综合

2026 年 11 月 22 日

投资评级：看好（维持）

海外科技：Anthropic 完成 IPO 上市，通用大模型切入机器人操作

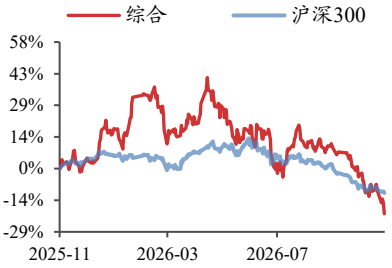
——行业周报

初敏（分析师）
chumin@kysec.cn
证书编号：S0790522080008

叶彬慧（分析师）
yebinhui@kysec.cn
证书编号：S0790526070007

李秋莹（联系人）
liqiuying@kysec.cn
证书编号：S0790126060038

行业走势图



数据来源：聚源

相关研究报告

《海外科技：通用算力需求提速，美国 AIDC 规划落地风险升温—行业周报》-2026.11.15

《大模型 10 月行业点评：中低档模型降价与 Agent 能力升级并进，国产模型密集发布—》-2026.11.10

《港股消费&应用：阜博紧抓 AI 短剧红利进入业绩加速期，美护三季报盈利延续分化 —行业周报》-2026.11.8

● **大模型：Anthropic IPO 上市，长程 Agent 能力持续强化，成本效率竞争加剧**

Anthropic 于 11 月 19 日在纳斯达克上市，发行价 83 美元/股，高于指导区间，发行总额约 935 亿美元，为史上规模最大的 IPO，首日收涨 17%，11 月 20 日收报 88.91 美元；2026Q3 收入 169.56 亿美元，环比增长约 47%，算力及基础设施费用占收入比重升至 58%。我们认为，Anthropic 收入高增的同时，算力成本与长期采购承诺持续扩大，算力供给、算力利用率及融资能力或将成为影响前沿模型厂商中长期盈利能力的关键变量。

● **物理 AI：通用大模型切入机器人操作，数据与推理效率成为落地关键**

NVIDIA 于 10 月发布的研究显示，GPT-6 Astra 在 RoboLab-120 仿真基准中取得 74.8%的任务成功率，高于研究发布时排行榜最佳策略 Lakeside 的 51.0%，但 Agentic Policy 方案成功回合的平均 API 成本约 36.40 美元，Direct Policy 方案成本降至 1.96 美元，单次推理平均耗时则达 18.56 秒；Anthropic 内部测试中，Claude Fable 5.2 在 RoboDojo Sim 获得 64.9%的平均成绩。Dyna 于 8 月发布的 Dyna-2 将人类操作视频预训练数据扩至 100 万小时，经少量机器人数据后训练，真机任务表现随数据规模持续提升。我们认为，通用大模型已在部分仿真机器人操作任务中展现较强能力，但推理成本、响应延迟及真实场景泛化仍是规模化应用的重要挑战，高质量数据供给与端侧部署效率有望成为具身智能竞争的重点。

● **云计算：美国封堵中国海外算力租用，云厂商竞争向运行与治理延伸**

美国将远程算力访问纳入出口管制，中国 AI 算力的外部半径被迫收缩。11 月 16 日 BIS 发布临时最终规则、19 日生效，境外数据中心向中国及 D:5 国家组实体提供超 1200 万 TPP 聚合算力须事先获许可，存量合同 90 天内补办；该门槛约合 750 颗 H100，而万亿参数预训练需数万卡集群，外训内推难以为继，国产算力链相对受益，独立模型公司训练不确定性上升。**云厂商竞争同步由 GPU 供给转向 Agent 运行与治理，**Ignite 2026 围绕 Agent 运行环境、数据连接与运维补齐 Cobalt 200、HorizonDB、SRE Agent 等全栈产品。我们认为，管制不改变本土算力需求方向，但先进算力可获得性成为模型梯队分化的核心变量，许可审批与国产芯片爬坡仍需跟踪。

● **投资建议：**

大模型与云计算：推荐标的：智谱、德适-B，受益标的：阿里巴巴、谷歌、亚马逊、微软、SpaceX、Meta、MiniMax。

● **风险提示：**模型及产品推出不及预期、AI 算力需求不及预期风险、算力供给超预期释放风险、宏观经济增长放缓。

目 录

1、 大模型：Anthropic 完成 IPO 上市，长程 Agent 能力持续强化，成本效率竞争加剧 3
1.1、 Anthropic：完成 IPO 上市，发布 Claude Fable 5.2 与 Mythos 5.2，收入高增与模型能力升级并行 3
1.2、 阿里巴巴：发布 Qwen4-Max，MoE 架构优化推理效率，长程 Agent 与专业工作能力升级 4
1.3、 OpenAI：发布 GPT-6.2 Astra 与 Sol，复杂专业任务能力升级，Token 效率及单任务成本进一步优化 5
2、 物理 AI：通用大模型切入机器人操作，数据与推理效率成为落地关键 6
3、 云计算：美国封堵中国海外算力租用，云厂商竞争向运行与治理延伸 7
3.1、 美国封堵中国海外算力租用，中国 AI 算力半径被迫收缩 7
3.2、 Microsoft Ignite 2026：Agent 平台化加速，云厂商竞争向运行与治理延伸 8
4、 港股周度更新 11
5、 模型与算力价格数据更新 13
5.1、 模型信息更新 13
5.2、 算力租赁价格更新 15
6、 投资建议 16
7、 风险提示 16

图表目录

图 1： Fable 5.2 在编程、业务工作流及科研四项评测中均领先 Opus 5.5 与 GPT-6 Astra 4
图 2： GPT-6 Astra 在 RoboLab-120 上成功率最高达 74.8%，高于同期榜首 Lakeside 的 51.0% 6
图 3： Direct Policy 设置下，改用低成本的 Sol、Luna 后，成功率由 Astra 的 72.7%降至 33.3%和 2.0% 7
图 4： Azure Chaos Studio 故障注入实验设计界面 9
图 5： Azure SRE Agent 可实现事件响应自动化 10
图 6： 本周恒生指数（2026.11.16-2026.11.20）在全球主要市场中相对跌幅较小（单位：%） 11
图 7： 本周（2026.11.16-2026.11.20）恒生科技跌幅居首（%） 11
图 8： 本周（2026.11.16-2026.11.20）恒生传媒、恒生创新药跌幅居前（%） 11
图 9： 2026 年 10 月底以来港股通当日买入成交净额有所波动（单位：亿元） 12
图 10： 恒生沪深港通 AH 股溢价指数有所回升 13
图 11： OpenRouter 的 token 调用量持续提升 14
图 12： 国产 AI 在 ZenMux 的 Token 市场份额约 50% 15
图 13： B200 租赁价格高位运行但有所回落（单位：$/小时） 15
图 14： H200 租赁价格高位运行但有所回落（单位：$/小时） 15
图 15： H100 租赁价格高位震荡（单位：$/小时） 16
图 16： A100 租赁价格基本持平（单位：$/小时） 16

表 1： 东南亚数据中心在新规则下发展前景受到影响 8
表 2： 本周（2026.11.16-2026.11.20）腾讯控股、中芯国际与华虹宏力流入资金居前 12
表 3： 10 月以来新上线模型 13

