下载日志已记录，仅供内部参考

传媒

报告日期：2026 年 09 月 01 日

人工智能数据周报：GLM-5.3-Flash 进入能力-成本 Pareto 最优区

——人工智能数据周报-20260831

投资要点

- **一、Artificial Analysis（数据截至：2026-08-31）**

  Artificial Analysis 综合能力指数覆盖 29 个代表模型（共 624 个）。Intelligence Index Top5 分别为 Claude Opus 5 (max)63 分、Claude Fable 5 (with fallback)62 分、GPT-5.6 Sol (max)61 分、Grok 4.6 (high)61 分、Kimi K3 (max)60 分。Agentic Index Top 5 分别为 Claude Opus 5 (max)59 分、GLM-5.3 (max)59 分、Grok 4.6 (high)59 分、GLM-5.3-Flash58 分、GPT-5.6 Sol (max)58 分。

  单任务成本：最高为 Claude Fable 5 (with fallback)$3.14；最低为 GPT-5.6 Luna (max)$0.05。

  单任务耗时：最低为 Gemini 3.5 Flash-Lite0.6；最高为 Qwen3.8 2.4T A95B13.8。

  能力-成本 Pareto 最优区本周覆盖模型：GLM-5.3-Flash、DeepSeek V4 Pro 0813 (max)、GPT-5.6 Luna (max)、Qwen3.8 27B (xhigh)、MiniMax-M3、Inkling。

  能力-速度 Pareto 最优区本周覆盖模型：Claude Opus 5 (max)、Claude Fable 5 (with fallback)、GPT-5.6 Sol (max)、Grok 4.6 (high)、Kimi K3 (max)、GLM-5.3 (max)、GPT-5.6 Terra (max)、Gemini 3.7 Flash (high)、DeepSeek V4 Pro 0813 (max)、GPT-5.6 Luna (max)、MiniMax-M3、Inkling。

- **二、OpenRouter 平台 Token 消耗量（数据区间：2026-08-24～2026-08-30）**

  上周平台总调用量 113.0T Token（环比+21.0%），API 调用次数 47.1 亿次（环比+10.2%）。

  Token 使用量 Top 5 模型：1）Ox Alpha15.7T（份额 13.9%，环比+35.6%）；2）DeepSeek V4 Flash 073112.3T（份额 10.9%，环比+6.3%）；3）MiMo-V2.59.1T（份额 8.1%，环比-8.1%）；4）GPT-5.6 Luna7.8T（份额 6.9%，环比+58.7%）；5）Hy36.7T（份额 5.9%，环比-18.9%）。

  市场份额 Top 5 提供商：deepseek23.5%（环比-1.7%）、google21.5%（环比+8.8%）、openai19.7%（环比+13.9%）、z-ai6.3%（环比新增）、qwen5.1%（环比+7.2%），合计 76.0%。

- **三、Epoch AI 全球数据中心与 AI 芯片（数据截至：2026-08-31）**

  全球追踪 86 个 AI 数据中心，合计等效算力 14.6M H100e，IT 总功率 13.3GW，累计资本支出$502B。最大单站 SpaceXAIColossus 2（1.11M H100e，946MW），前 5 大站点合计占 27%。主要运营商：Google20 个（4.61M H100e）、Meta15 个（2.69M H100e）、Microsoft7 个（1.69M H100e）。本周新增入列站点 3 座，在建 GW 级项目 9 座。

  全球 AI 芯片累计出货等效算力 30.5M H100e。供应商：Nvidia22160k H100e（累计占比 73%）、Google TPU4837k、Amazon1021k、AMD1740k、华为 708k。主力芯片：B300 累计 10704k H100e（出货 4237k 颗）、H1004344k H100e、B2006423k H100e。

- **四、Silicon Data（数据截至：2026-08-30）**

  GPU 租赁价格指数（USD/小时）：H100 新云$2.65（周环比+0.0%）、超大规模云$7.23（+0.0%），价差$4.58；H200$3.26（+0.6%）；A100 新云$1.60（+0.6%）、超大规模云$3.74（+0.0%）；B200$5.58（-1.6%）；MI300X$2.65（+0.0%）。LLM Token 支出指数$1.06（周环比+1.3%）。内存价格指数：RAM Index™（GDDR6 现货）$18.54（周环比+0.3%）。

- **风险提示：**第三方数据统计范围等数据偏差问题；AI 技术迭代不及预期；用户与收入增长不及预期；监管及安全风险；行业竞争加剧。

行业评级：看好(维持)

