

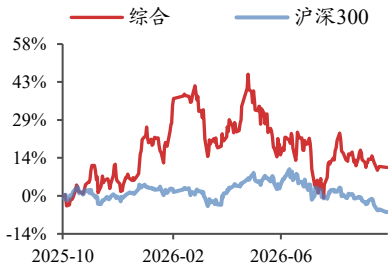
综合

2026 年 10 月 11 日

投资评级：看好（维持）

——行业周报

行业走势图



数据来源：聚源

相关研究报告

《海外科技：AI Agent 打开 CPU 空间，海外 X86 服务器供应链受益——行业周报》-2026.9.27

《港股消费周报：微软升级 Copilot 加速 AI 渗透，上美合作丽驻兰——行业周报》-2026.9.27

《市场更加重视模型帕累托最优，模型厂&云厂&主权基金资金弹药充足——港股行业深度报告》-2026.9.24

海外科技：AMD 布局世界模型，模型能力升级与成本优化同步推进

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

● AI 芯片：AMD 布局物理 AI，Marvell 指引上调，光互连带价齐升

AMD 收购 World Labs，强化软硬协同布局物理 AI：9 月 28 日，AMD 宣布拟以 82 亿美元全股票收购李飞飞创立的 World Labs，布局空间智能与世界模型。后者已与 AMD 开展 GPU 训练及推理优化合作。我们认为，此次收购有望强化 Instinct 与 ROCm 软硬件协同，拓展物理 AI 市场，提升 AMD 中长期竞争力。

Marvell 大幅上调指引，光互连带价齐升：Marvell 上调 FY2028 收入目标至 200 亿美元，FY2031 目标中值 800 亿美元，互联占比 47%。预计 2030 年互联市场达 650 亿美元，复合增速 65%。我们认为，速率升级及 Scale-Up 光化将推动光通信量价齐升，利好硅光、激光器及高速模拟电芯片。

● 服务器：HPE 斩获 Helios AI 机架大单，半导体设备持续增强定价

服务器：HPE 获 Vultr 12 亿美元 AMD Helios AI 机架订单，并推出搭载第六代 EPYC 的 ProLiant Gen13 服务器，验证 AI 整机需求。

半导体设备：ASML 对韩光刻机零部件提价 10%，设备厂商议价能力增强；三星预计 2027 年 HBM 占 DRAM 产能近 30%，应用材料与 Besi 加码先进封装，AI 扩产持续拉动设备、存储及封装需求。

● 大模型：能力升级与成本优化同步推进，复杂 Agent 任务成本进一步下降

Anthropic 发布 Haiku 5.5，在 10 万 Token 以内的输入场景下，输入、输出价格较前代均下降 90%，知识工作与编程能力同步改善；Sonnet 5.5 缓存读取价格下调 50%，多数 Agent 任务成本降低约 20%。OpenAI 推出 GPT-6.1 Sol，复杂任务表现接近 GPT-6 Astra，但标准输入、输出价格仅为后者的五分之一，缓存输入价格较 GPT-6 Sol 下降 50%。我们认为，本轮模型更新推动能力与成本同步优化：中档模型以更低成本逼近旗舰模型，轻量模型拓展高频、成本敏感型场景，模型竞争进一步转向任务执行效果与综合成本。

● 英伟达洽谈收购 Reflection AI，拓展 AI 软硬件协同布局

英伟达正与 Reflection AI 就潜在收购初步洽谈，尚未达成协议。若收购推进，英伟达有望将 Reflection AI 的开源权重模型纳入自身 AI 生态：以 GPU 算力支持模型迭代、推动开放权重模型能力前沿，并结合其政府项目资源，为政企及主权客户提供可私有化部署的“模型+算力”方案，强化英伟达对政府 AI 预算及软硬件生态的绑定。

● 投资建议：

AI 硬件：推荐标的：Tower 半导体，受益标的：Lumentum、AMD；**服务器：**受益标的：联想、Dell、HPE；**大模型与云计算：**推荐标的：智谱、德适-B，受益标的：谷歌、亚马逊、微软、Nebius、CoreWeave、MiniMax。

● **风险提示：**AI 资本开支增速放缓风险、CPU 竞争格局变化风险、CPO 产业化不及预期风险、模型及产品推出不及预期，AI 算力需求不及预期风险。

目 录

1、 AI 硬件：AMD 加码世界模型，Marvell 看好光互联增长.....	4
1.1、 AMD 斥资 82 亿美元收购 World Labs，世界模型补齐 AI 生态版图.....	4
1.2、 Marvell 大幅上调中长期收入目标，光互联有望成为 AI 基础设施下一轮增长核心	4
2、 AI 设备：HPE 获 HeliosAI 机架公开大单，品牌 OEM 份额加速扩张.....	5
2.1、 服务器：HPE 获 AMDHeliosAI 机架首个公开大单，持续夯实企业端服务器布局.....	5
2.2、 半导体设备：ASML 供应韩国提价 10%，应用材料与 Besi 扩键合设备合作关系	6
2.2.1、 ASML 供应韩国提价 10%，设备厂商议价能力显著增强	6
2.2.2、 三星预计 2027 年 HBM4 出货比例提升至 80%，应用材料与 Besi 扩大研发合作关系	7
3、 大模型：能力升级与成本优化同步推进，Agent 任务执行成本进一步下降	7
3.1、 Anthropic：Haiku 5.5 以更低成本提升知识工作与 Agent 能力，价格优化覆盖轻量模型及缓存调用	7
3.2、 OpenAI：GPT-6.1 Sol 以五分之一价格提供接近 Astra 的模型能力，复杂 Agent 任务成本进一步下降.....	9
4、 英伟达洽谈收购 Reflection AI，拓展 AI 软硬件协同布局	10
4.1、 Beam：501B 总参数、单次激活 23B，主打每 token 智能的推理效率	11
4.2、 Reflection AI 已深度扎根政府项目，英伟达有望借助开源模型绑定政企预算	12
5、 港股周度更新	13
6、 模型与算力价格数据更新	15
6.1、 模型信息更新	15
6.2、 算力租赁价格更新	17
7、 投资建议	18
8、 风险提示	18