1、 大模型：Anthropic 完成 IPO 上市，长程 Agent 能力持续强化，成本效率竞争加剧

1.1、 Anthropic：完成 IPO 上市，发布 Claude Fable 5.2 与 Mythos 5.2，收入高增与模型能力升级并行

IPO 定价高于指导区间，超额配售权全额行使后发行规模创历史纪录。Anthropic 以 83 美元/股定价，高于此前 70 至 78 美元/股的指导区间，并于 11 月 19 日在纳斯达克全球精选市场上市，股票代码 ANT；按发行价计算，完全稀释估值约 1.5 万亿美元。11 月 20 日，公司公告承销商已全额行使超额配售权，最终公开发行 11.27 亿股 A 类普通股，发行总额约 935 亿美元，超过 SpaceX 于 2026 年 6 月创下的纪录，其中公司取得约 916 亿美元。同期完成 260 亿美元战略私募，投资方包括 AMD、博通、Google、Micron、Microsoft、NVIDIA 及 TSMC Partners，两项交易总额合计约 1,195 亿美元，均为扣除发行费用前金额。

上市首日高开后回落，收盘较发行价上涨 17%。11 月 19 日，公司股票开盘价 111.50 美元/股，较发行价上涨 34%，盘中最高约 115 美元/股，随后涨幅收窄，收盘报 97.17 美元/股，较发行价上涨 17%；首日成交股数超过 SpaceX 上市首日水平。11 月 20 日，公司股票收报 88.91 美元/股，较发行价仍上涨 7.1%，首日涨幅已回吐过半。

收入快速增长，企业客户规模与单客户贡献持续提升。2026Q3 公司实现收入 169.56 亿美元，环比增长约 47%，较 2026Q2 的约 144%增速放缓。截至 2026 年 9 月 30 日，过去 12 个月收入贡献达到或超过 10 万美元的客户超过 11,900 家，其中达到或超过 100 万美元的客户超过 1,900 家、达到或超过 1,000 万美元的客户约 250 家，净收入留存率约 900%。截至 2026 年 10 月 31 日，公司年化收入运行率（Run-Rate Revenue）约 810 亿美元，较 9 月底增加 100 亿美元，反映 Claude 整体商业化规模继续扩大。

算力费用增长快于收入，调整后利润率环比回落。2026Q3 公司算力及基础设施费用达 98.63 亿美元，环比增长约 77%，占收入比重由 2026Q2 的 48%升至 58%。同期 GAAP 营业利润由 2026Q2 盈利 9.74 亿美元转为亏损 5.67 亿美元；调整后营业利润为 24.63 亿美元，环比增长约 10%，但调整后营业利润率由 19%降至 15%。

长期算力承诺进一步扩大，基础设施投资与融资需求值得关注。最新招股书披露，截至 2026 年 9 月 30 日及期后，公司采购承诺与租赁协议对应的累计承诺支出约 9,009 亿美元，其中约 4,516 亿美元预计于 2030 年底前发生，自 2026 年 9 月 30 日起的 12 个月内到期金额为 [illegible] 亿美元。相关承诺绝大部分涉及算力及基础设施。公司通过 AWS Trainium、Google TPU、NVIDIA 及 AMD GPU 等多元算力资源支持模型训练和推理。

战略融资强化资金支持，特殊治理结构需要纳入估值考量。战略私募配售已于 IPO 后完成交割，进一步补充资金来源。治理方面，Anthropic 采用公益公司（PBC）结构，创始人通过 Founder LLC 持有的 F 类股享有一般股东表决事项 50.1%的投票权，长期利益信托则通过 T 类股选举和罢免 T 类董事，T 类董事占董事会多数。

Anthropic 发布 Claude Fable 5.2 与 Mythos 5.2，复杂编程、推理及视觉理解能

力进一步提升。11 月 9 日，Anthropic 推出 Claude Fable 5.2 和 Claude Mythos 5.2，两者采用相同底层模型，主要区别在于安全防护和开放范围。Fable 5.2 配置网络安全及生物安全防护，面向普通用户及开发者开放；Mythos 5.2 则面向通过网络安全或生命科学审核计划的机构提供。根据官方评测，Fable 5.2 在 FrontierCode 2.0 和 FrontierSWE v2 智能体编程评测中分别取得 49.1%和 69.9%的成绩，高于 Opus 5.5 的 45.5%和 62.3%；在 AutomationBench 业务工作流评测中取得 51.4%，高于 Opus 5.5 的 42.5%。

图1：Fable 5.2 在编程、业务工作流及科研四项评测中均领先 Opus 5.5 与 GPT-6 Astra

	Fable 5.2	Fable 5.1	Opus 5.5	GPT-6 Astra	Gemini 4 Carbon
Agentic coding FrontierCode 2.0	**49.1%** Xhigh effort	32.4% Xhigh effort	45.5%	42.8%	41.1% Medium effort
Agentic coding FrontierSWE v2	**69.9%**	56.3%	62.3%	65.5%	61.2%
Business workflows AutomationBench	51.4%	31.4%	42.5%	41.4%	**53.9%**
Agentic scientific research Terminal-Bench-Science 0.1	**74.6%** 75.3% (Mythos 5.2)	52.6%	58.7%	68.1%	64.8%

资料来源：Anthropic 官网

长程编程与计算机操作能力提升，复杂工程任务执行表现改善。Fable 5.2 在 OSWorld 2.1 离线测试子集中取得 89.3%的成绩，高于 Opus 5.5 的 87.2%和 GPT-6 Astra 的 79.5%。在 Remote Labor Index 评测中，新模型得分为 41.3%，高于 Opus 5.5 的 21.3%。编程效率方面，Fable 5.2 在 FrontierCode 2.0 中采用 High 推理强度时，以约一半单任务成本达到接近 Opus 5.5 最高推理强度的表现；在 Xhigh 推理强度下得分达到 49.1%，进一步提升复杂软件工程任务表现。

科学研究与硬件工程能力拓展，支持更复杂的跨领域任务。Fable 5.2 在 Terminal-Bench Science 0.1 中取得 74.6%的成绩，Mythos 5.2 得分为 75.3%，高于 Opus 5.5 的 58.7%。此外，Anthropic 披露，新模型在硬件设计、数学证明形式化及理论物理研究等任务中取得进展，进一步扩展模型在复杂科学与工程工作流中的应用范围。

推理速度提升、Token 价格下降，快速模式进一步优化任务执行效率。相比 Fable 5.1 和 Mythos 5.1，新模型 Token 输出速度最高提升 30%，每百万 Token 输入和输出价格分别降至 8 美元和 40 美元。同时，Anthropic 推出可选 Fast Mode，Token 生成速度最高提升至标准模式的 2.5 倍，对应输入、输出价格分别为每百万 Token 16 美元和 80 美元。产品方面，Fable 5.2 已通过 Claude 平台及 AWS、Google Cloud 和 Microsoft Azure 提供服务，Mythos 5.2 则通过网络安全及生命科学审核计划开放。

1.2、 阿里巴巴：发布 Qwen4-Max，MoE 架构优化推理效率，长程 Agent 与专业工作能力升级

阿里巴巴发布 Qwen4-Max，原生视觉理解与复杂 Agent 能力进一步提升。11 月 16 日，阿里巴巴发布 Qwen4-Max，定位目前 Qwen 系列能力最强的模型，总参数量为 2.4 万亿，原生支持视觉理解及 100 万 Token 上下文。新模型重点提升编程、Agent 任务及视觉推理能力，在 Code Arena 真实编程排行榜中位列第 6，并在 Agents' Last Exam、Vals Index、JobBench 及 Chartography 等评测中取得较好表现。根据阿里自研 Work360 评测，模型覆盖 721 个专业领域的复杂任务，整体表现优于