分析师：冯翠婷
执业证书号：S1230525010001
fengcuiting@stocke.com.cn

相关报告

1 《国产“模算协同”开启规模化验证，AI 精品剧、电影始引关注》 2026.08.31

2 《NeoCloud 看算力租赁新叙事，Harness 看国产大模型估值上修》 2026.08.24

3 《NeoCloud：从"紧缺"叙事走向"涨价与回报"兑现》 2026.08.17

下载日志已记录，仅供内部参考

图表目录

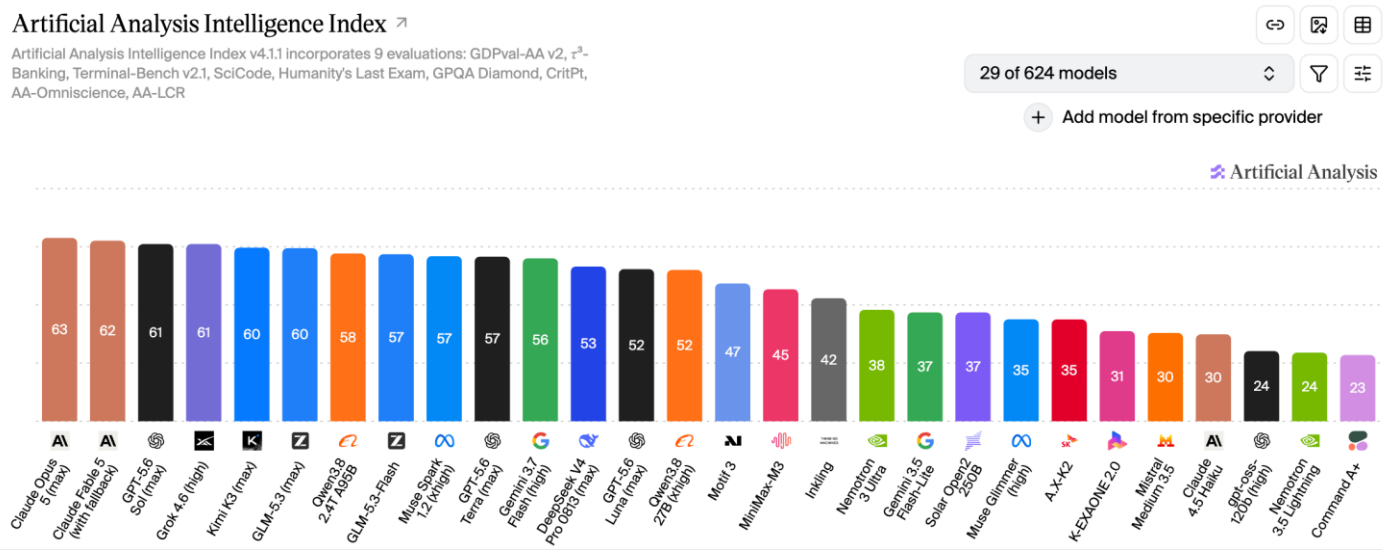
图 1: Artificial Analysis 人工智能指数（数据截至 2026-08-31）.....3
图 2: Artificial Analysis 智能体指数（数据截至 2026-08-31）.....3
图 3: 单任务推理成本对比（数据截至 2026-08-31）.....4
图 4: 能力-成本象限图（数据截至 2026-08-31）.....4
图 5: 单任务耗时对比（数据截至 2026-08-31）.....5
图 6: 能力-速度象限图（数据截至 2026-08-31）.....5
图 7: Openrouter 周度模型使用量趋势（2026.8.24-2026.8.30，本周各模型分配比重见图 8）.....6
图 8: Openrouter 周度模型 Token 使用量占比（2026.8.24-2026.8.30）.....6
图 9: 周度主要公司模型使用量折线图（2026.8.24-2026.8.30）.....7
图 10: 周度各公司市场份额百分比堆积图（2026.8.24-2026.8.30）.....7
图 11: 周度重点公司市场份额占比（2026.8.24-2026.8.30）.....8
图 12: 各厂商算力份额（2026.8.24-2026.8.30）.....8
图 13: 头部数据中心算力爬坡（2026.8.24-2026.8.30）.....9
图 14: 各厂商 AI 芯片累计出货等效算力（2026.8.24-2026.8.30）.....9
图 15: 各品类 AI 芯片累计出货等效算力（2026.8.24-2026.8.30）.....10

表 1: Silicon Data GPU 租赁价 / LLM Token 价 / 内存现货价基准指数及变化.....10

下载日志已记录，仅供内部参考

附录：

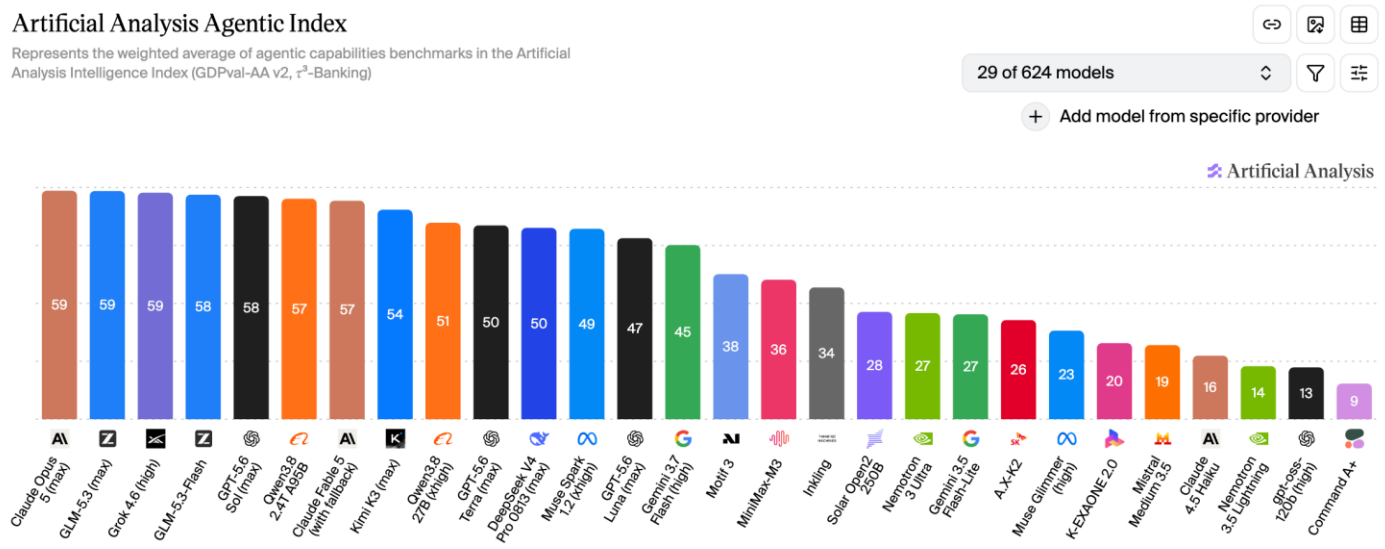
图1： Artificial Analysis 人工智能指数（数据截至 2026-08-31）



资料来源：Artificial Analysis，浙商证券研究所

注：综合能力指数由 GDPval-AA v2(真实经济价值任务)与 r2-Banking(金融代理任务)加权得出；单任务成本/耗时:按 Answer、Reasoning、Cache Write、Cache Hit、Input 五类 Token 加权平均；象限图:横纵坐标虚线交叉处为 Pareto 边界，绿区为综合最优。