图表目录

图 1： 互联市场预计在 2025-2030 年复合增速达到 65%.....	5
图 2： Claude Haiku 5.5 较 Haiku 4.5 的 Token 价格明显下降	8
图 3： Claude Haiku 5.5 在知识工作与计算机操作评测中较前代显著提升	8
图 4： DeepSWE v1.1 中，GPT-6.1 Sol 以约五分之一成本达到与 GPT-6 Astra 相当的水平	9
图 5： Beam 的编程与 Agent 能力与 GLM 5.2 相近，提升了美国开放权重模型的能力前沿	11
图 6： Beam 将编程与 Agent 能力和高效推理相结合.....	12
图 7： Reflection AI 已深度扎根政府项目	12
图 8： 本周恒生指数（2026.10.05-2026.10.09）在全球主要市场中涨跌幅靠前（单位：%）	13
图 9： 本周（2026.10.05-2026.10.09）恒生港股通高股息率涨幅较大（%）	13
图 10： 本周（2026.10.05-2026.10.09）恒生航运主题、消费涨幅居前（%）	13
图 11： 2026 年 9 月以来港股通当日买入成交净额整体为正（单位：亿元）	14
图 12： 恒生沪港通 AH 溢价指数或已触底.....	15
图 13： OpenRouter 的 token 调用量持续提升	16
图 14： 国产 AI 在 ZenMux 的 Token 市场份额占比接近 50%	17
图： B200 租赁价格高位运行（单位：\$/小时）	17
图： H200 租赁价格高位运行（单位：\$/小时）	17
图： H100 租赁价格高位震荡（单位：\$/小时）	18
图： A100 租赁价格有所下降（单位：\$/小时）	18

表 1： HPE 推出四款 ProLiantGen13 平台服务器	6
---	---

表 2: 本周 (2026.10.05-2026.10.09) 友邦保险、腾讯控股与阿里巴巴-W 流入资金居前	14
表 3: 9 月以来新上线模型与待发布模型	15

1、AI 硬件：AMD 加码世界模型，Marvell 看好光互联增长

1.1、AMD 斥资 82 亿美元收购 World Labs，世界模型补齐 AI 生态版图

9 月 28 日，AMD 宣布以约 82 亿美元全股票交易收购由李飞飞创立的 World Labs，预计 2026 年底前完成交割。收购完成后，李飞飞将出任 AMD 执行副总裁兼首席科学家，直接向苏姿丰汇报。

World Labs 聚焦空间智能与世界模型，推动 AI 从理解语言走向理解物理世界。 World Labs 成立于 2024 年，致力于开发能够理解三维空间、物体运动和环境变化的世界模型。不同于主要处理语言及符号关系的大语言模型（LLM），世界模型旨在让 AI 理解、生成并模拟真实或虚拟三维环境。World Labs 于 2025 年推出的 Marble 可根据文本、图像及视频生成可交互的三维场景；2026 年 9 月发布的新一代 Atlas 则采用多模态自回归扩散 Transformer 架构，原生融合文本、图像、视频及三维空间信息，实现空间重建、可控视频生成和机器人仿真。

我们认为，此次收购的核心是通过引入世界模型研发团队，推动 AMD 从 GPU、CPU 算力供应商，进一步向理解前沿模型需求、实现软硬件协同设计的 AI 基础设施平台演进。对于 AMD 而言，此次收购最重要的意义在于获得前沿模型研发能力，使其能够提前识别下一代 AI 工作负载对计算、内存、软件框架及系统架构的需求。

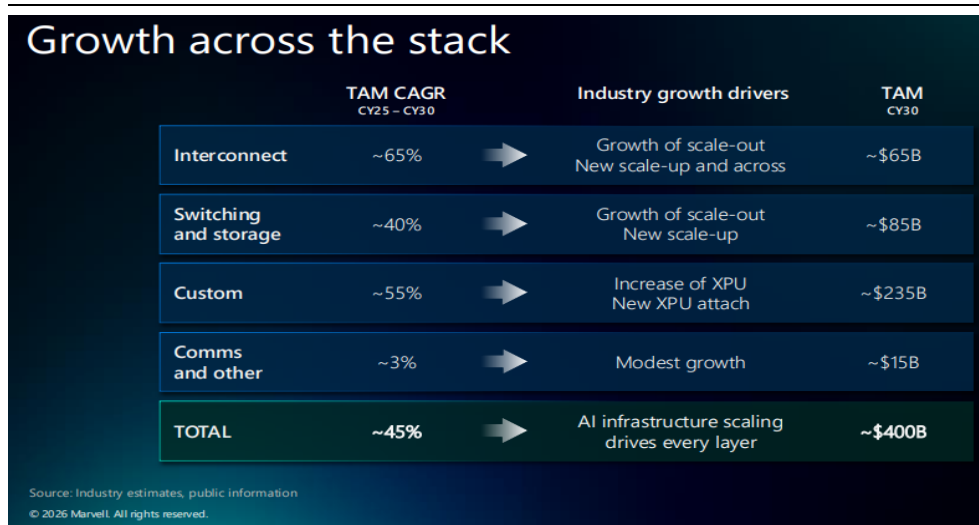
受益标的：收购世界模型的 AMD，以及未来或存在自研世界模型的英伟达。

1.2、Marvell 大幅上调中长期收入目标，光互联有望成为 AI 基础设施下一轮增长核心

10 月 6 日，Marvell 在 2026 年投资者日宣布，将 FY2028（对应 2027 自然年为主）收入目标从 180 亿美元上调至 200 亿美元，同比增长约 67%，其中数据中心业务预计贡献约 180 亿美元，同比增长约 80%；同时首次提出 FY2031 收入达到 700 亿—900 亿美元的长期目标，以中值 800 亿美元计算，相较 FY2026 的 82 亿美元增长近 10 倍，对应未来五年约 58% 的收入复合增速。管理层指出，此次 FY2028 收入目标新增的 20 亿美元，主要来自 Scale-Out 光互联、Scale-Up 光互联及交换芯片业务需求超预期。