Qwen3.8-Max，在 High 推理强度下 Token 消耗减少 35%，平均单任务成本降低约 60%。

升级稀疏 MoE 架构，减少激活参数并降低推理成本。Qwen4-Max 采用新一代 Qwen4 架构，对稀疏混合专家（MoE）模型的注意力机制、残差连接及嵌入层进行优化。在 2.4 万亿总参数规模下，每 Token 仅激活 710 亿参数，较同等规模的 Qwen3.8-Max 减少 25%，在提升模型能力的同时降低推理计算开销。Qwen4-Max 已通过 QwenCloud 的 Token Plan 及 API 开放，并接入 QwenWork 和 QwenStudio 等产品。定价方面，API 缓存命中输入价格为每百万 Token 0.075 美元，阿里计划后续开放 Qwen4 系列较小规模模型的权重。

专业工作流与云资源调度能力扩展，强化复杂任务的自主规划与执行。Qwen4-Max 新增云资源调度能力，能够通过闭环优化自主调整调度流程。根据官方基准测试，模型使数据标注环节的模型调用成本平均下降 9%，离线 CPU 重新调度速度提升超过 14 倍。此外，模型支持 3D 建模、游戏开发、Agent 视频创作及制造业生产流程等任务。在汽车制造场景的官方演示中，模型通过反馈驱动的迭代优化，将工件编程时间由约 475 分钟缩短至 24 分钟，降幅达 95%。

长程 Agent 执行能力进一步提升，支持持续运行的科学研究工作流。Qwen4-Max 能够执行长周期、多步骤科研任务，通过规划、工具调用及反馈迭代推进研究流程。官方披露，在芯片冷却材料研究中，模型连续运行 48 小时，经过四轮分子设计、合成测试及主动学习闭环，发现并通过实验验证一种冷却液分子。在阿尔茨海默病相关研究中，模型完成 454 次工具调用及 3,239 次指标评估，设计并优化 24 种候选分子，体现了长程 Agent 在科研工作流中的应用能力。

1.3、 OpenAI: 发布 GPT-6.2 Astra 与 Sol，复杂专业任务能力升级，Token 效率及单任务成本进一步优化

OpenAI 发布 GPT-6.2 Astra 与 Sol，重点提升复杂专业任务及日常工作流的执行能力。11 月 17 日，OpenAI 推出 GPT-6.2 系列，其中 Astra 定位前沿旗舰模型，面向复杂专业项目；Sol 定位高效通用模型，面向日常工作及编程任务。根据官方评测，GPT-6.2 Astra 在 Agents' Last Exam v1.1 长程专业工作流评测中，得分较 GPT-6 Astra 提高 5.2 个百分点，单任务成本降至其三分之一以下；较 Claude Fable 5.2 高 5.5 个百分点，成本仅为其 5%。在 Terminal-Bench-Science 0.1 科研任务评测中，Astra 较 GPT-6 Astra 提升 9 个百分点，单任务成本降至其三分之一以下。

编程与专业工作能力进一步提升，Sol 以较低成本接近 Astra 的编程表现。在 Cognition 的 FrontierCode 2.0 评测中，GPT-6.2 Sol 与 Astra 得分相差不到 0.5 个百分点，单任务成本约为 Astra 的五分之一。在覆盖金融、数据科学及生物技术等领域的 GDPval v2 专业工作流评测中，Astra 较 GPT-6 Astra 提升 6.2 个百分点，输出 Token 用量减少 22%；Sol 较 GPT-6.1 Sol 提升 5.3 个百分点。此外，Astra 在法律合同审查、财务分析、电子表格及幻灯片生成等任务中取得进展，在 Artificial Analysis 的 AA-AnalystAgent 评测中，62.5%的问题在全部五次独立尝试中均得到正确回答。

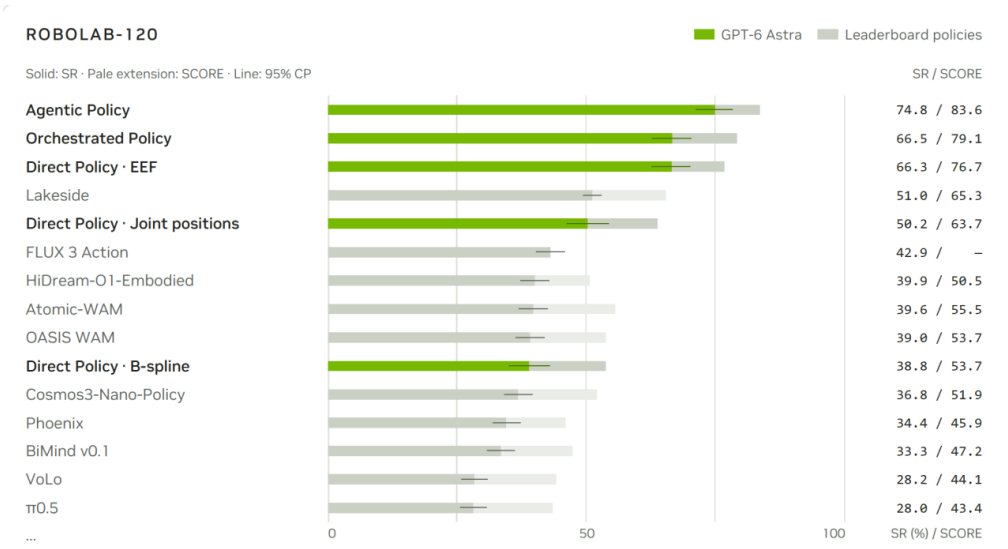
计算机操作与复杂工程任务能力拓展，视觉推理及交互式设计表现改善。GPT-6.2 Astra 强化了计算机操作、CAD 建模、机械装配及仿真设计能力，可根据文字需求构建交互式网页、3D 模型及工程工作流。在 Artificial Analysis 的 AA-3D 评测中，Astra 以 1,318 的 Elo 评分排名第一，高于 Claude Fable 5.2 的 1,271 分及 GPT-6.2 Sol 的 1,259 分。模型同时优化了长程任务规划、工具选择及代码生成效率，减少完成复杂工作流所需的 Token 消耗和交互轮次。

API 价格下降与 Token 效率改善同步推进，强化前沿模型的成本竞争力。GPT-6.2 Astra 每百万 Token 输入、缓存输入及输出价格分别为 8 美元、0.40 美元和 40 美元，较 GPT-6 Astra 分别下降 20%、60%和 20%；GPT-6.2 Sol 沿用 GPT-6.1 Sol 价格，对应输入、缓存输入及输出价格分别为 2 美元、0.10 美元和 10 美元。两款模型均已通过 OpenAI API、ChatGPT Work 及 Codex 向符合条件的用户开放，并支持超高速模式。安全方面，OpenAI 披露两款模型网络安全能力均达到其准备框架中的 Critical 阈值，并针对高风险网络安全任务加强了访问限制和安全防护。