图2： Artificial Analysis 智能体指数（数据截至 2026-08-31）

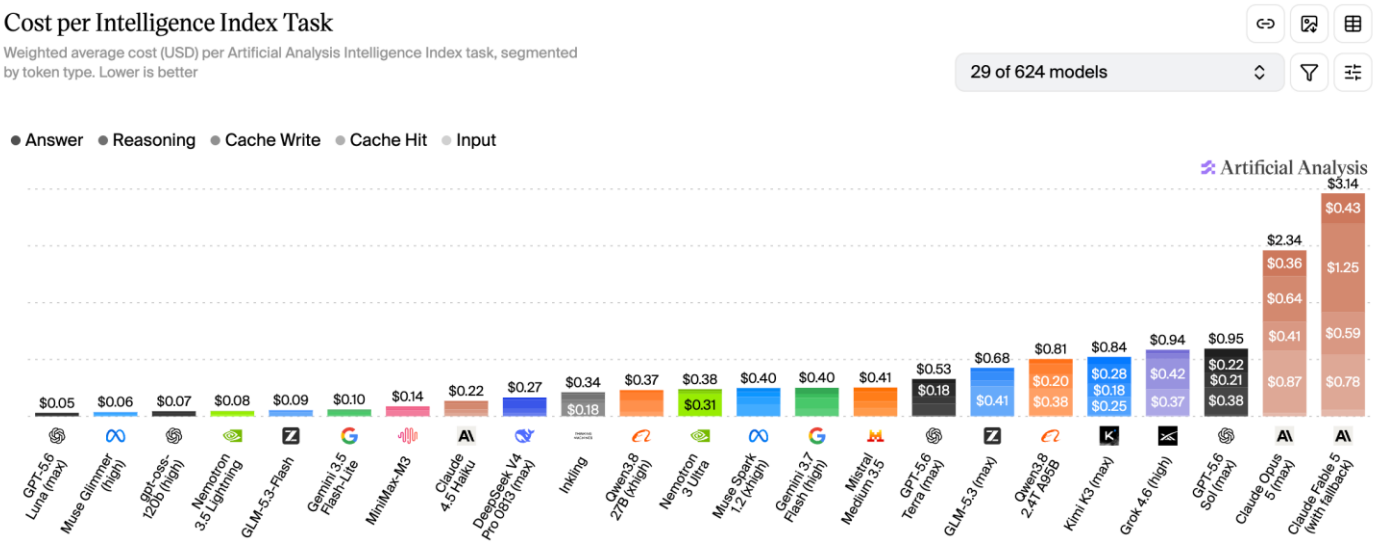


资料来源：Artificial Analysis，浙商证券研究所

注：综合能力指数由 GDPval-AA v2(真实经济价值任务)与 r2-Banking(金融代理任务)加权得出；单任务成本/耗时:按 Answer、Reasoning、Cache Write、Cache Hit、Input 五类 Token 加权平均；象限图:横纵坐标虚线交叉处为 Pareto 边界，绿区为综合最优。

下载日志已记录，仅供内部参考

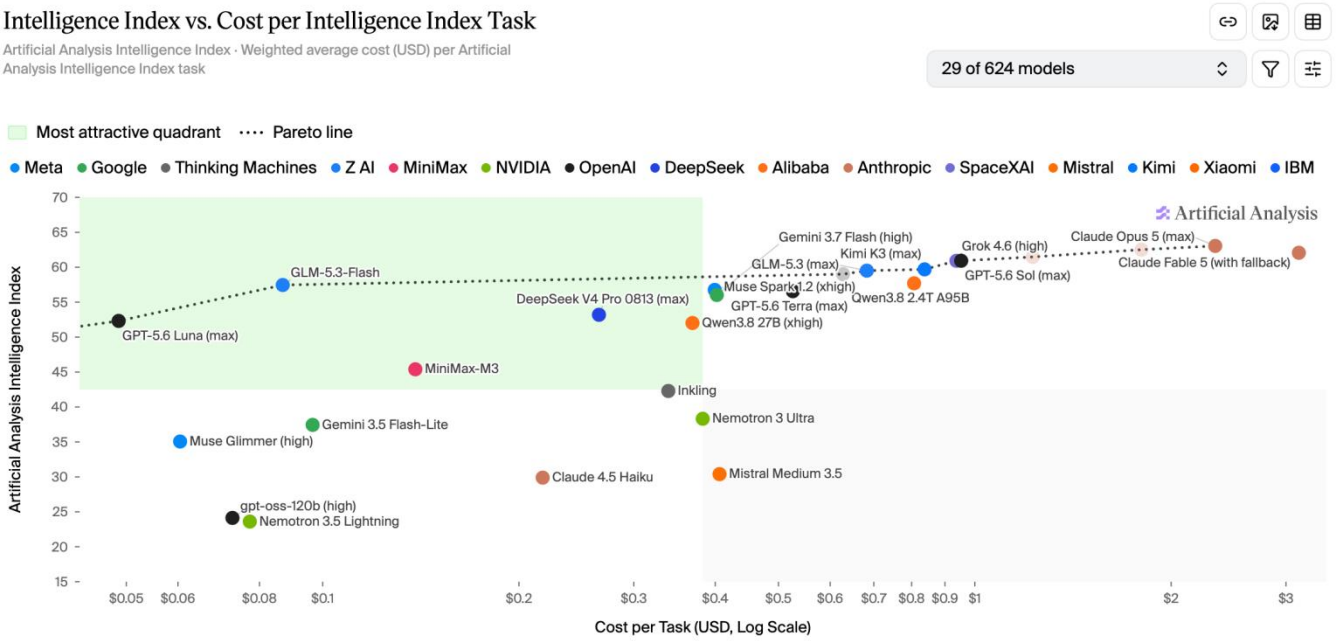
图3： 单任务推理成本对比（数据截至 2026-08-31）



资料来源：Artificial Analysis，浙商证券研究所

注：综合能力指数由 GDPval-AA v2(真实经济价值任务)与 r2-Banking(金融代理任务)加权得出；单任务成本/耗时:按 Answer、Reasoning、Cache Write、Cache Hit、Input 五类 Token 加权平均；象限图:横纵坐标虚线交叉处为 Pareto 边界，绿区为综合最优。

图4： 能力-成本象限图（数据截至 2026-08-31）

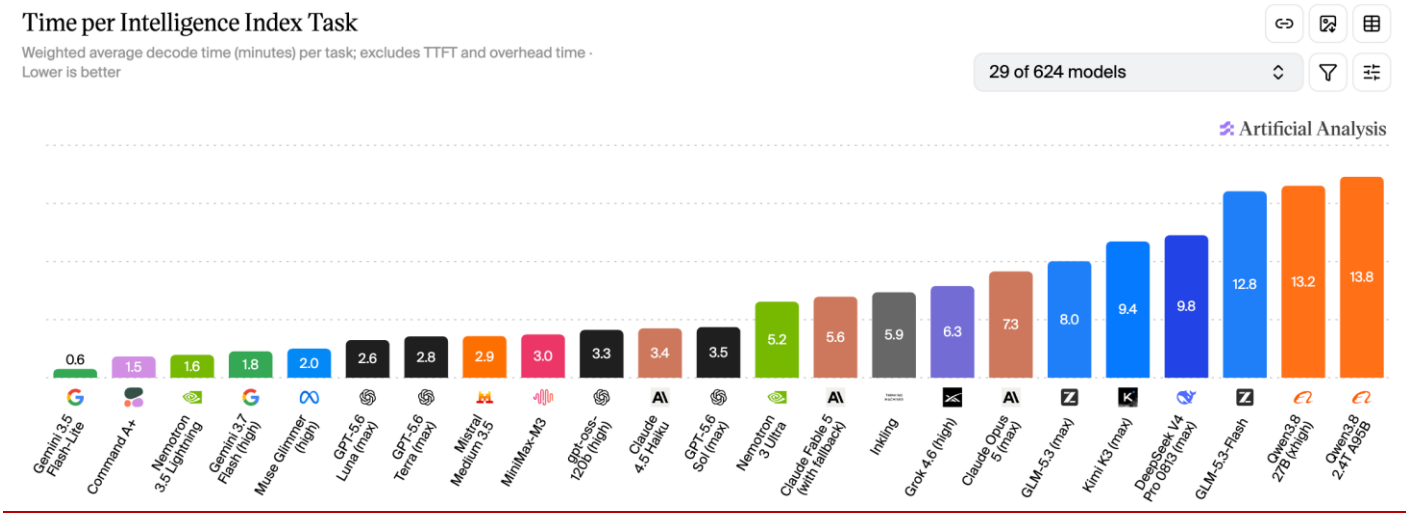


资料来源：Artificial Analysis，浙商证券研究所

注：综合能力指数由 GDPval-AA v2(真实经济价值任务)与 r2-Banking(金融代理任务)加权得出；单任务成本/耗时:按 Answer、Reasoning、Cache Write、Cache Hit、Input 五类 Token 加权平均；象限图:横纵坐标虚线交叉处为 Pareto 边界，绿区为综合最优。