Marvell 预计，2030 年公司可服务市场（TAM）将达到约 4000 亿美元，其中互联市场规模约 650 亿美元，对应 2025—2030 年约 65% 的复合增速，明显快于整体 AI 基础设施资本开支增长。根据公司长期规划，FY2031 互联业务收入预计达到 375 亿美元，占 800 亿美元整体收入目标中值的约 47%；相比之下，定制芯片业务、交换及存储业务收入预计分别达到 300 亿、100 亿美元。

图1：互联市场预计在 2025-2030 年复合增速达到 65%



资料来源：Marvell 2026 投资者日 PPT

Marvell 指出，现有 Scale-Out 光互联业务过去每年保持接近翻倍增长，虽然长期增速预计逐渐趋缓，但 1.6T 正进入规模部署阶段，3.2T 升级周期也已开启。公司认为，NPO 和 CPO 将成为未来 Scale-Up 光互联的重要实现路径，并通过此前收购的 Celestial AI，将 Photonic Fabric 与自有硅光、TIA、Driver 及交换芯片技术整合，提供完整的光引擎解决方案，Scale-Up 光互联对应每颗 XPU 的光学连接价值量可达到数千美元量级，而且链路两端均需要部署光学器件。

我们认为，Marvell 将互联业务放在接近半数的长期收入权重，说明 AI 集群扩张不仅增加 GPU 及 ASIC 数量，更将带动每颗计算芯片所对应的网络连接价值量持续提升，光互联正在从 AI 基础设施的配套环节，逐渐成为决定整体系统效率的重要组成部分。相比传统 Scale-Out 主要受益于网络端口速率提升，Scale-Up 引入光互联有望进一步扩大光学连接的应用边界，为硅光芯片、激光器、光引擎及高速模拟电芯片创造增量需求。Marvell 的长期指引再次验证光互联产业的量价齐升逻辑，重点关注硅光、激光器及高速模拟前端。

投资建议：推荐标的：硅光晶圆、电芯片代工环节的 Tower 半导体。受益标的：激光器环节的 Lumentum、Coherent。

2、AI 设备：HPE 获 HeliosAI 机架公开大单，品牌 OEM 份额

加速扩张

2.1、服务器：HPE 获 AMDHeliosAI 机架首个公开大单，持续夯实企业端服务器布局

HPE 获得 AMDHelios 平台首个公开大单，上调 FY2027 网络业务收入增速指引。9 月 30 日 HPE 宣布获得 Vultr 价值 12 亿美元的 AMDHeliosAI 机架订单该系统用于 Vultr 位于美国的云数据中心，每个机架集成了 72 个 AMDInstinct™MI455XGPU、AMDEPYC™“Venice”CPU、AMDPensando™VulcanoAI 网卡以及 AMDROCm™软件，同时配备包括每个机架六个 HPEJuniperNetworkingQFX5252 可扩展以太网交换机托

架，通过高带宽、低延迟、基于标准的以太网连接所有 72 个 GPU，验证 HPE 把瞻博网络整合进 AI 数据中心网络的能力。公司同步将上调 FY2027 网络业务收入增速指引由 14-17%上调至“高十几-低二十”，并表示到 2029 财年数据中心网络业务有望以 50%以上的复合增速持续增长。

HPE 推出新一代 X86 服务器，助力企业 AI 创新。10 月 HPE 发布四台全新的 HPEProLiantGen13 服务器，搭载第六代 AMD EPYC 服务器处理器，在更高核心密度、内存带宽与 I/O 容量，助力企业加速智能体人工智能和推理、分析、虚拟化、仿真和建模的基础上，同时通过自动自加密驱动器（Auto-SED）、多人授权（MPA）、后量子密码（PQC）固件级防护、硅信任根组件验证，进一步增强安全性、简化运维，并在企业分布式 IT 环境中提供可靠性。

表1: HPE 推出四款 ProLiantGen13 平台服务器

产品型号	预计上市时间	形态规格	处理器配置	扩展与 I/O	典型适用场景
DL585aGen13	2027 年初	高密度机架式	最多 2 颗第六代 AMD EPYC 服务器处理器	最多支持 8 个双宽 PCIe GPU, PCIe Gen6 连接; 最大化 GPU 利用率、令牌吞吐量与 PCIe 带宽	AI 推理、虚拟化、检索增强生成 (RAG) 模型微调、分析、智能体 AI
DL525Gen13	2026 年第四季度	1U 单路紧凑机架式	1 颗最高 256 核第六代 AMD EPYC 服务器处理器	高带宽 16 通道内存, 支持 DDR5 和 MRDIMM; PCIe Gen6 连接	欺诈检测、市场数据处理、风险和定价模型、电子设计自动化 (EDA)、仿真、AI 推理
XD245	2027 年	10U (单机架单元) 四节点	最多 8 颗第六代 AMD EPYC 服务器处理器	符合 OCP 标准, 共享电源与散热; 高密度计算+高 I/O 容量	数据密集型 AI、高性能计算 (HPC)、大规模仿真、复杂计算模型
XD285	2027 年	20U (双机架单元) 双节点	第六代 AMD EPYC 服务器处理器	符合 OCP 标准, 共享电源与散热架构	计算机辅助工程 (CAE)、天气和气候建模、科学研究、金融建模

资料来源: HPE 官网、开源证券研究所

2.2、半导体设备: ASML 供应韩国提价 10%，应用材料与 Besi 扩键合设备合作关系

2.2.1、ASML 供应韩国提价 10%，设备厂商议价能力显著增强

ASML 上调设备零件价格，设备厂商寻求强定价权。ASML 总部通过韩国子公司向三星电子、SK 海力士下发涨价方案，对供应韩国市场的 EUV、DUV 光刻机全部零部件统一上调 10% 价格，2027 年 1 月起交付的货品执行新价格。此举一方面因零部件上游原材料成本普涨，另一方面 ASML 希望提升自身利润空间，实际利润提升幅度仍将取决于成本变化、采购量和产品组合。应用材料、东京电子、AdvancedEnergy 等均在推进或宣布产品调价。多轮涨价根源在于供需失衡与通胀推高成本，设备供应商议价能力显著增强，反映在 AI 相关产能扩张背景下，半导体设