2、 物理 AI：通用大模型切入机器人操作，数据与推理效率成为落地关键

通用大模型机器人操作能力提升，部分仿真任务表现超过专用模型。NVIDIA 于 10 月发布的 Astranomy 研究显示，GPT-6 Astra 在 RoboLab-120 仿真基准中取得 74.8%的任务成功率，高于研究发布时排行榜最佳策略 Lakeside 的 51.0%。Anthropic 公布的内部测试显示，Claude Fable 5.2 在 RoboDojo Sim 评测中取得 64.9%的平均成绩，高于同一测试设置下 GPT-6 Astra 的 29.8%。两项研究表明通用大模型已在部分机器人操作任务中展现较强能力，具身模型竞争格局或将发生变化。

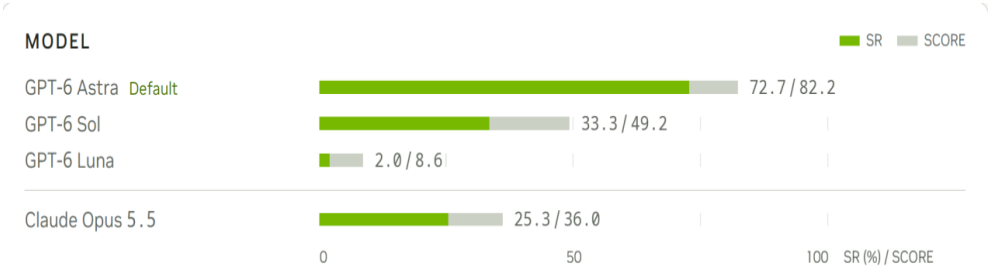
图2：GPT-6 Astra 在 RoboLab-120 上成功率最高达 74.8%，高于同期榜首 Lakeside 的 51.0%



资料来源：NVIDIA Research

推理成本与延迟仍是机器人实际部署的重要约束。NVIDIA 研究显示，GPT-6 Astra 采用 Agentic Policy 方案时，成功回合的平均 API 成本约 36.40 美元；Direct Policy 方案降至 1.96 美元，但平均单次推理耗时 18.56 秒。在 Direct Policy 设置下的 30 项任务子集对比实验中，GPT-6 Astra、Sol 和 Luna 的成功率分别为 72.7%、33.3%和 2.0%，显示换用低成本模型可能带来较大性能损失。当前高任务成功率与低成本、低延迟尚未同时实现，单纯换用低成本模型，或难以兼顾成本与成功率，推理效率仍是实际部署的重要瓶颈；上述结果均来自仿真评测，真实场景表现仍有待验证。

图3：Direct Policy 设置下，改用低成本的 Sol、Luna 后，成功率由 Astra 的 72.7%降至 33.3%和 2.0%



资料来源：NVIDIA Research

人类视频数据规模扩大有助于提升机器人任务表现，数据处理价值逐步显现。Dyna 于 8 月发布的 Dyna-2 研究显示，随着人类操作视频预训练数据由 1,000 小时扩张至 100 万小时，经少量机器人数据后训练（每项任务不超过 10 小时），模型在 14 项真机任务中的平均归一化得分由 20%提升至 53%，且预训练阶段未使用机器人数据。Dyna 指出，人类视频相较于遥操作数据更易规模化获取，但手部姿态提取与大规模标注仍存在挑战。人类视频有望成为机器人训练数据的重要补充，视频采集、数据清洗与手部姿态标注等环节的产业价值值得关注，数据质量及多样性仍是影响模型泛化能力的重要因素。

跨场景数据采集与跨本体训练持续推进，端侧推理有望成为重要技术方向。DROID 数据集包含约 7.6 万条机器人操作轨迹，覆盖 564 个场景；Open X-Embodiment 汇集超过 100 万条机器人轨迹、22 种机器人本体，研究显示跨机器人数据训练能够改善模型任务表现。此外，Google DeepMind 于 7 月推出 Gemini Robotics On-Device 2，针对机器人本地运行进行优化，并可在数小时内、通常以少于 200 个示例适配新的双臂机器人本体。跨场景数据积累、跨本体数据适配与端侧推理优化，有望共同改善具身模型的泛化能力及实际部署效率。

具身智能竞争逐步从模型能力延伸至数据与部署效率。现有关键结果仍主要来自研发方测试，如 RoboDojo Sim 与 Dyna-2。我们认为，通用大模型在部分仿真机器人操作任务中已取得较强表现，但推理成本、响应延迟及真实场景泛化能力仍是规模化应用的重要挑战。随着端侧模型持续发展，“云端规划、端侧执行”的混合架构有望成为重要技术方向。通用模型能力提升可能改变专用具身模型的竞争环境，第三方真机评测需求有望上升。具身智能产业链的竞争重点有望进一步拓展至高质量数据供给、模型泛化能力及实际部署效率，相关技术进展值得持续关注。

3、 云计算：美国封堵中国海外算力租用，云厂商竞争向运行与治理延伸

3.1、 美国封堵中国海外算力租用，中国 AI 算力半径被迫收缩

11 月 16 日，美国商务部工业与安全局（BIS）发布《关于远程访问先进半导体》的临时最终规则，并于 11 月 19 日正式生效，将远程访问先进计算能力纳入出口管

制范围。新规的核心在于将管制范围从芯片实物出口延伸至算力服务，实质上收紧了通过境外数据中心租赁先进算力的合规空间。具体而言，一是针对远程算力访问，若境外数据中心向总部或最终母公司位于中国（包括中国澳门）及 D:5 国家组的实体提供超过 1200 万 TPP 的聚合算力，相关服务提供方须在提供服务前申请出口许可，存量合同则需在规则生效后 90 天内补办许可；二是对印尼、马来西亚、新加坡、泰国及巴西五国进口受控 AI 芯片及服务器增加许可管控。

我们认为，此次规则调整意味着美国对华先进算力限制进一步从芯片供应端延伸至算力使用端，中国 AI 产业的发展约束正由先进芯片的外部采购能力，进一步转向本土先进芯片的制造与供给能力。与此同时，东南亚数据中心的行业逻辑亦面临重估：此前其主要受益于中国 AI 算力需求外溢，承接境外部署及算力租赁需求；新规落地后，相关数据中心不仅面临美国出口管制许可带来的合规不确定性，也可能受到中国客户租赁需求收缩的影响，行业增长逻辑由增量需求驱动转向合规约束与存量需求变化并存。

表1：东南亚数据中心在新规则下发展前景受到影响

地区	数据中心开发管线（MW）	在建容量（MW）	运营容量（MW）
马来西亚·柔佛	3088	602	1110
泰国·曼谷	2084	859	134
印尼·雅加达	1699	395	—

数据来源：CUSHMAN &WAKEFIELD、开源证券研究所

外训内推模型难以维系，国产算力更加重要。据蓉和半导体咨询测算，单颗英伟达 H100 芯片的 TPP 约 1.6 万，1200 万 TPP 门槛大致对应 750 颗 H100；若换成新一代英伟达 Blackwell 芯片，仅 300 余颗，体量大约是四五个机柜。而万亿参数级大模型预训练通常需要数万卡集群，这或会对国产大模型公司造成冲击，而国产算力链则有望受到国产替代红利。

公司层面的抗压能力差异被放大。具备自研芯片与充沛现金的厂商明显占优：阿里平头哥 2026 年上半年出货 13.3 万颗、位居国产第二，且公司计划在中长期将提升自研算力占比。更严格的算力租赁限制可能导致拥有更多算力资源的互联网大厂相对独立模型公司的竞争优势，缺乏自有算力储备的独立模型公司，其下一代模型训练的不确定性显著上升。

