	0.0364	0.0278	0.0154	0.2296	0.0893	0.0303	0.0238
Decode 内存时间2	0.0227	0.0186	0.0139	0.0098	0.0104	0.0091	0.0050	0.0152
Decode通信时间	0.0085	0.0028	0.0053	0.0027	0.0699	0.0329	0.0155	0.0044
Decode_batch	227	227	227	227	1706	1706	1706	227
KV Cache占比	30%	30%	30%	30%	30%	30%	30%	30%
TPS Decode (Tokens/s/GPU)	90	163	196	437	38	204	864	308
单卡一小时的 Decode Token数 (百万)	0.32	0.59	0.70	1.57	0.14	0.73	3.11	1.11
单卡一小时处理的 请求数=Deocde Token/单次请求的 输出Token	162	293	352	787	68	367	1,554	555
单卡一小时的缓存 命中Token数	2.3	4.1	4.9	11.0	1.0	5.1	21.8	7.8
单卡一小时的缓存 未命中Token数	0.6	1.2	1.4	3.1	0.3	1.5	6.2	2.2
整体利用率	31%	39%	44%	49%	55%	60%	66%	58%
毛利率	49%	53%	53%	65%	52%	79%	92%	40%

资料来源：英伟达，DeepSeek，国信证券经济研究所整理测算

➤ 云厂：

- 1. 收入端：1GW GB200数据中心大约43万张卡，假设租赁价格为6美元，对应租赁收入为224亿美元。
- 2. 成本端：①1GW GB200数据中心对应CAPEX大约400亿美元；②假设运营成本大约为建设成本的20%。因此总成本约为137亿美元
- 3. OPM：测算得到1GW B200数据中心的OPM大约为39%。

➤ 模型厂商：以Opus5为例

- 1. 收入端：根据前面测算，Opus5在GB200上的TPS为2037个Token/s/GPU, Token的平均价格大约3.9美元，1GW算力中，50%用于推理，由此计算得到1GW GB200对应大约528亿美元收入
 - 2. 成本端：①云厂收入对应模型公司的算力成本，假设模型公司的算力成本中训练和推理比例为1:1；②假设其他运营成本占收入比例为20%/15%
 - 3. OPM：测算得到Opus5在GB200上API 调用OPM达到37%
- 结论：美国1GW GB200数据中心需要400亿美元CAPEX，对应云厂224亿美元年化租赁收入，对应模型厂528亿美元年收入。

表：1GW数据中心商业模型测算

1GW数据中心商业闭环	
单位：亿美元	
模型	Opus5
芯片	GB200
云厂	
总成本	124
CAPEX	400
综合折旧年限	9
折旧	44
运营成本	80
运营成本/CAPEX	20%
收入（亿美元）	224
1GW数据中心对应芯片数量	426,888
租赁价格（美元/h/GPU）	6
毛利率	80%
OP	100
OPM	45%
模型厂商	
收入（亿美元）	528
TPS（Token/GPU）	2037
Token价格(美元/百万Tokens)	3.9
降价幅度	
毛利率	79%
收入/GW（亿美元）	528
算力成本	224
训练成本	112
占比	50%
推理成本	112
占比	50%
其他运营成本	106
占收入比	20%
OP	198
OPM	37%

资料来源：英伟达，国信证券经济研究所整理测算

如何理解价格战对模型厂和云厂的影响？

- 假设降价20%：则模型厂在成本固定的情况下，收入削减20%，毛利率从79%降低至73%。OPM从37%降低至27%
- 假设降价50%：毛利率从79%降低至57%。OPM从37%降低至亏损
- 结论1：价格战会直接冲击模型厂的毛利率，且由于训练开支和运营成本的相对刚性，会导致模型的OPM迅速下降。当价格减半时有可能导致模型厂亏损。
- 结论2：随着模型价格战加剧，模型厂为了维持毛利率可能将价格影响传递到云厂。则假设降价50%，且维持65%的毛利率/10%的OPM，测算得到，云厂的毛利率将从80%下降至76%，对应的OPM则将从接近45%下降至33%。