备供应商持续寻求更大的定价权。

2.2.2、三星预计 2027 年 HBM4 出货比例提升至 80%，应用材料与 Besi 扩大研发合作关系

HBM4 扩产持续推进，良率和产能仍为供给缺口制约。三星管理层表示预计 2027 年 HBM 将占 DRAM 行业总晶圆产能的近 30%，高于目前的约 20%，其中三星的 HBM 晶圆月均产量可能从 2026 年的约 18 万片增长近 40%，到 2027 年达到约 25 万片，随着 HBM4E 量产的推进，HBM4 系列可能占三星 HBM 2027 年总出货量的 80% 左右，高于 2026 年的约 40%。TrendForce 上调其 2027 年 HBM 价格预期，以反映持续的供应限制、高价 HBM4 份额的不断增长以及 HBM4e 在 2027 年下半年的逐步提升，目前预计综合平均售价将同比增长 121%。

应用材料公司与 BESI 扩大战略合作伙伴关系，加码先进封装系统布局。Besi 以创新合作伙伴的身份加入 Applied 的 EPIC 中心，进一步拓展了双方自 2020 年两家公司联合成立以来的合作关系，预计的研发项目将包括：（1）开发用于 D2W 混合键合的新平台；（2）热压焊接(TCB)及其周边工艺生态系统的规模化；（3）新 D2W、D2D 和 D2P 集成平台；（4）支持共封装光学器件和人工智能系统扩展的光子互连应用。

3、大模型：能力升级与成本优化同步推进，复杂 Agent 任务执行成本进一步下降

3.1、Anthropic: Haiku 5.5 以更低成本提升知识工作与 Agent 能力，价格优化覆盖轻量模型及缓存调用

Haiku 5.5 在知识工作、计算机操作及编程任务中实现明显提升，轻量模型的适用范围进一步扩大。10 月 7 日，Anthropic 发布 Claude Haiku 5.5，定位高频、成本敏感型任务。相比 Haiku 4.5，新模型在多项评测中实现显著提升：GDPval-AA v2.1 得分由 735 提高至 1620；OSWorld 2.1 离线子集得分由 15.7%提升至 72.4%，Terminal-Bench 4.0 由 0.0%提升至 39.2%。

价格下降覆盖输入、输出及缓存调用，多数请求场景下 Token 单价大幅下降。对于不超过 10 万 Token 的输入，Haiku 5.5 每百万 Token 输入、输出价格分别为 0.10 美元和 0.50 美元，较 Haiku 4.5 的 1 美元和 5 美元均下降 90%；缓存读取及写入价格分别由 0.10 美元和 1.25 美元降至 0.01 美元和 0.125 美元。对于超过 10 万 Token 的输入，模型输入、输出价格分别为 0.50 美元和 2.50 美元，较前代下降 50%。Anthropic 披露，Haiku 4.5 约 90% 的请求属于不超过 10 万 Token 的场景，考虑新模型分词器变化带来的任务 Token 消耗差异后，Haiku 5.5 平均运行成本较前代降低约 75%。

图2: Claude Haiku 5.5 较 Haiku 4.5 的 Token 价格明显下降

Price per 1 million tokens	Haiku 5.5 prompts up to / over 100k	Haiku 4.5	Sonnet 5.5
Cache reads	\$0.01 / \$0.05	\$0.10	\$0.10
Cache writes	\$0.125 / \$0.625	\$1.25	\$2.50
Input tokens	\$0.10 / \$0.50	\$1.00	\$2.00
Output tokens	\$0.50 / \$2.50	\$5.00	\$10.00

资料来源: Anthropic 官网

Haiku 5.5 在知识工作和计算机操作评测中超过 GPT-6 Luna, 但仍低于 Sonnet 5.5。在 GDPval-AA v2.1 中, Haiku 5.5 得分为 1620, 高于 GPT-6 Luna 的 1437, 低于 Sonnet 5.5 的 1840; 在 OSWorld 2.1 离线子集中, Haiku 5.5 得分为 72.4%, 高于 GPT-6 Luna 的 48.9%, 但低于 Sonnet 5.5 的 83.9%。编程方面, Haiku 5.5 在 Terminal-Bench 4.0 中得分 39.2%, 高于 GPT-6 Luna 的 16.4%, 但仍明显低于 Sonnet 5.5 的 70.6%。Anthropic 指出, Sonnet 5.5 和 Opus 5.5 仍更适合复杂 Agent 编程任务, Haiku 5.5 则主要面向上下文压缩、摘要及子 Agent 等范围相对明确的工作。

图3: Claude Haiku 5.5 在知识工作与计算机操作评测中较前代显著提升

	Haiku 5.5	Haiku 4.5	GPT-6 Luna	Sonnet 5.5 For reference
Knowledge work GDPval-AA v2.1	1620	735	1437	1840
Knowledge work AA-Briefcase v1.1	1578	614	1336	1824
Computer use OSWorld 2.1	72.4% Offline subset	15.7% Offline subset	48.9% Offline subset	83.9% Offline subset
Multidisciplinary reasoning Humanity's Last Exam	45.9% no tools	10.2% no tools	—	56.9% no tools
	57.4% with tools	18.7% with tools	—	64.5% with tools
Agentic coding Terminal-Bench 4.0	39.2%	0.0%	16.4%	70.6%
Agentic coding FrontierCode 1.1 (Main)	46.4%	—	42.4%	52.1% Xhigh
Visual reasoning Chartography	46.4% no tools	6.4% no tools	29.1% no tools	61.6% no tools