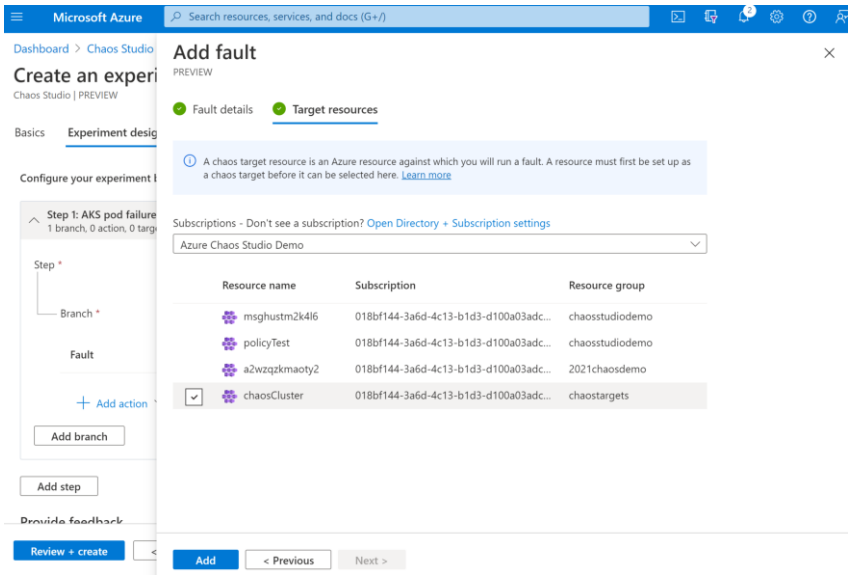
3.2、 Microsoft Ignite 2026：Agent 平台化加速，云厂商竞争向运行与治理延伸

Azure 加速构建从芯片到 Agent 的全栈基础设施体系，完善企业级 AI 规模化部署能力。AI 已从实验阶段进入生产部署阶段，Agent 的规模化运行需要底层芯片、计算基础设施、数据平台及运维管理等全栈能力协同支持。在本次 Ignite 大会中，微软围绕 Agent 运行环境、企业数据连接、安全治理及底层基础设施等完善产品体系。Azure Cobalt 200 虚拟机、Azure HorizonDB、Rayfin 及 Fabric Apps 正式商用；同时，微软推出多项预览功能，将 Microsoft Foundry 模型及 Agent 引入 Azure Kubernetes Service（AKS），并通过 Azure SRE Agent 和扩展的机密计算能力，强化 Agent 运行期间的运维管理与安全保障。我们认为，随着企业 AI 应用进入生产部署阶段，云厂商的竞争将从单一 GPU 算力供给进一步拓展至芯片、计算、数据、运维及安全等全栈能力，具备软硬件协同和企业级服务能力的云厂商有望获得更强的竞

争优势。

基础设施韧性与数据安全能力同步升级，为企业级 Agent 部署提供可靠性保障。微软进一步扩展 Azure Chaos Studio，新增 Chaos Studio Workspaces 及面向 Azure Kubernetes Service（AKS）的扩展功能，均已正式商业。前者支持企业定义待测试资源，并通过可复用场景模拟可用区故障、数据库故障切换等突发事件，持续验证应用的恢复能力；后者则支持对 AKS 中的 Kubernetes Pod 进行受控故障注入，并通过更少的权限要求提升测试安全性。此外，微软扩展机密计算产品体系，推出三项预览能力：VR72 多 GPU 机密计算支持扩展至单 GPU 之外的大规模模型保护；Azure Confidential Enclaves 及 AKS Kata CoCo Confidential Containers 通过在虚拟机内部构建更小范围的可信执行环境，降低安全信任边界。

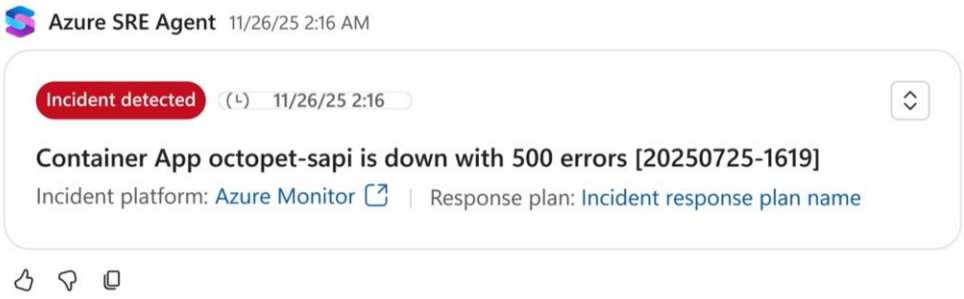
图4：Azure Chaos Studio 故障注入实验设计界面



资料来源：微软博客

云运维由人工监控向 Agent 辅助决策与自动化执行演进，Agent 自身也成为云平台的运维主体。微软进一步扩展 Azure Monitor 及 Azure SRE Agent 的 Agentic 运营能力。其中 Azure Monitor Health Models 可将遥测数据转化为实时健康关系图，关联应用架构、依赖关系及业务优先级，帮助运维人员识别故障根因、评估业务影响并制定修复方案；Azure SRE Agent 则进一步强化企业级安全与治理能力，新增客户管理密钥、Entra Agent ID 及 Agent 365 集成，并支持跨 Agent 通信、Microsoft Teams 集成和多云环境故障排查，相关能力处于预览阶段。Agentic 运营有望将 AI 能力从应用层延伸至云基础设施运维，通过 Agent 持续分析运行状态、识别异常及协同处理故障，提升云资源运营效率，同时推动 Agent 身份、权限及审计能力成为云运维体系的重要组成部分。

图5：Azure SRE Agent 可实现事件响应自动化

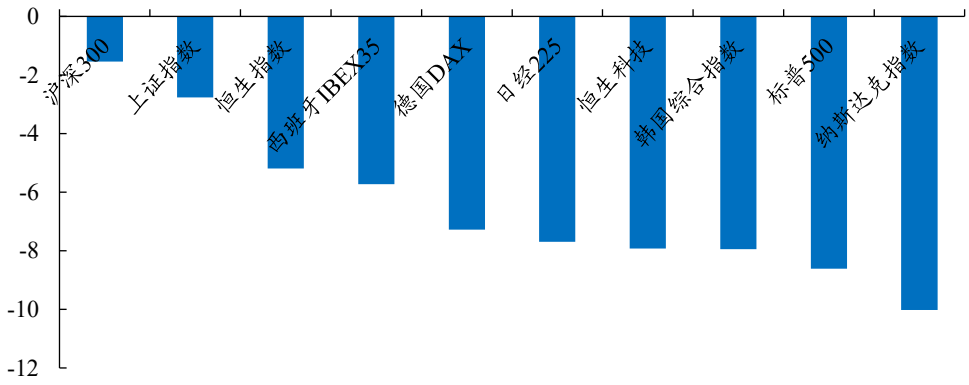


资料来源：微软博客

4、 港股周度更新

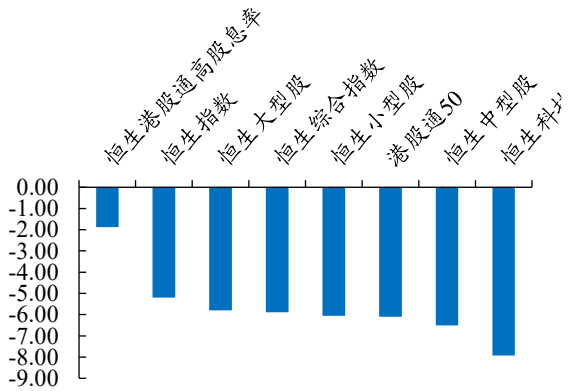
本周恒生指数下跌 5.19%，在全球主要市场中涨跌幅靠前，从成分上看，恒生传媒、恒生创新药跌幅居前，恒生科技本周下跌 7.92%。