表：1GW数据中心商业模型测算

模型芯片	Opus5 GB200	Opus5 GB200	Opus5 GB200	Opus5 GB200
云厂				
总成本	124	124	124	124
CAPEX	400	400	400	400
综合折旧年限	9	9	9	9
折旧	44	44	44	44
运营成本	80	80	80	80
运营成本/CAPEX	20%	20%	20%	20%
收入 (亿美元)	224	224	224	185
1GW数据中心对应芯片数量	426,888	426,888	426,888	426,888
租赁价格 (美元/h/GPU)	6	6	6	4.9
毛利率	80%	80%	80%	76%
OP	100	100	100	60
OPM	45%	45%	45%	33%
模型厂商				
收入 (亿美元)	528	422	264	264
TPS (Token/GPU)	2037	2037	2037	2037
Token价格(美元/百万Tokens)	3.9	3.1	1.9	1.9
降价幅度		20%	50%	50%
毛利率	79%	73%	57%	65%
收入/GW (亿美元)	528	422	264	264
算力成本	224	224	224	185
训练成本	112	112	112	92
占比	50%	50%	50%	50%
推理成本	112	112	112	92
占比	50%	50%	50%	50%
其他运营成本	106	84	53	53
占收入比	20%	20%	20%	20%
OP	198	113	-13	26
OPM	37%	27%	-5%	10%

资料来源：英伟达，国信证券经济研究所整理测算

芯片升级对模型毛利率的影响？

表：1GW数据中心商业模式测算

- 假设1：1GW
H100/GB200/Rubin/910C/950DT
（国内）对应的资本开支分别为
370/400/470/200/300亿美元，对
应的芯片数量分别为
78/43/26/48/44.5万张。
- 假设2：H100、GB200、Rubin、
910C、950DT租赁价格分别按
2.5/6/10/1.5/2.5美元/h/GPU测算。
- 结论：随着芯片的升级，相同模型
的毛利率持续提升。测算得到
Opus5在Rubin上的毛利率约88%，
较GB200高约9pcts，较H100高出约
36pcts。GLM5.2在950DT上的毛利
率53%也高于910C的49%

模型 芯片 云厂	Opus5 H100	Opus5 GB200	Opus5 Rubin	GLM5.2 910C	GLM5.2 950DT
总成本	115	124	146	52	78
CAPEX	370	400	470	200	300
综合折旧年限	9	9	9	9	9
折旧	41	44	52	22	33
运营成本	74	80	94	30	45
运营成本/CAPEX	20%	20%	20%	15%	15%
收入 (亿美元)	171	224	314	63	98
1GW数据中心对应芯片数量	779,200	426,888	256,104	480,000	445,217
租赁价格 (美元/h/GPU)	2.5	6	14	1.5	2.5
毛利率	76%	80%	83%	65%	66%
OP	56	100	168	11	19
OPM	33%	45%	53%	17%	20%
模型厂商					
收入 (亿美元)	179	528	1343	61	103
TPS (Token/GPU)	377	2037	8636	900	1627
Token价格(美元/百万Tokens)	3.9	3.9	3.9	0.9	0.9
降价幅度					
毛利率	52%	79%	88%	49%	53%
收入/GW (亿美元)	179	528	1343	61	103
算力成本	171	224	314	63	98
训练成本	85	112	157	32	49
占比	50%	50%	50%	50%	50%
推理成本	85	112	157	32	49
占比	50%	50%	50%	50%	50%
其他运营成本	36	106	269	9	15
占收入比	20%	20%	20%	15%	15%
OP	-28	198	760	-11	-10
OPM	-16%	37%	57%	-18%	-10%

资料来源：英伟达，国信证券经济研究所整理测算

- 海外闭源模型拥有高于开源模型的利润率，但对于云厂而言，部署闭源模型还需要对模型厂进行分成，最终对利润的影响可能取决于分成比例和相关运营支出的刚性成本规模。
- 以Opus5和GLM5.2为闭源/开源模型代表，假设云厂部署闭源模型的抽成分别为20%/40%。毛利率角度云厂MaaS业务只需要承担1GW数据中心的折旧以及分成成本，可以计算出采用开源模型有更好的毛利率。
- 当考虑了相对刚性的运营支出后，在20%的分成比例下，采用闭源模型与开源模型的OPM相当，40%分成模式云厂使用开源模型有更大的利润空间。