资料来源: Anthropic 官网

早期客户测试显示模型推理速度提升、任务延迟明显下降。Asana 在其 AI Teammates 测评集中测试了 Haiku 5.5, 相比其当前使用的模型, 任务完成延迟下降超过 30%, Agent 推理速度最高提升至 2.5 倍。Haiku 5.5 已通过 Claude Platform、AWS、Google Cloud 和 Microsoft Azure 提供服务。

Sonnet 5.5 同步降低缓存读取价格, Agent 工作流的成本优化进一步扩展至中档模型。Anthropic 将 Sonnet 5.5 每百万 Token 缓存读取价格由 0.20 美元下调至 0.10 美

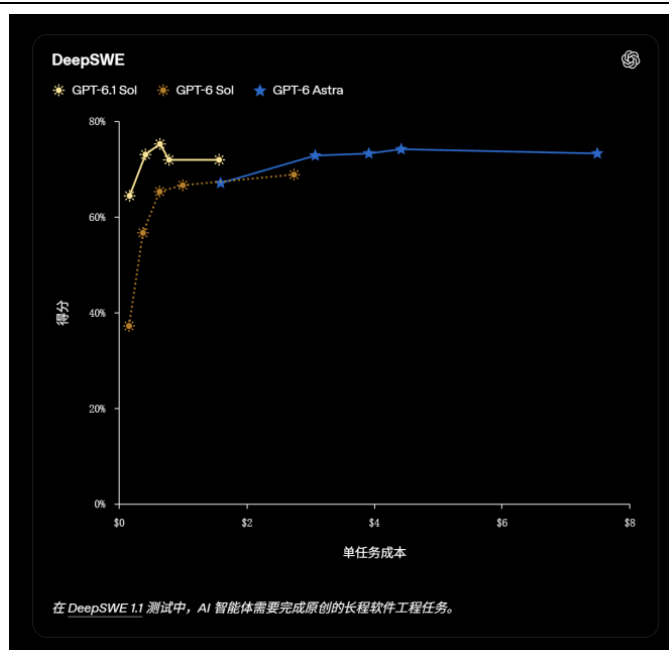
元，降幅达到 50%。由于缓存读取在模型 Token 消耗中占比较高，官方表示此次调整可使多数 Agent 工作负载的运行成本下降约 20%。此外，Anthropic 宣布自发布当周起向 Max 和 Team 订阅用户推出月度 API 额度。

本轮更新体现了轻量模型能力提升与使用成本下降的同步推进，模型分工仍与任务复杂度密切相关。Haiku 5.5 在知识工作、计算机操作和部分编程测评中的提升，使其能够以较低价格承担更多高频、范围明确的任务；Sonnet 5.5 的缓存降价则进一步降低复杂 Agent 工作流的运行成本。与此同时，Haiku 5.5 在复杂编程任务中仍明显落后于 Sonnet 5.5。

3.2、OpenAI: GPT-6.1 Sol 以五分之一价格提供接近 Astra 的模型能力，复杂 Agent 任务成本进一步下降

GPT-6.1 Sol 在智能体编程、专业工作与计算机操作中实现能力提升，以更低成本接近旗舰模型表现。9 月 29 日，OpenAI 发布 GPT-6.1 Sol，重点提升代码编写与调试、复杂文档理解及多步骤业务工作流程执行能力。在 DeepSWE v1.1 中，GPT-6.1 Sol 以约五分之一的单任务成本达到与 GPT-6 Astra 相当的水平，同时以更低的推理强度和成本，较 GPT-6 Sol 的最高得分提高 6.4 个百分点。

图4: DeepSWE v1.1 中，GPT-6.1 Sol 以约五分之一成本达到与 GPT-6 Astra 相当的水平



资料来源: OpenAI 官网

基础 Token 价格维持中档模型水平，缓存输入价格下降 50%，与旗舰模型形成明显成本差异。GPT-6.1 Sol 每百万 Token 输入、输出价格分别为 2 美元和 10 美元，仅为 GPT-6 Astra 对应价格的五分之一；缓存输入价格为 0.10 美元，较 GPT-6 Sol 下降 50%。对于需要在不同请求间复用上下文的 Agent 任务，更低的缓存输入价格有助于减少重复处理上下文的调用成本。

专业文档理解与多步骤业务工作流程能力提升，部分任务表现超过 Opus 5.5 且成本更低。在 GDP.pdf 测评中，GPT-6.1 Sol 需要根据金融、医疗、法律等 10 个专业领域的复杂 PDF 文档回答问题。官方披露，在所测试的各档推理强度下，GPT-6.1 Sol

得分均高于采用回退机制的 Opus 5.5，而单任务成本不到其一半，同时以约五分之一的任务成本接近 Astra 的表现。在 AutomationBench 1.0.6 评测中，GPT-6.1 Sol 在中等推理强度下得分较 Opus 5.5 高 2.2 个百分点，成本约为其三分之一，较相同推理设置下的 GPT-6 Sol 高 4.8 个百分点。

计算机操作与科学研究任务的得分与成本表现同步改善，但最高难度科学研究任务仍由 Astra 保持领先。在 OSWorld 2.0 离线测试集中，GPT-6.1 Sol 在最高推理强度下得分较 GPT-6 Sol 提高 7 个百分点，成本不到其一半；与 Astra 的得分差距缩小至 2.1 个百分点以内，单任务成本约为其七分之一。科学研究方面，在 Terminal-Bench Science 0.1 中，GPT-6.1 Sol 在最高推理强度下得分超过 GPT-6 Sol 的两倍，单任务成本不到其一半；其平均单任务成本为 5.47 美元，低于 Opus 5.5 的 23.21 美元和 Astra 的 23.80 美元，降幅均超过 75%。不过，Astra 仍以 68.1% 的得分在受测模型中排名第一。