图6：本周恒生指数（2026.11.16-2026.11.20）在全球主要市场中相对跌幅较小（单位：%）



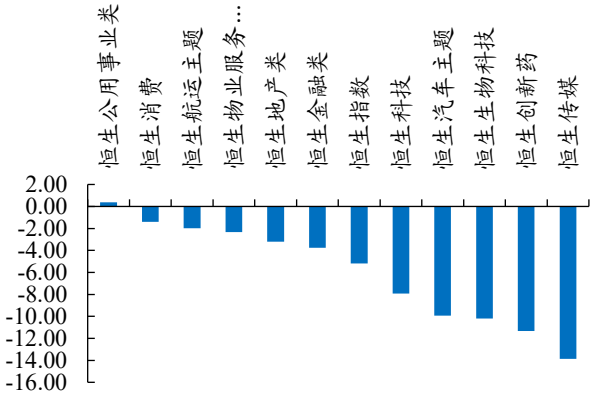
数据来源：Wind、开源证券研究所

图7：本周（2026.11.16-2026.11.20）恒生科技跌幅居首（%）



数据来源：Wind、开源证券研究所

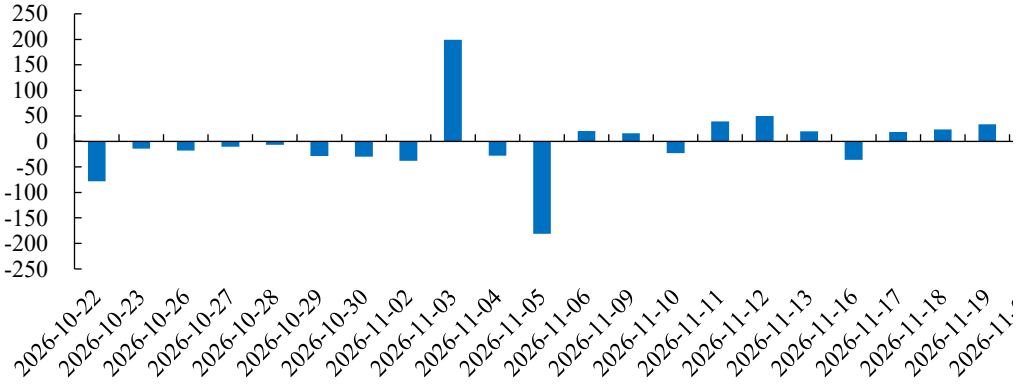
图8：本周（2026.11.16-2026.11.20）恒生传媒、恒生创新药跌幅靠前（%）



数据来源：Wind、开源证券研究所

本周港股通资金净流入人民币 140.46 亿元，规模较上周（+100.97 亿元）回升；但从头部活跃个股来看，腾讯控股（+77.5 亿港元）、中芯国际（+24.9 亿港元）、华虹宏力（+11.6 亿港元）资金净流入居前。

图9：2026 年 10 月底以来港股通当日买入成交净额有所波动（单位：亿元）



数据来源：Wind、开源证券研究所

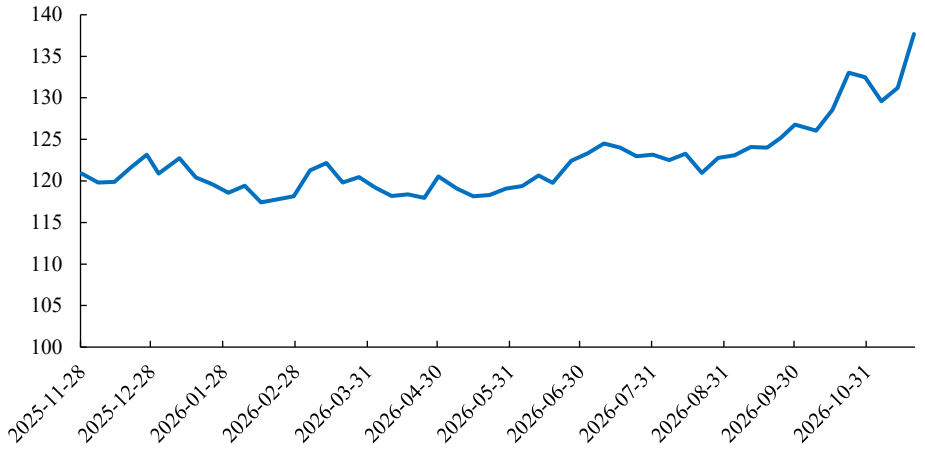
表2：本周（2026.11.16-2026.11.20）腾讯控股、中芯国际与华虹宏力流入资金居前

股票代码	股票名称	成交净买入（亿港元）	港股通持股数占港股股本比例
0700.HK	腾讯控股	77.5	12.0%
0981.HK	中芯国际	24.9	27.1%
1347.HK	华虹宏力	11.6	15.1%
0100.HK	MINIMAX-W	6.0	14.5%
0939.HK	建设银行	5.5	15.1%
1810.HK	小米集团-W	4.9	18.5%
9903.HK	天数智芯	2.4	8.6%
1024.HK	快手-W	1.9	20.3%
6166.HK	剑桥科技	1.8	63.0%
3308.HK	中际旭创	1.3	10.4%
9926.HK	康方生物	0.0	46.4%
1888.HK	建滔积层板	-0.3	9.6%
2513.HK	智谱	-0.8	17.8%
2269.HK	药明生物	-3.0	23.0%
6869.HK	长飞光纤光缆	-4.6	37.4%
2388.HK	中银香港	-5.8	3.6%
0883.HK	中国海洋石油	-7.0	22.4%
9988.HK	阿里巴巴-W	-17.0	9.6%

数据来源：Wind、开源证券研究所

本周恒生沪深港通 AH 股溢价指数为 137.68，环比有所上升，随着后续更多 A 股科技类标的赴港上市，或带动港股板块的“成长性溢价”持续向 A 股靠拢。

图10：恒生沪深港通 AH 股溢价指数有所回升



数据来源：Wind、开源证券研究所

5、 模型与算力价格数据更新

5.1、 模型信息更新

10 月以来全球大模型行业进入新品密集发布期，海内外头部厂商先后推出数十款新模型，产品定位与定价梯队分化明显。本周（2026.11.16-2026.11.20）新上线模型为阿里巴巴 Qwen4-Max、腾讯 Hy4 flash、OpenAI 的 GPT-6.2 Astra 与 GPT-6.2 Sol，以及谷歌的 Gemini 4 Neon。