下载日志已记录，仅供内部参考

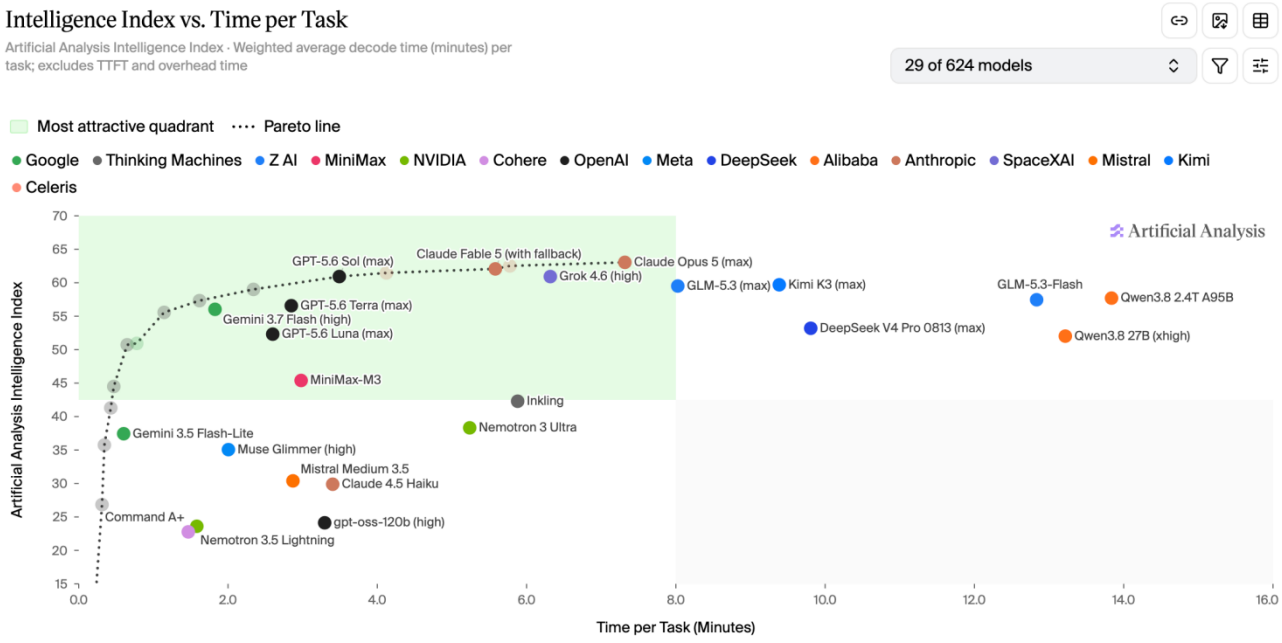
图5： 单任务耗时对比（数据截至 2026-08-31）



资料来源：Artificial Analysis，浙商证券研究所

注：综合能力指数由 GDPval-AA v2(真实经济价值任务)与 r2-Banking(金融代理任务)加权得出；单任务成本/耗时:按 Answer、Reasoning、Cache Write、Cache Hit、Input 五类 Token 加权平均；象限图:横纵坐标虚线交叉处为 Pareto 边界，绿区为综合最优。

图6： 能力-速度象限图（数据截至 2026-08-31）

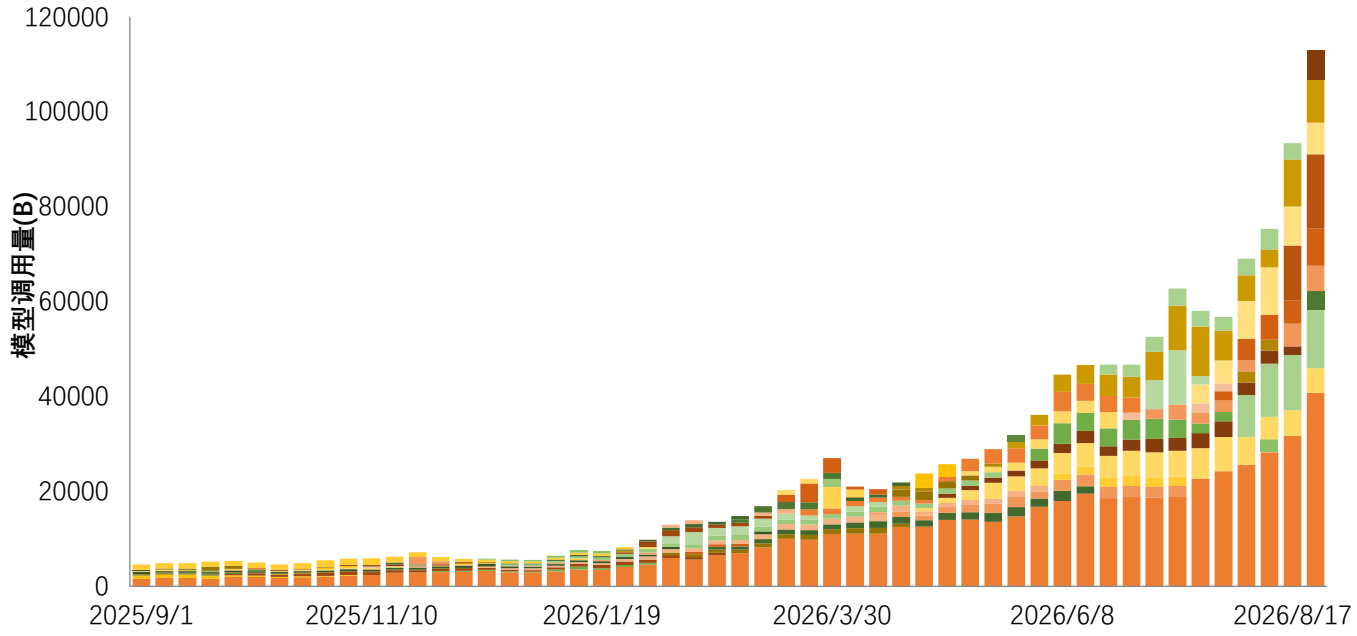


资料来源：Artificial Analysis，浙商证券研究所

注：综合能力指数由 GDPval-AA v2(真实经济价值任务)与 r2-Banking(金融代理任务)加权得出；单任务成本/耗时:按 Answer、Reasoning、Cache Write、Cache Hit、Input 五类 Token 加权平均；象限图:横纵坐标虚线交叉处为 Pareto 边界，绿区为综合最优。