高难度场景评测中，GPT-6.1 Sol 的事实错误率下降，搜索工具故障告知能力有所改善。在高难度事实准确性评测中，GPT-6.1 Sol 在低推理强度下包含事实错误的回复比例由 GPT-6 Sol 的 11.4% 降至 7.7%，降幅约 32%；在全部受测推理强度下，其错误率与 Astra 的差距均不超过 1.9 个百分点，单任务成本不到后者的五分之一。安全评估方面，在最高推理强度下，当搜索工具发生故障时，GPT-6.1 Sol 未向用户说明的比例为 2.1%，低于 GPT-6 Sol 的 4.9%，但仍高于 Astra 的 1.5%。OpenAI 指出，上述评估采用刻意筛选的高难度场景，并不代表典型使用情况下的错误率。

产品部署覆盖 ChatGPT 工作、Codex 及 API，进一步支持专业任务与智能体应用。GPT-6.1 Sol 已面向 Plus、Pro、Business、Enterprise 和 Edu 用户开放，可在 ChatGPT 工作及 Codex 中使用。开发者可通过 OpenAI API 调用模型。此外，OpenAI 随后推出 GPT-6.1 Sol Ultrafast，在 Codex 中的 Token 生成速度最高可达到标准版本的 8 倍。

本轮升级体现了中档模型向旗舰能力靠近的同时，复杂任务的实际执行成本进一步下降。GPT-6.1 Sol 在部分专业任务中以更低成本达到接近 Astra 的表现。与此同时，Astra 在最高难度科学研究任务中仍保持领先，说明不同能力档位的模型仍存在适用场景差异。

4、英伟达洽谈收购 Reflection AI，拓展 AI 软硬件协同布局

据《金融时报》10 月 10 日报道，英伟达正与美国 AI 初创公司 Reflection AI 就进一步投资或潜在收购进行初步洽谈，可能的交易形式包括股权投资、算力合作、收购及人才与技术整合等；目前双方尚未达成协议，交易仍存在不确定性。英伟达此前已对 Reflection AI 进行大额投资，已向该初创公司投资了 8 亿美元。

Reflection AI 成立于 2024 年，由前 Google DeepMind 研究人员创立，是一家专注于开放权重基础模型研发的美国人工智能初创公司。公司以构建高性能开放模型为核心，重点布局代码生成、复杂推理及 Agent 等应用场景，可为企业和政府客户提供可自主部署、可结合专有数据定制的 AI 系统。

我们认为，英伟达收购 Reflection AI 有两点重要意义：（1）为 Reflection 开源模型提供算力保障，推动其模型迭代，同时推进美国开放权重模型的能力前沿。（2）借助开源可私有化部署特性与 Reflection AI 的政府资源，让英伟达跳过公有云中介，

直接在物理设施层将政企预算锁定在自己的生态内。

4.1、Beam: 501B 总参数、单次激活 23B，主打每 token 智能的推理效率

2026 年 10 月 5 日，Reflection AI 发布了首个开源权重模型 Beam。该模型采用稀疏混合专家（MoE）架构，总参数量达 501B，单次推理激活参数为 23B。旨在与成本较低的中国模型 DeepSeek、Kimi 等竞争编码和代理任务。

Beam 的能力来自预训练与强化学习（RL）两方面的大规模投入。在预训练阶段，使用了来自互联网及专有授权数据集的 23.8 万亿个多样化、经过筛选的高质量 Token 进行训练。与此同时，开发了相应的算法、训练环境及基础设施，以支持大规模、高算力消耗的强化学习训练。在为期 4 周的高算力强化学习训练中使用了 10500 张 NVIDIA GB300 GPU，累计生成超过 1 亿次 Rollout。

Beam 在训练时注重其编程能力与 Agent 任务表现，推进了美国开放权重模型的能力前沿。在编程和 Agent 任务上的表现可与参数规模更大的开放模型（如 GLM 5.2）相竞争，并逐步接近 Qwen 3.8-Max。尽管 Kimi K3 等前沿开放模型在整体原始能力上仍具优势，但 Beam 的优势在于推理阶段的效率。

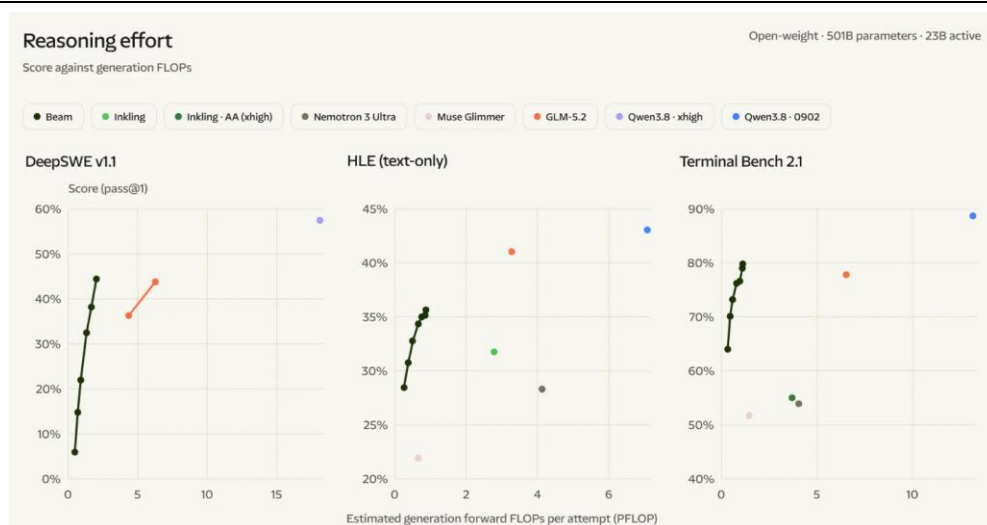
图5: Beam 的编程与 Agent 能力与 GLM 5.2 相近, 提升了美国开放权重模型的能力前沿