表3：10 月以来新上线模型

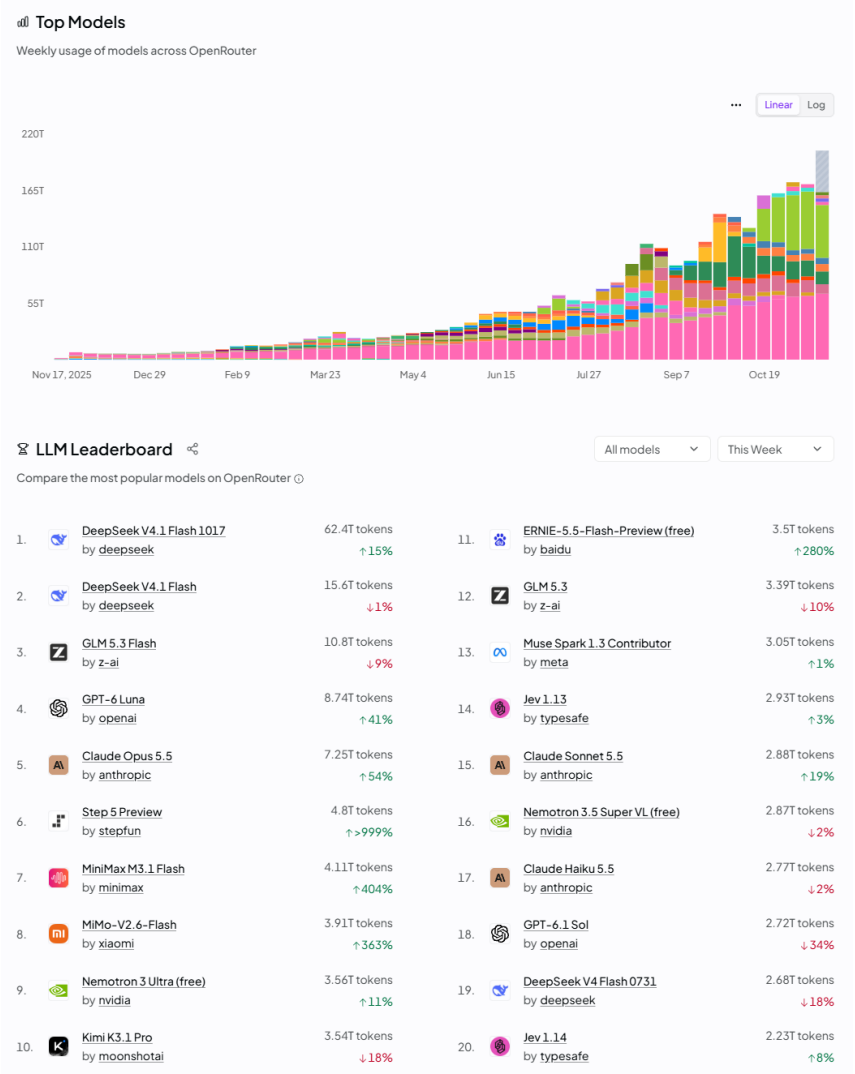
已发布模型	发布时间	厂商	输入价格（/M）	缓存命中输入价格（/M）	输出价格（/M）	上下文窗口
Gemini 4 Argon	2026-10-01	Google	$2.00	$0.10	$10.00	1M
Beam	2026-10-05	Reflection AI		暂无		1M
Claude Haiku 5.5	2026-10-08	Anthropic	$0.10	$0.01	$0.50	1M
Muse Wonder Full	2026-10-12	Meta		目前只在 Muse Code 订阅中开放		1M
Muse Wonder	2026-10-12	Meta	$4.00	$0.20	$20.00	1M
Muse Spark Open	2026-10-15	Meta	$1.25	$0.15	$4.25	1M
DeepSeek-V4.1-Flash-1017(高峰)	2026-10-17	DeepSeek	$0.30	$0.006	$1.20	1M
DeepSeek-V4.1-Flash-1017(空闲)	2026-10-17	DeepSeek	$0.15	$0.003	$0.60	1M
MiniMax M3.1 Flash	2026-10-19	MiniMax	$0.15	$0.015	$0.60	1M
Kimi K3.1 Pro Highspeed	2026-10-24	月之暗面	$4.00	$0.20	$20.00	1M
Kimi K3.1 Pro	2026-10-24	月之暗面	$2.00	$0.20	$10.00	1M
GPT-6.1 Terra	2026-10-27	OpenAI	$1.00	$0.05	$5.00	1.05M
Gemini 4 Carbon	2026-10-28	Google	$2.00	$0.10	$10.00	1M
LongCat-2.5	2026-11-02	美团	$0.75	$0.015	$1.20	1M
Hy 4	2026-11-04	腾讯	$0.83	$0.042	$2.50	1M

已发布模型	发布时间	厂商	输入价格（/M）	缓存命中输入价格（/M）	输出价格（/M）	上下文窗口
Gemini 4 Hydrogen	2026-11-12	Google	$1.50	$0.075	$7.50	1M
Qwen4-Max	**2026/11/16**	**阿里巴巴**	**$1.41**	**$0.071**	**$4.24**	**1M**
Hy4 flash	**2026/11/16**	**腾讯**	**$0.07**	**$0.003**	**$0.28**	**1M**
GPT-6.2 Astra	**2026/11/18**	**OpenAI**	**$8.00**	**$0.400**	**$40.00**	**1.05M**
GPT-6.2 Sol	**2026/11/18**	**OpenAI**	**$2.00**	**$0.100**	**$10.00**	**1.05M**
Gemini 4 Neon	**2026/11/18**	**Google**	**$3.00**	**$0.150**	**$15.00**	**1M**

资料来源：各模型厂商官网、IT 之家、开源证券研究所

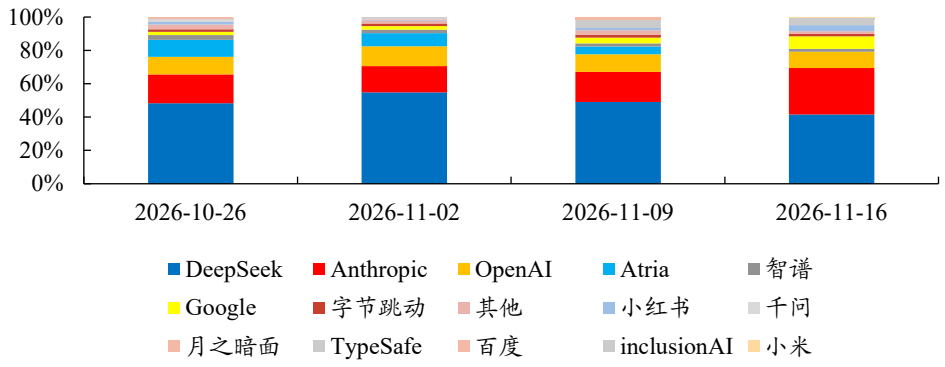
本周（2026.11.16-2026.11.20）OpenRouter 的 token 调用量持续提升，其中国产模型 DeepSeek V4.1 Flash 1017 与 GLM 5.3 Flash 位居前列。截至 11 月 16 日，在 ZenMux 统计的 Token 市场份额中，国产 AI 占比约 50%，DeepSeek 市场份额保持第一，占比为 41.55%。依托开源生态及出色的性价比优势，国产模型持续扩大市场影响力。