下载日志已记录，仅供内部参考

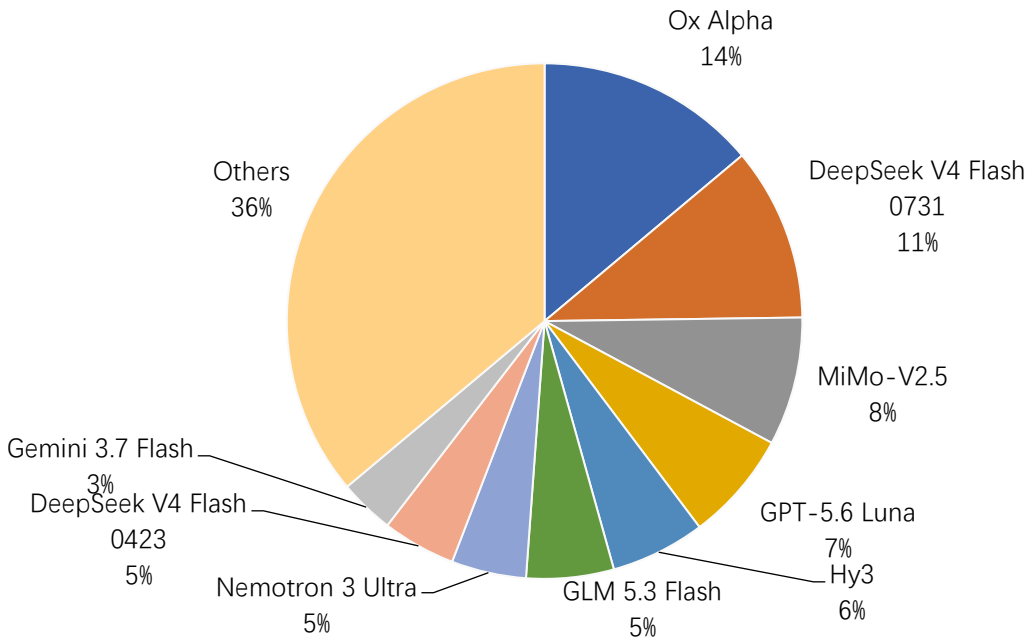
图7： Openrouter 周度模型使用量趋势（2026.8.24-2026.8.30，本周各模型分配比重见图 8）



资料来源：Openrouter，浙商证券研究所

注：OpenRouter 的数据仅为平台内生的第一方 API 调用元数据量，仅用作趋势观察，并不覆盖以下场景：1）各大模型原生平台（ChatGPT、Claude App、Gemini 等）的消费级应用流量；2）企业级 Copilot 产品（Microsoft 365 Copilot、GitHub Copilot 等）；3）通过其他 API 聚合平台或直连的模型调用；4）本地私有部署及离线推理。

图8： Openrouter 周度模型 Token 使用量占比（2026.8.24-2026.8.30）

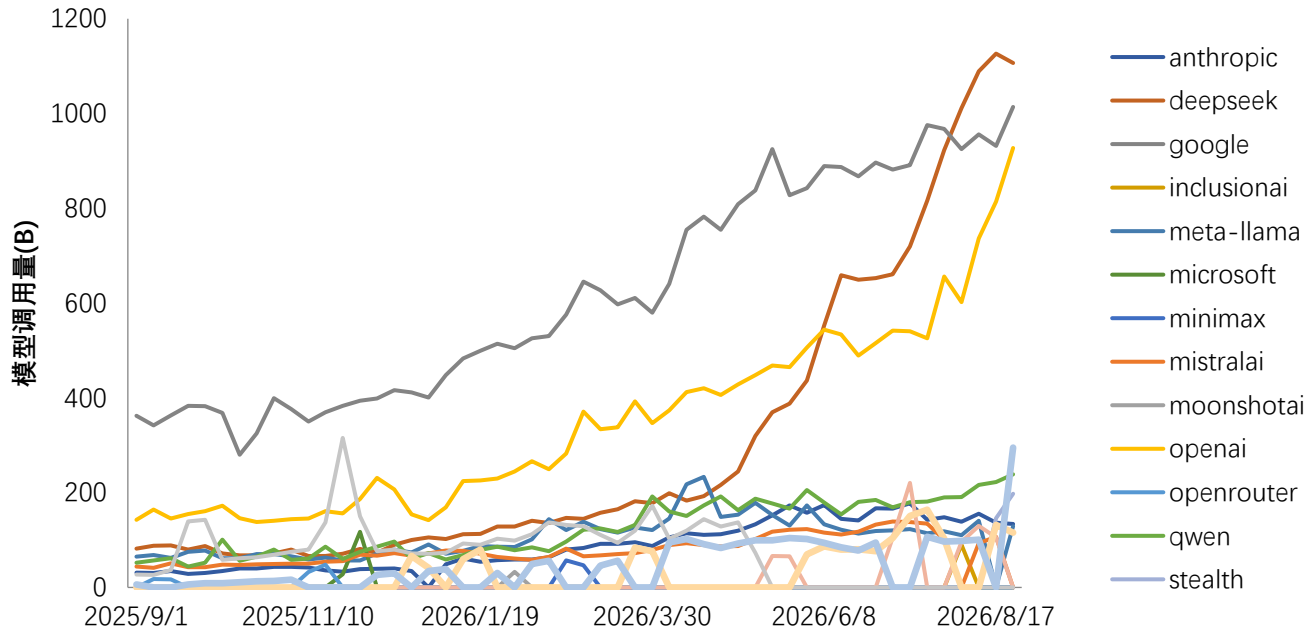


资料来源：Openrouter，浙商证券研究所

注：OpenRouter 的数据仅为平台内生的第一方 API 调用元数据量，仅用作趋势观察，并不覆盖以下场景：1）各大模型原生平台（ChatGPT、Claude App、Gemini 等）的消费级应用流量；2）企业级 Copilot 产品（Microsoft 365 Copilot、GitHub Copilot 等）；3）通过其他 API 聚合平台或直连的模型调用；4）本地私有部署及离线推理。