表：1GW数据中心 云厂MaaS业务使用开源&闭源模型ROI对比

	Opus5 GB200	Opus5 GB200	GLM5.2 GB200
总成本	597	386	174
分成成本	422	211	0
分成比例	40%	20%	0%
推理成本	44	44	44
数据中心运营成本	80	80	80
MaaS平台运营成本	50	50	50
收入	1,056	1,056	531
TPS (Token/GPU)	2,037	2,037	4,371
Token价格(美元/百万Tokens)	3.85	3.85	0.9
毛利润	589	800	486
毛利率	56%	76%	92%
OP	459	670	356
opm	43%	63%	67%

资料来源：英伟达，国信证券经济研究所整理测算

- [01] 中美AI投入规模及进展
- [02] 国产模型性价比优势凸显，融资规模持续扩大
- [03] 模型商业化闭环与盈利能力测算
- [04] 国内外云厂收入结构拆分
- [05] 风险提示

谷歌云拆分

表：谷歌云收入利润拆分

- 26年预计谷歌云收入同比+85%，绝对值增长约500亿美元，其中100亿来自传统云，400亿来自AI云的TPU、GPU租赁、MaaS增长，TPU租赁及外售是主要驱动因素。
- 27年预计谷歌云收入同比+95%，绝对值增长约1000亿美元，其中150亿来自传统云，AI云贡献绝大部分增量，最核心的增长来自TPU租赁增量约300亿+TPU外售增量约250亿美元。我们预计26年TPU出货量接近1000万张，推动公司AI云收入占比达到70%。
- 利润端，我们测算26Q3可能出现小幅OPM下滑，主要是部分采用算力租赁方式满足需求。