Agentic Coding/Terminal								
	Beam	Inkling	Nemotron 3 Ultra	GLM 5.2	GLM 5.3	Kimi K3	Qwen 3.8 Max	DeepSeek V4.1 Flash Max
TOTAL PARAMETERS	501B	975B	550B	753B	753B	2.8T	2.4T	552B
ACTIVE PARAMETERS	23B	41B	55B	40B	40B	104B	95B	88/16B
CONTEXT LENGTH	1M	1M	1M	1M	1M	1M	1M	1M
LICENSE	Apache 2.0	Apache 2.0*	Custom	MIT	Custom	Custom	Custom	MIT
DeepSWE v1.1	44.4	-	-	44.0	61.0	68.0	51.0	74.2
SWE-Bench Pro v2-Hard	77.2	56.9	-	-	84.3	88.2	-	-
SWE-Bench Pro v1	65.5	54.3	46.4	62.1	-	-	67.7	-
Terminal Bench v2.1	80.1	63.8	56.4	81.0	88.2	88.3	86.6	90.6
SWE Atlas Codebase QnA	34.6	-	-	-	61.0	68.0	-	-
SWE-Bench Multilingual	78.0	-	67.7	-	-	-	-	-
SWE-Bench Verified	80.9	77.6	70.7	-	-	-	-	-

资料来源：Reflection AI 官网

Beam 将编程与 Agent 能力和高效推理相结合。在高级推理基准测试中，Beam 的得分可与 Qwen 3.8-Max 媲美，同时推理计算量降低约 3-4 倍。与 Qwen 3.8-Max 等参数规模超过 2 万亿的模型相比，Beam 的效率优势更加显著，后者每生成一个 Token 所需的推理计算量明显更高。这意味着 Beam 能够以更低的推理成本提供较强的模型能力，实现更高的单位 Token 智能产出，有望使其成为适用于企业级编程与智能体工作负载的高性价比主力模型。

图6: Beam 将编程与 Agent 能力和高效推理相结合



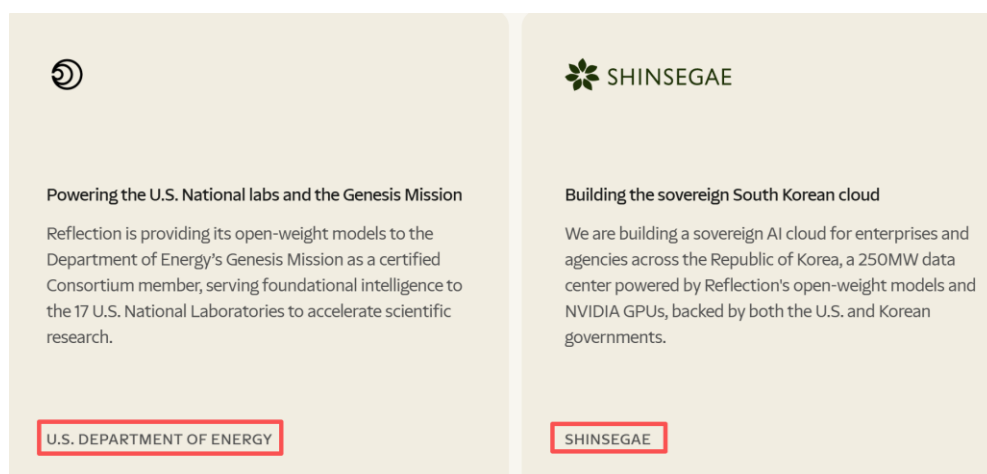
资料来源: Reflection AI 官网

4.2、Reflection AI 已深度扎根政府项目，英伟达有望借助开源模型绑定政企预算

Reflection AI 已经深度扎根政府与主权 AI 云项目。业务包括为韩国推进 250MW 的主权 AI 云项目，由 Reflection AI 的开放权重模型和 NVIDIA GPU 驱动，已得到美韩两国政府的支持。并为美国能源部 Genesis 任务以及 17 所国家实验室供应定制模型。

政府客户对公有云闭源 API 较为忌惮，不会把敏感数据上传给云厂商与模型厂商，私有化部署和完全可控的权重是其采购的底线。如果英伟达掌控了顶级的开源权重，就能向全球主权客户提供一体化方案：购买英伟达的机柜，跑英伟达软硬件协同的 Reflection 模型。面对微软、亚马逊和谷歌加速推进自研芯片的现状，这种打法让英伟达跳过公有云中介，直接在物理设施层将政企预算锁定在自己的生态内。