图11：OpenRouter 的 token 调用量持续提升



资料来源：OpenRouter 官网

图12：国产 AI 在 ZenMux 的 Token 市场份额约 50%



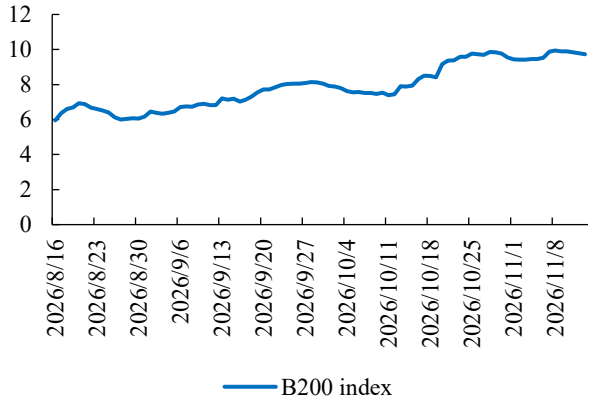
数据来源：ZenMux、开源证券研究所

5.2、 算力租赁价格更新

GPU 租赁价格整体高位运行，但本周有所回落。根据 Ornn 统计的基于 neocloud 市场的租赁交易价格，GPU 租赁价格整体保持高位运行，不过不同性能 GPU 价格走势有所分化，高性能型号 B200 与 H200 租赁价格处于高位但近期有所回落，H100 与 A100 租赁价格震荡运行。

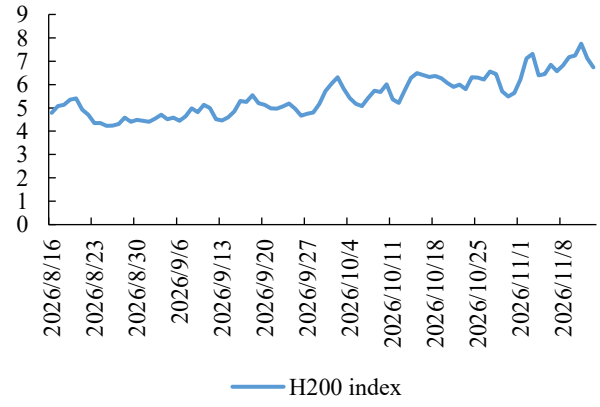
截至 11 月 20 日，B200 租赁价格为 9.10 $/小时；H200 租赁价格为 6.14 $/小时；H100 租赁价格为 2.9 $/小时；A100 租赁价格为 1.00 $/小时。

图13：B200 租赁价格高位运行但有所回落（单位：$/小时）



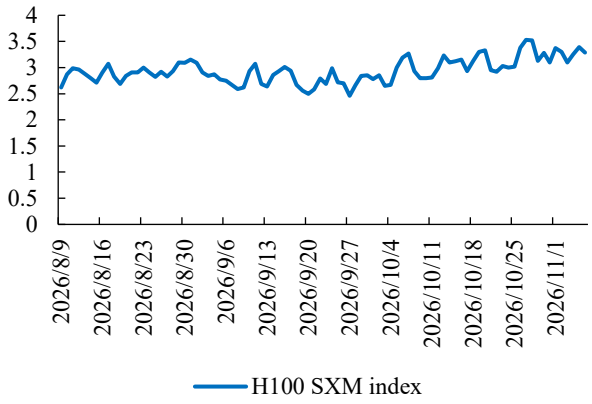
数据来源：Ornn、开源证券研究所

图14：H200 租赁价格高位运行但有所回落（单位：$/小时）



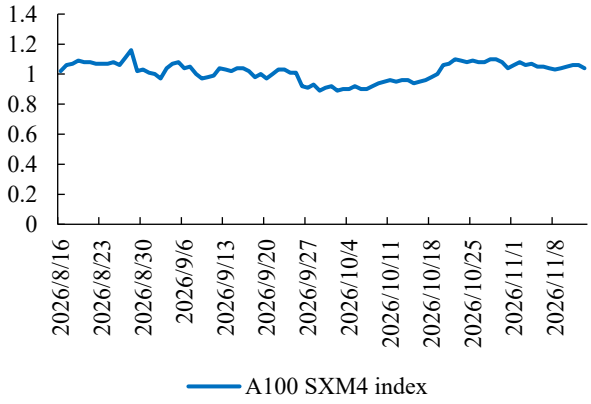
数据来源：Ornn、开源证券研究所

图15：H100 租赁价格高位震荡（单位：$/小时）



数据来源：Ornn、开源证券研究所

图16：A100 租赁价格基本持平（单位：$/小时）



数据来源：Ornn、开源证券研究所

6、 投资建议

大模型：Haiku 5.5 与 GPT-6.1 Sol 提升知识工作、编程及 Agent 能力并降低实际调用成本，改善复杂工作流的应用经济性，有望带动 Token 调用和推理算力需求增长。大模型端，国产模型流量增长与商业化推进相互促进，建议关注具备模型迭代、产品落地和客户拓展能力的厂商，推荐标的：智谱、德适-B，受益标的 MiniMax。

云计算：推理需求和企业 AI 支撑算力建设，但高额资本开支的回报仍需跟踪，收入兑现与利润率压力可能带来短期估值波动。受益标的：谷歌、亚马逊、微软、SpaceX、Meta、阿里巴巴。

7、 风险提示

模型及产品推出不及预期：大模型行业更新较快，市场竞争激烈，若产品推出不及时或不及预期，或影响企业未来竞争力和创收情况。

AI 算力需求不及预期风险：若下游模型迭代放缓、客户融资或算力利用率低于预期，可能导致新增订单放缓、RPO 转化周期拉长或数据中心产能交付延后。

算力供给超预期释放风险：本轮按需算力价格上涨建立在供需缺口之上，若 GPU 与内存供给快速释放，或新云厂商产能集中投放，按需与长约价格可能回落，算力租赁的盈利弹性随之收窄。

宏观经济增长放缓：若宏观经济增长放缓影响终端消费需求、企业 IT 支出等，将影响产业链相关企业的创收能力。

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已于2017年7月1日起正式实施。根据上述规定，开源证券评定此研报的风险等级为R4（中高风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为境内专业投资者及风险承受能力为C4、C5的普通投资者。若您并非境内专业投资者及风险承受能力为C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。
因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师承诺

本研究报告的署名人员具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告，并对内容和观点负责。本报告清晰准确地反映了署名人员的研究观点，所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。本报告署名人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现 20%以上；
	增持（outperform）	预计相对强于市场表现 5%～20%；
	中性（Neutral）	预计相对市场表现在－5%～＋5%之间波动；
	减持（underperform）	预计相对弱于市场表现 5%以下。
行业评级	看好（overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（underperform）	预计行业弱于整体市场表现。
备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数、港股基准指数为恒生指数、新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）、美股基准指数为标普 500 或纳斯达克综合指数。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。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。		

分析、估值方法的局限性说明

本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用。

本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动，过往的业绩表现不应作为其日后表现的预示。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。投资者应自主作出投资决策并自行承担投资风险，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。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

本报告的版权归本公司所有。本公司对本报告保留一切权利。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

开源证券研究所

上海
地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层
邮编：200120
邮箱：research@kysec.cn

深圳
地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层
邮编：518000
邮箱：research@kysec.cn

北京
地址：北京市西城区西直门外大街18号金贸大厦C2座9层
邮编：100044
邮箱：research@kysec.cn

西安
地址：西安市高新区锦业路1号都市之门B座5层
邮编：710065
邮箱：research@kysec.cn