单位: 百万美元	2024	2025	2026E	2027E	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26E	4Q26E	1Q27E	2Q27E	3Q27E	4Q27E
谷歌云收入	43,229	58,705	108,414	211,209	12,260	13,624	15,157	17,664	20,028	24,768	28,802	34,817	41,329	49,611	56,347	63,923
YoY	31%	36%	85%	95%	28%	32%	34%	48%	63%	82%	90%	97%	106%	100%	96%	84%
员工薪酬	20,519	22,078	27,870	36,231	5,412	5,517	5,586	5,563	6,443	6,933	7,262	7,232	8,376	9,013	9,440	9,401
YoY	8%	8%	26%	30%	5%	7%	8%	11%	19%	26%	30%	30%	30%	30%	30%	30%
占收入比例	47%	38%	26%	17%												
其他运营成本 (含折旧)	16,598	22,717	42,582	99,829	4,671	5,281	5,977	6,788	6,987	9,021	11,400	15,175	18,186	22,986	26,846	31,811
YoY	35%	37%	87%	134%	33%	32%	41%	40%	50%	71%	91%	124%	160%	155%	136%	110%
占收入比例	38%	39%	39%	47%												
OP	6,112	13,910	37,962	75,149	2,177	2,826	3,594	5,313	6,598	8,814	10,140	12,410	14,767	17,612	20,060	22,710
OPM	14%	24%	35.0%	35.6%	18%	21%	24%	30%	33%	35.6%	35.2%	35.6%	35.7%	35.5%	35.6%	35.5%
Google Cloud Platform	27,374	40,804	88,466	189,160	8,322	9,328	10,503	12,652	15,639	19,981	23,615	29,231	36,478	44,319	50,614	57,749
YoY	39%	49%	117%	114%	37%	43%	45%	68%	88%	114%	125%	131%	133%	122%	114%	98%
占比	63%	70%	82%	90%												
OP	-1,815	4,959			208	678	1,267	2,807	4,404	6,420	7,547	9,617	12,342	14,966	17,194	19,623
OPM	-7%	12%			2%	7%	12%	22%	28%	32%	32%	33%	34%	34%	34%	34%
传统云	23,833	31,407	42,573	57,473	6,997	7,675	8,098	8,637	9,366	10,615	10,932	11,660	12,644	14,331	14,758	15,740
YoY	32%	32%	36%	35%	30%	34%	29%	34%	34%	38%	35%	35%	35%	35%	35%	35%
占GCP	87%	77%	48%	30%												
OP	-2,988	1,468	-16,763	-48,240	-243	98	360	1,253	2,047	2,934	3,022	3,223	3,495	3,961	4,079	4,351
OPM	-13%	5%	-39%	-84%	-3%	1%	4%	15%	22%	28%	28%	28%	28%	28%	28%	28%
AI云	3,541	9,397	45,894	131,687	1,324	1,653	2,405	4,015	6,274	9,365	12,683	17,572	23,835	29,988	35,855	42,009
YoY	117%	165%	388%	187%	87%	103%	152%	278%	374%	466%	427%	338%	280%	220%	183%	139%
占GCP	13%	23%	52%	70%	16%	18%	23%	32%	40%	47%	54%	60%	65%	68%	71%	73%
占云收入比例	8%	16%	42%	62%	11%	12%	16%	23%	31%	38%	44%	50%	58%	60%	64%	66%
OP	1,173	3,492	16,763	48,240	451	579	907	1,554	2,357	3,486	4,525	6,395	8,847	11,005	13,115	15,273
OPM	33.1%	37.2%	36.5%	36.6%	34.1%	35.1%	37.7%	38.7%	37.6%	37.2%	35.7%	36.4%	37.1%	36.7%	36.6%	36.4%
AI IaaS收入	3,541	7,049	34,154	96,467	1,207	1,418	1,700	2,724	4,513	7,017	9,748	12,876	16,791	21,887	26,346	31,443
YoY	117%	99%	385%	182%	70%	74%	78%	156%	274%	395%	473%	373%	272%	212%	170%	144%
OP	1,173	2,318	10,893	30,630	393	462	555	908	1,476	2,312	3,058	4,047	5,325	6,955	8,360	9,990
OPM	33.1%	32.9%	31.9%	31.8%	32.5%	32.6%	32.6%	33.4%	32.7%	33.0%	31.4%	31.4%	31.7%	31.8%	31.7%	31.8%
TPU收入	2,214	4,059	23,827	80,130	609	731	893	1,826	2,448	4,642	6,960	9,778	13,523	18,130	21,935	26,542
YoY	92%	83%	487%	236%	38%	43%	49%	175%	302%	535%	679%	435%	453%	291%	215%	171%
占云收入比例	5%	7%	22%	38%	5%	5%	6%	10%	12%	19%	24%	28%	33%	37%	39%	42%
TPU租赁	2,214	4,059	18,827	50,130	609	731	893	1,826	2,448	4,142	5,460	6,778	9,023	11,530	13,535	16,042
YoY	168%	83%	364%	166%	38%	43%	49%	175%	302%	467%	511%	271%	269%	178%	148%	137%
TPU外售			5,000	30,000						500	1500	3000	4500	6600	8400	10500
GPU收入	1,327	2,990	10,327	16,336	598	688	807	897	2,065	2,375	2,788	3,098	3,267	3,757	4,411	4,901
YoY	168%	125%	245%	58%												
MaaS收入		2,348	11,740	35,220	117	235	704	1,291	1,761	2,348	2,935	4,696	7,044	8,101	9,509	10,566
YoY			400%	200%												
OP		1,174	5,870	17,610	59	117	352	646	881	1,174	1,468	2,348	3,522	4,050	4,755	5,283
OPM		50%	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%

- **折旧影响可能相对滞后：**谷歌26/27年预计收入增长接近翻倍，与CAPEX增速接近，在此基础上，26Q2折旧增速仅40%+。另一方面，谷歌的资本开支大约90%在当年转为在建工程，意味着在次年开始折旧，因此整体折旧对利润影响相对滞后。假设28年开始收入增速放缓，CAPEX增速降至20%，折旧增速达到70%，对利润端的影响开始逐渐显现。