下载日志已记录，仅供内部参考

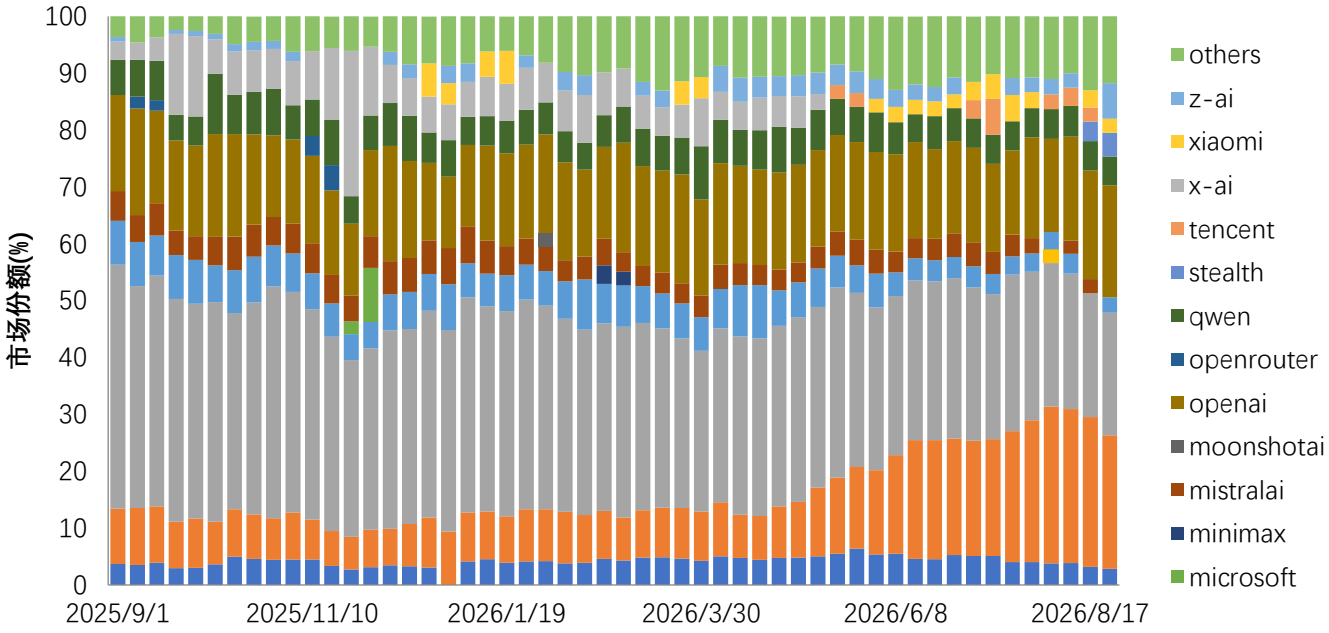
图9： 周度主要公司模型使用量折线图（2026.8.24-2026.8.30）



资料来源：Openrouter，浙商证券研究所

注：OpenRouter 的数据仅为平台内生的第一方 API 调用元数据量，仅用作趋势观察，并不覆盖以下场景：1）各大模型原生平台（ChatGPT、Claude App、Gemini 等）的消费级应用流量；2）企业级 Copilot 产品（Microsoft 365 Copilot、GitHub Copilot 等）；3）通过其他 API 聚合平台或直连的模型调用；4）本地私有部署及离线推理。

图10： 周度各公司市场份额百分比堆积图（2026.8.24-2026.8.30）

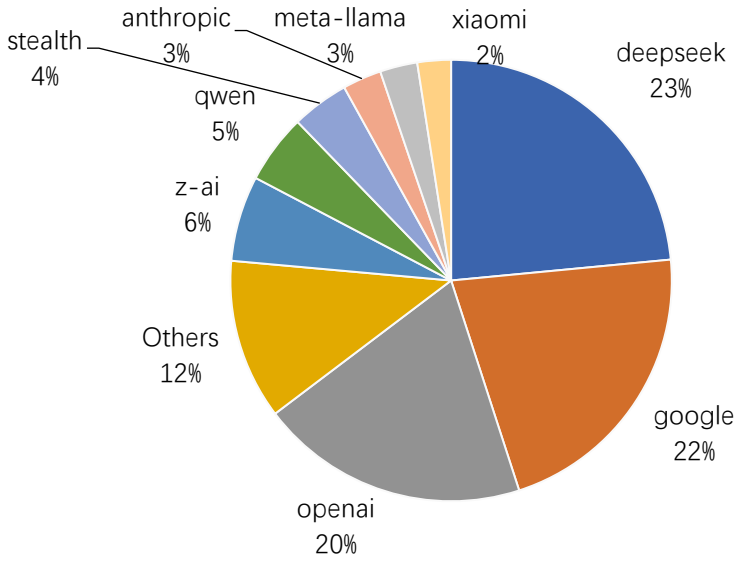


资料来源：Openrouter，浙商证券研究所

注：OpenRouter 的数据仅为平台内生的第一方 API 调用元数据量，仅用作趋势观察，并不覆盖以下场景：1）各大模型原生平台（ChatGPT、Claude App、Gemini 等）的消费级应用流量；2）企业级 Copilot 产品（Microsoft 365 Copilot、GitHub Copilot 等）；3）通过其他 API 聚合平台或直连的模型调用；4）本地私有部署及离线推理。

下载日志已记录，仅供内部参考

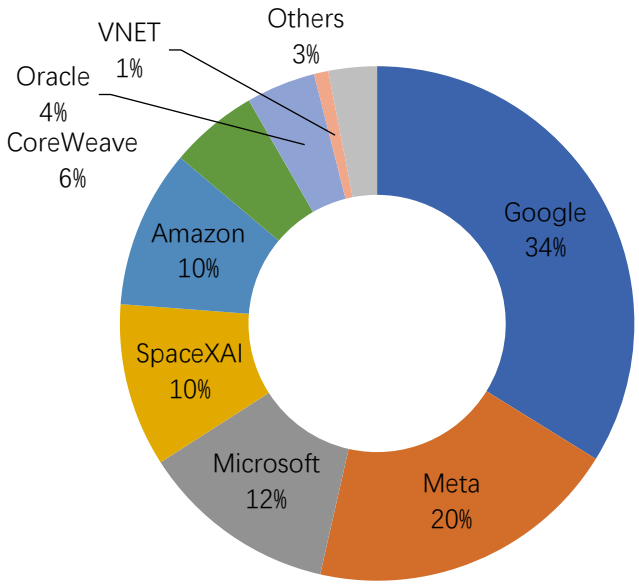
图11： 周度重点公司市场份额占比（2026.8.24-2026.8.30）



资料来源：Openrouter，浙商证券研究所

注：OpenRouter 的数据仅为平台内生的第一方 API 调用元数据量，仅用作趋势观察，并不覆盖以下场景：1）各大模型原生平台（ChatGPT、Claude App、Gemini 等）的消费级应用流量；2）企业级 Copilot 产品（Microsoft 365 Copilot、GitHub Copilot 等）；3）通过其他 API 聚合平台或直连的模型调用；4）本地私有部署及离线推理。