表：谷歌云折旧摊销预测

期间 实际 / 预测	2024A 实际	2025A 实际	2026E 预测	2027E 预测	2028E 预测	2029E 预测	2030E 预测	2031E 预测
资本开支、在建工程与转固								
资本开支	\$52,535	\$91,447	\$200,000	\$400,000	\$480,000	\$528,000	\$554,400	\$554,400
年末在建工程 / 当年CapEx	96.3%	85.9%	90.0%	90.0%	90.0%	90.0%	90.0%	90.0%
年末在建工程	\$50,597	\$78,592	\$180,000	\$360,000	\$432,000	\$475,200	\$498,960	\$498,960
模型期可用CapEx			\$119,402	\$400,000	\$480,000	\$528,000	\$554,400	\$554,400
模型期期初在建工程			\$122,814	\$180,000	\$360,000	\$432,000	\$475,200	\$498,960
模型期在建工程增加			\$57,186	\$180,000	\$72,000	\$43,200	\$23,760	-
转入已投产资产			\$62,216	\$220,000	\$408,000	\$484,800	\$530,640	\$554,400
技术基础设施（已投产原值）	\$141,852	\$203,679	\$309,393	\$529,393	\$937,393	\$1,422,193	\$1,952,833	\$2,507,233
办公空间（已投产原值）	\$45,403	\$48,348	\$50,635	\$50,635	\$50,635	\$50,635	\$50,635	\$50,635
公司及其他资产（已投产原值）	\$12,574	\$14,463	\$6,498	\$6,498	\$6,498	\$6,498	\$6,498	\$6,498
已投产固定资产合计	\$199,829	\$266,490	\$366,526	\$586,526	\$994,526	\$1,479,326	\$2,009,966	\$2,564,366
固定资产原值（含在建）	\$250,426	\$345,082	\$546,526	\$946,526	\$1,426,526	\$1,954,526	\$2,508,926	\$3,063,326
折旧预测								
技术基础设施折旧年限			9.0	8.0	8.0	8.0	8.0	8.0
其他已投产资产折旧年限			20.0	20.0	20.0	20.0	20.0	20.0
平均技术基础设施原值			\$256,536	\$419,393	\$733,393	\$1,179,793	\$1,687,513	\$2,230,033
技术基础设施折旧			\$28,504	\$52,424	\$91,674	\$147,474	\$210,939	\$278,754
平均其他已投产资产原值			\$59,972	\$57,133	\$57,133	\$57,133	\$57,133	\$57,133
其他资产折旧			\$2,999	\$2,857	\$2,857	\$2,857	\$2,857	\$2,857
折旧合计	\$15,311	\$21,136	\$31,503	\$55,281	\$94,531	\$150,331	\$213,796	\$281,611
折旧同比增速		38.0%	49.0%	75.5%	71.0%	59.0%	42.2%	31.7%
折旧 / CapEx	29.1%	23.1%	15.8%	13.8%	19.7%	28.5%	38.6%	50.8%

资料来源：公司财报，国信证券经济研究所整理测算

亚马逊云拆分

- 预计26/27年公司AI云收入有望达到316/767亿美元，占云收入比例达到18%/31%。AWS AI云收入结构与其他公司不同，26年的重要增长驱动来自Bedrock平台Claude模型贡献的MaaS收入，预计在26Q1 150亿AI云ARR中占比达到1/3。IaaS层主要依赖26年起Trainium芯片放量，因此26-27年预计公司收入有望迎来加速期。
- 利润端：由于MaaS业务占比更高，有望帮助亚马逊保持更好的OPM

表：亚马逊云收入利润拆分

单位：百万美元	2024	2025	2026E	2027E	1Q25	2Q25	3Q25	4Q25	1Q26	2Q26	3Q26E	4Q26E	1Q27E	2Q27E	3Q27E	4Q27E
AWS收入	107,556	128,725	179,225	260,495	29,267	30,873	33,006	35,579	37,587	42,232	46,814	52,592	55,995	62,917	67,865	73,718
YOY	19%	20%	39%	45%	17.0%	17.5%	20.2%	24.0%	28%	37%	42%	48%	49%	49%	45%	40%
OP	39,834	45,606	68,027	98,293	11,547	10,160	11,434	12,465	14,161	16,621	17,511	19,734	21,066	23,721	25,630	27,876
OPM	37.0%	35.4%	38.0%	37.7%	39.5%	32.9%	34.6%	35.0%	37.7%	39.4%	37.4%	37.5%	37.6%	37.7%	37.8%	37.8%
传统云	104,806	122,034	147,676	177,212	28,392	29,598	31,019	33,025	33,837	35,982	37,709	40,148	40,604	43,178	45,251	48,178
YOY	16%	16%	21%	20%	16%	16%	16%	18%	19%	22%	22%	22%	20%	20%	20%	20%
OP	38,829	43,064	55,714	65,568	11,217	9,678	10,679	11,490	12,711	14,196	13,952	14,855	15,024	15,976	16,743	17,826
OPM	37.0%	35.3%	37.7%	37.0%	39.5%	32.7%	34.4%	34.8%	37.6%	39.5%	37%	37%	37%	37%	37%	37%
AI云收入	2,750	6,691	31,548	83,283	875	1,275	1,987	2,554	3,750	6,250	9,105	12,444	15,391	19,739	22,614	25,540
YOY	400%	143%	371%	164%					329%	390%	358%	387%	310%	216%	148%	105%
占比	3%	5%	18%	32%	3.0%	4%	6%	7%	10%	15%	19%	24%	27%	31%	33%	35%
ARR									15,000	25,000						
OP	1,005	2,542	12,313	32,724	330	482	755	975	1,450	2,425	3,559	4,879	6,042	7,745	8,887	10,050
OPM	37%	38%	39%	39%	38%	38%	38%	38%	39%	39%	39%	39%	39%	39%	39%	39%
AI IaaS	2750	5,739	15,798	36,158	796	1,154	1,698	2,091	2,500	3,750	4,355	5,194	6,328	8,864	9,926	11,040
YOY	400%	109%	175%	129%					214%	225%	156%	148%	153%	136%	128%	113%
AI IaaS OP	1,005	2,161	6,013	13,874	299	434	639	789	950	1,425	1,659	1,979	2,417	3,395	3,812	4,250
OPM	36.6%	37.7%	38.1%	38.4%	37.5%	37.6%	37.7%	37.8%	38.0%	38.0%	38.1%	38.1%	38.2%	38.3%	38.4%	38.5%
GPU		2,689	6,128	11,780	396	554	798	941	1,000	1,500	1,655	1,974	2,278	3,014	3,176	3,312
YOY			128%	92%					152%	171%	107%	110%	128%	101%	92%	68%
OP		941	2,145	4,123	139	194	279	329	350	525	579	691	797	1,055	1,112	1,159
OPM		35%	35%	35%	35%	35%	35%	35%	35%	35%	35%	35%	35%	35%	35%	35%
占AI IaaS比例		47%	39%	33%	50%	48%	47%	45%	40%	40%	38%	38%	36%	34%	32%	30%
Trn		3,050	9,670	24,378	400	600	900	1,150	1,500	2,250	2,700	3,220	4,050	5,850	6,750	7,728
YOY			217%	152%					275%	275%	200%	180%	170%	160%	150%	140%
占AI IaaS比例			53%	67%												
OP		1,220	3,868	9,751	160	240	360	460	600	900	1,080	1,288	1,620	2,340	2,700	3,091
OPM		40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%
MaaS (Bedrock)	952	15,750	47,125		79	121	289	463	1,250	2,500	4,750	7,250	9,063	10,875	12,688	14,500
YOY			1554%	199%					1487%	1963%	1542%	1466%	625%	335%	167%	100%
占AI云比例			50%	57%	9%	10%	15%	18%	33%	40%	52%	58%	59%	55%	56%	57%
ARR									5000	10000	19000	29000	36250	43500	50750	58000
OP		381	6,300	18,850	32	48	116	185	500	1,000	1,900	2,900	3,625	4,350	5,075	5,800
OPM		40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%	40%

资料来源：公司财报，国信证券经济研究所整理测算

阿里云拆分

- 阿里云收入加速和利润率的提升主要来自以下AI云的驱动，具体来看：
 1. GPU租赁：提价+自研芯片的放量缓解供给压力，推动收入增长的同时，毛利率也将持续优化
 2. MaaS：随着Token调用量的爆发，预计6月底ARR已达到100亿元，MaaS收入占比的提升也将推动整体利润率改善。
- 未来三年，预计阿里云收入中，AI云收入占比将超过外部云收入的60%，也将推动阿里云收入保持复合约45%的增长。同时，利润端伴随AI云Margin对整体的带动，将提升至约18%。