图12： 各厂商算力份额（2026.8.24-2026.8.30）

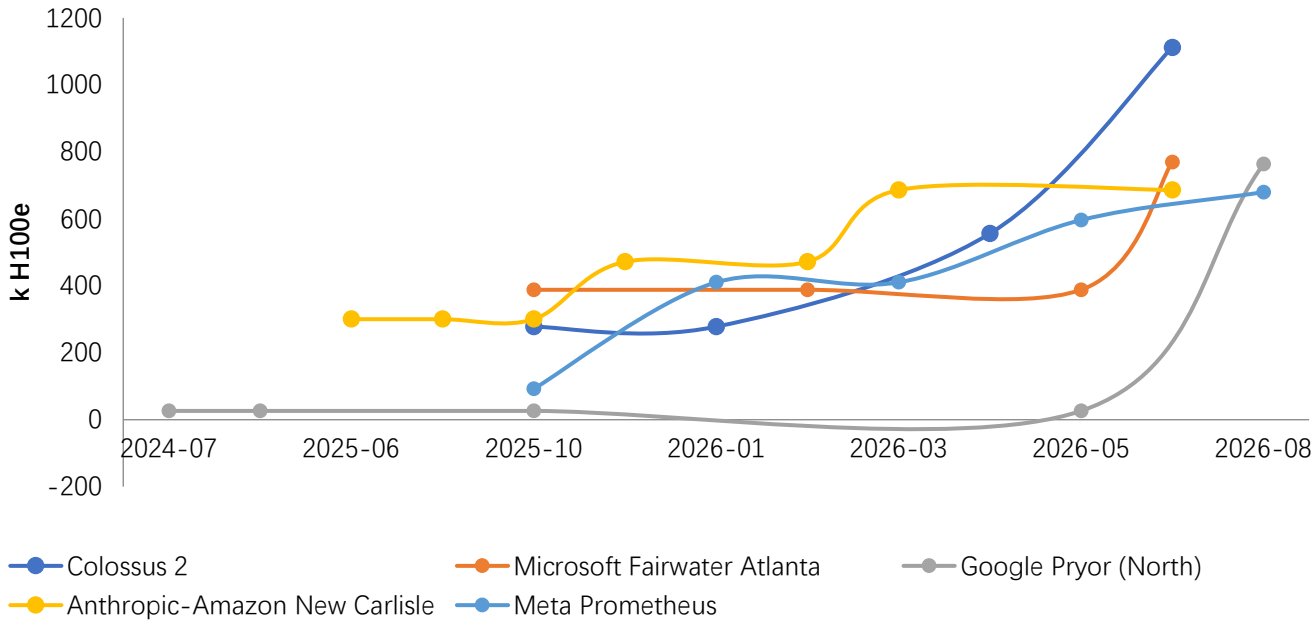


资料来源：Epoch AI，浙商证券研究所

注：Epoch AI 数据中心数据基于卫星图像、政府许可文件、SEC 文件及公司公告等公开渠道的交叉估算，非运营商实测数据。已追踪的站点高度集中于美国，不覆盖中国境内集群、中小型私有推理集群及边缘计算。所有算力/功率/成本均为模型估算值，芯片型号由公开信息推断。

下载日志已记录，仅供内部参考

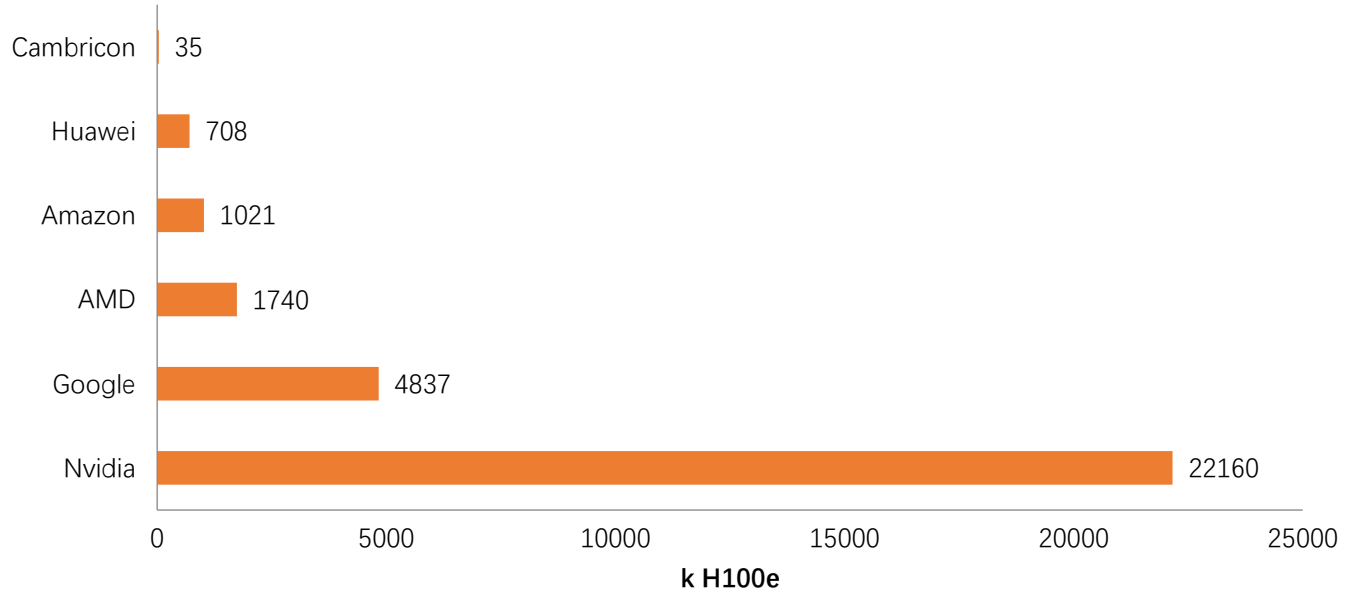
图13：头部数据中心算力爬坡（2026.8.24-2026.8.30）



资料来源：Epoch AI，浙商证券研究所

注：Epoch AI 数据中心数据基于卫星图像、政府许可文件、SEC 文件及公司公告等公开渠道的交叉估算，非运营商实测数据。已追踪的站点高度集中于美国，不覆盖中国境内集群（已匿名化）、中小型私有推理集群及边缘计算。所有算力/功率/成本均为模型估算值，芯片型号由公开信息推断。