表：阿里云收入利润拆分

单位：百万元人民币	FY2025	FY2026	FY2027E	FY2028E	FY2029E	1QFY26	2QFY26	3QFY26	4QFY26	1QFY27E
	2024	2025	2026	2027	2028	2Q25	3Q25	4Q25	1Q26	2Q26
收入	118,028	158,132	235,312	349,236	500,088	33,398	39,824	43,284	41,626	48,518
YoY	11%	34%	49%	48%	43%	26%	34%	36%	38%	45%
外部	83,698	111,043	169,388	256,941	370,875	24,092	27,791	29,331	29,829	35,489
yoy	10%	33%	53%	52%	44%	26%	29%	35%	40%	47%
占云收入										
AI云收入	11,260	28,263	71,121	143,935	243,178	4,914	6,578	7,800	8,971	12,284
yoy	135%	151%	152%	102%	69%	160%	160%	160%	134%	150%
占外部云收入	13%	25%	42%	56%	66%	20%	24%	27%	30%	35%
GPU租赁		19,687	38,967	62,347	93,521	3,440	4,605	5,460	6,182	8,256
YoY			98%	60%	50%				130%	140%
占AI云比例		70%	55%	43%	38%	70%	70%	70%	69%	67%
毛利		5,015	12,304	21,886	33,668	860	1,059	1,365	1,731	2,477
毛利率		25%	32%	35%	36%	25%	23%	25%	28%	30%
MaaS (含百炼+API调用)		5,658	25,300	69,250	127,450	983	1,316	1,560	1,800	2,800
YoY			347%	174%	84%				134%	185%
占AI云比例		20%	36%	48%	52%	20.0%	20.0%	20.0%	20.1%	22.8%
毛利		1,698	8,855	24,238	44,608	295	395	468	540	980
毛利率		30%	35%	35%	35%	30%	30%	30%	30%	35%
其他		2,918	6,854	12,337	22,207	491	658	780	989	1,229
YoY			135%	80%	80%				157%	150%
占AI云比例		10%	10%	9%	9%	10%	10%	10%	11%	10%
毛利		1,751	4,112	7,402	13,324	295	395	468	593	737
毛利率		60%	60%	60%	60%	60%	60%	60%	60%	60%
AI云毛利		8,463	25,272	53,526	91,599	1,450	1,848	2,301	2,864	4,194
AI云毛利率		30%	36%	37%	38%	29.5%	28.1%	29.5%	31.9%	34.1%
传统云收入	72,438	82,780	98,266	113,006	127,697	19,178	21,213	21,531	20,858	23,205
yoy	2%	14%	19%	15%	13%	11%	12%	15%	19%	21%
占外部云收入	87%	75%	58%	44%	34%					
传统云毛利		20,695	24,567	28,252	31,924	4,795	5,303	5,383	5,215	5,801
传统云毛利率		25%	25%	25%	25%	25%	25%	25%	25%	25%
外部云综合毛利		29,158	49,838	81,778	123,524	6,244	7,152	7,684	8,079	9,995
外部云综合毛利率		26%	29%	32%	33%	26%	26%	26%	27%	28%
内部	34,330	47,089	65,925	92,294	129,212	9,306	12,033	13,953	11,797	13,028
yoy	13%	37%	40%	40%	40%	25%	49%	39%	34%	40%
占云收入										
内部云综合毛利		11,772	16,481	23,074	32,303	2,327	3,008	3,488	2,949	3,257
内部云毛利率		25%	25%	25%	25%	25%	25%	25%	25%	25%
阿里云综合毛利		40,931	66,320	104,851	155,827	8,571	10,160	11,172	11,028	13,252
阿里云综合毛利率		26%	28%	30%	31%	26%	26%	26%	26%	27%
经调EBITA	10556	14265	30,538	57,307	90,815	2,954	3,604	3,911	3,796	5,489
YoY	89%	35%	114%	88%	58%	26%	35%	25%	57%	86%
经调EBITA Margin	8.9%	9.0%	13.0%	16.4%	18.2%	8.8%	9.0%	9.0%	9.1%	11.3%

资料来源：公司财报，国信证券经济研究所整理测算

- [01] 中美AI投入规模及进展
- [02] 国产模型性价比优势凸显，融资规模持续扩大
- [03] 模型商业化闭环与盈利能力测算
- [04] 国内外云厂收入结构拆分
- [05] 风险提示

第一，宏观经济波动。若宏观经济波动，公司业务、产业变革及新技术的落地节奏或将受到影响。

第二，下游需求不及预期。若下游AI需求不及预期，相关的AI研发投入增长或慢于预期，致使行业增长不及预期。

第三，核心技术水平升级不及预期的风险。AI大模型研发进度落后，AIGC相关产业技术壁垒较高，核心技术难以突破，影响整体进度。

第四，AI快速迭代、平权化下竞争加剧，影响云业务利润率。

第五，高资本开支下，利率波动可能带来融资成本压力风险。

国信证券投资评级			
投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）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到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（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。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数（HSI.HI）作为基准；美国市场以标普500指数（SPX.GI）或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。



国信证券

GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046 总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032