图14：各厂商 AI 芯片累计出货等效算力（2026.8.24-2026.8.30）

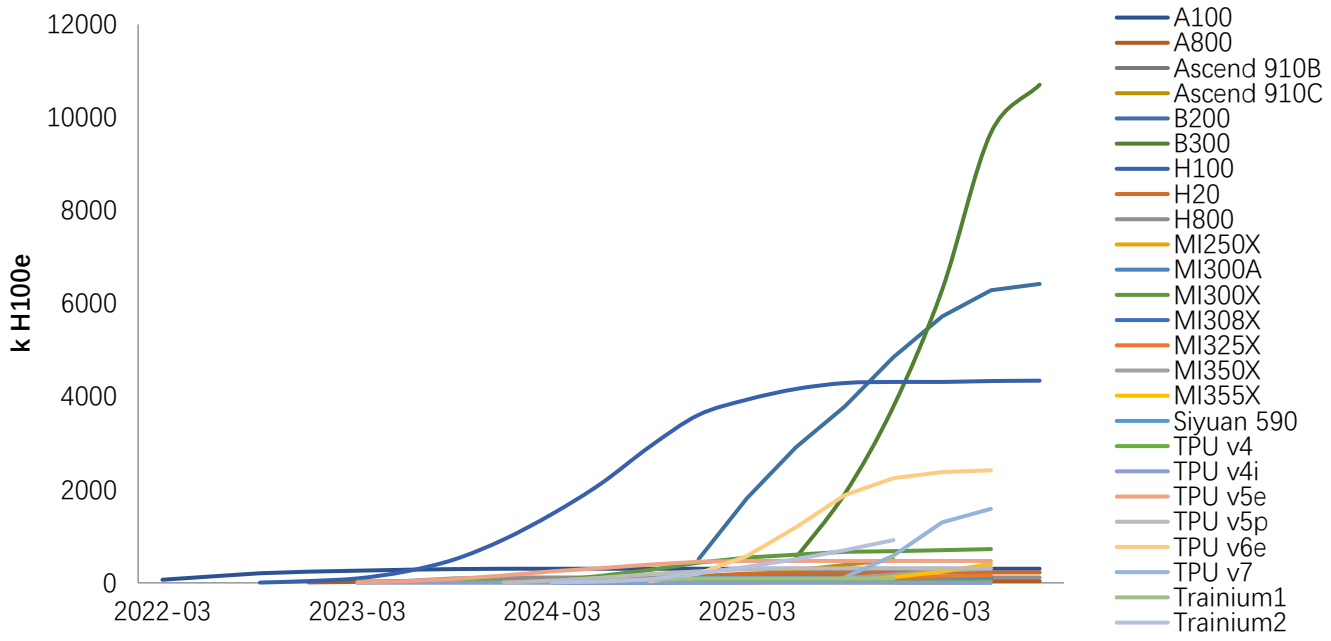


资料来源：Epoch AI，浙商证券研究所

注：Epoch AI 芯片数据基于厂商财报营收反推、第三方研究机构估算及公开技术规格，非厂商官方出货清单。Nvidia 数据自 2022 年起追溯、TPU 自 2023 年起，早期出货未纳入。华为昇腾/寒武纪数据来自零星媒体报道，可信度低于 Nvidia/Google 产品线；走私及非正规渠道芯片不在统计内。所有数值含中位数/5th/95th 百分位三档估值。H100e 折算基准：NVIDIA H100 SXM5 FP8 峰值算力（1,979 TFLOP/s）。

下载日志已记录，仅供内部参考

图15： 各品类 AI 芯片累计出货等效算力（2026.8.24-2026.8.30）



资料来源：Epoch AI，浙商证券研究所

注：Epoch AI 芯片数据基于厂商财报营收反推、第三方研究机构估算及公开技术规格，非厂商官方出货清单。Nvidia 数据自 2022 年起追溯、TPU 自 2023 年起，早期出货未纳入。华为昇腾/寒武纪数据来自零星媒体报道，可信度低于 Nvidia/Google 产品线；走私及非正规渠道芯片不在统计内。所有数值含中位数/5th/95th 百分位三档估值。H100e 折算基准：NVIDIA H100 SXM5 FP8 峰值算力（1,979 TFLOP/s）。

表1： Silicon Data GPU 租赁价 / LLM Token 价 / 内存现货价基准指数及变化

类别 / 名称	市场分层	价格 (USD)	7 日环比
GPU 租赁价格			
H100 Rental	Neo-Cloud	$2.65	▲ 0.0%
H100 Rental	Hyperscaler	$7.23	▲ 0.0%
H200 Rental	Neo-Cloud	$3.26	▲ 0.6%
A100 Rental	Neo-Cloud	$1.60	▲ 0.6%
A100 Rental	Hyperscaler	$3.74	▲ 0.0%
B200 Rental	Neo-Cloud	$5.58	▼ 1.6%
MI300X Rental	Neo-Cloud	$2.65	▲ 0.0%
LLM Token			
LLM Token Expenditure Index	—	$1.06	▲ 1.3%
内存 / 显存			
RAM Index™	GDDR6 现货现货	$18.54	▲ 0.3%

资料来源：Silicon Data，浙商证券研究所整理

注：1）覆盖范围：95% 主流新云（neo-cloud）、100% 主要超大规模云厂商（hyperscaler）；2）统计方法：每日采集 Cloud / Colocation / Brokered Cluster / Private Platform 四类原始报价-标准化为同口径 on-demand 价（按 GPU 型号、内存、租期、互联、地域对齐）-异常剔除 + 独立验证后计算指数；3）GPU 切片：H100 与 A100 同时发布 Neo-Cloud 与 Hyperscaler 两档指数以跟踪新云与超大规模云价差 4）数据截止日期 2026-08-30

下载日志已记录，仅供内部参考

股票投资评级说明

以报告日后的 6 个月内，证券相对于沪深 300 指数的涨跌幅为标准，定义如下：
1.买 入 ： 相对于沪深 300 指数表现 + 20 % 以上；
2.增 持 ： 相对于沪深 300 指数表现 + 10 % ~ + 20 %；
3.中 性 ： 相对于沪深 300 指数表现 - 10 % ~ + 10 % 之间波动；
4.减 持 ： 相对于沪深 300 指数表现 - 10 % 以下。

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。

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。
本报告版权均归本公司所有，未经本公司事先书面授权，任何机构或个人不得以任何形式复制、发布、传播本报告的全部或部分内容。经授权刊载、转发本报告或者摘要的，应当注明本报告发布人和发布日期，并提示使用本报告的风险。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

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
上海总部邮政编码：200127
上海总部电话：(8621) 80108518
上海总部传真：(8621) 80106010
浙商证券研究所：https://www.stocke.com.cn