图7: Reflection AI 已深度扎根政府项目

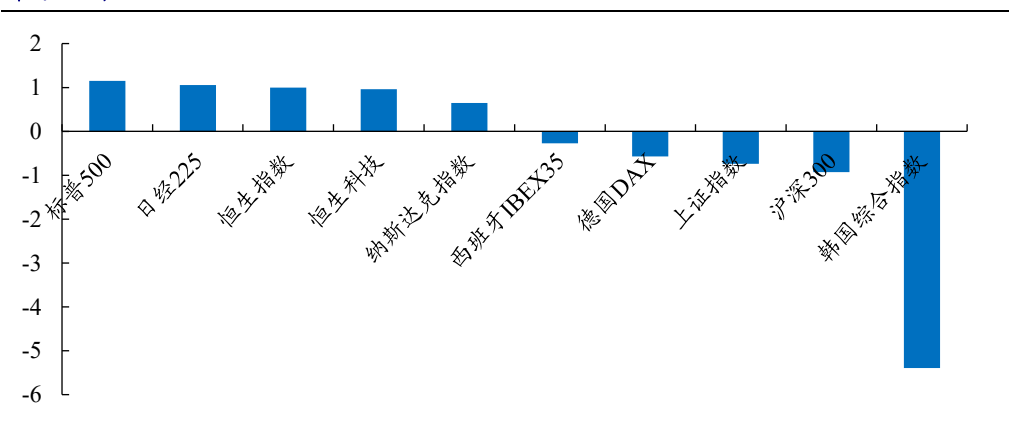


资料来源: Reflection AI 官网、开源证券研究所

5、港股周度更新

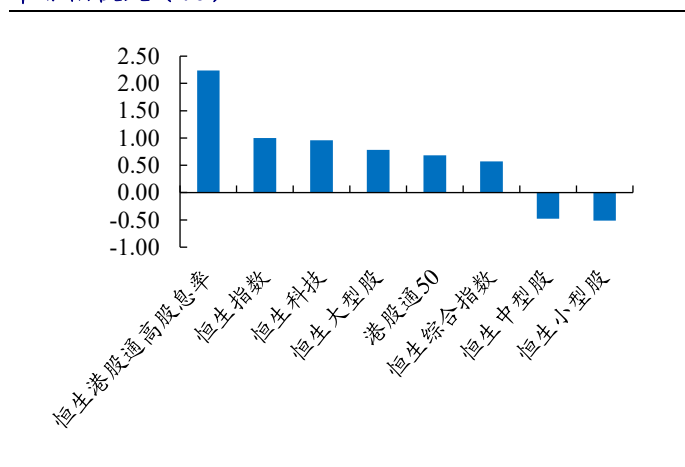
本周恒生指数上涨 1.00%，在全球主要市场中涨跌幅靠前，从成分上看，恒生港股通高股息率等涨幅居前，恒生科技本周上涨 0.96%。

图8：本周恒生指数（2026.10.05-2026.10.09）在全球主要市场中涨跌幅靠前（单位：%）



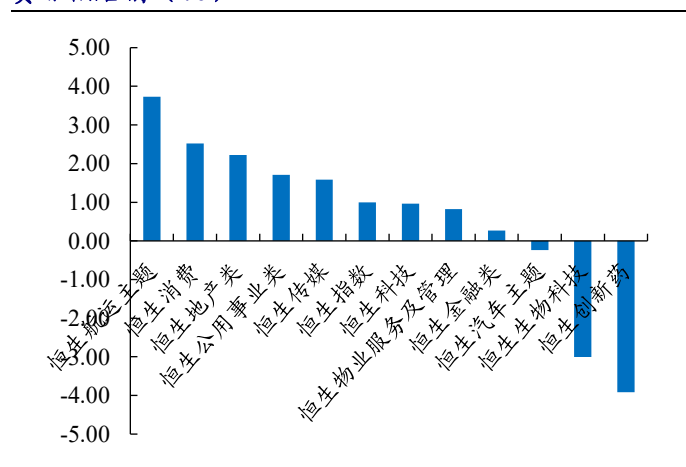
数据来源：Wind、开源证券研究所

图9：本周（2026.10.05-2026.10.09）恒生港股通高股息率涨幅较大（%）



数据来源：Wind、开源证券研究所

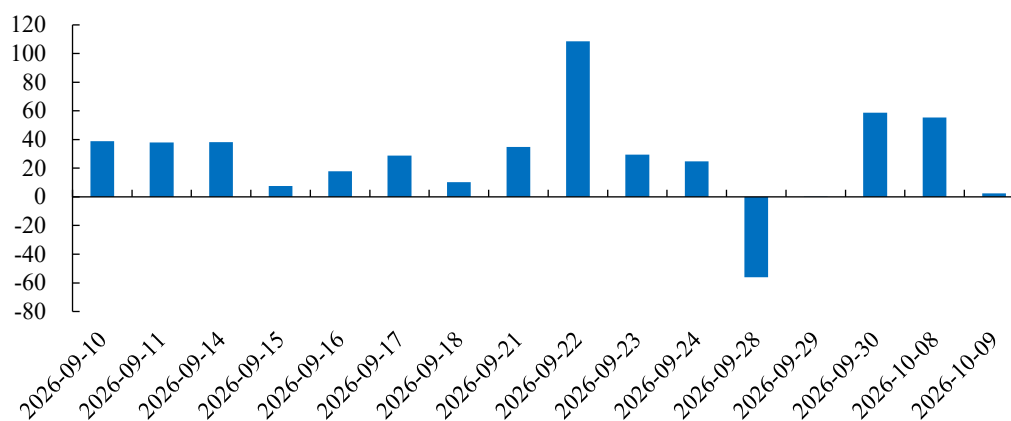
图10：本周（2026.10.05-2026.10.09）恒生航运主题、消费涨幅居前（%）



数据来源：Wind、开源证券研究所

本周港股通成交净额为流入 57.8 亿元，环比上升，从头部活跃个股交易层面来看，友邦保险（+23.8 亿港元）、腾讯控股（+21.9 亿港元）、阿里巴巴-W（+4.7 亿港元）资金净流入居前。

图11：2026年9月以来港股通当日买入成交净额整体为正（单位：亿元）



数据来源：Wind、开源证券研究所（注：10月5日、10月6日、10月7日港股通受假期影响关闭）

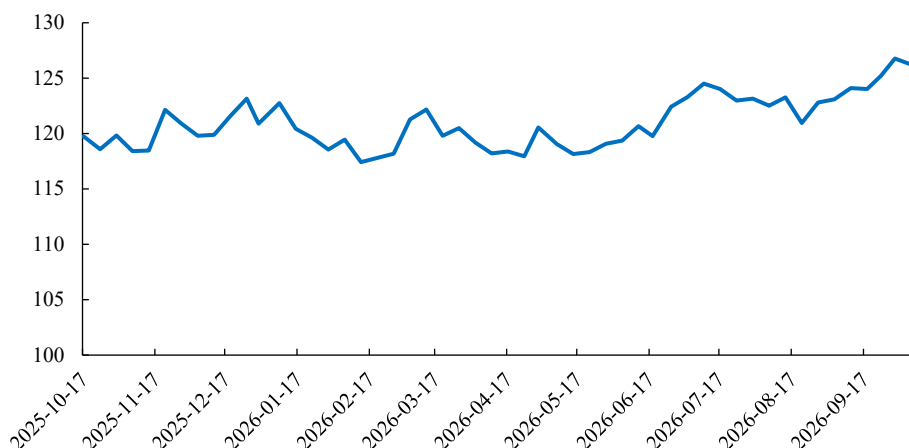
表2：本周（2026.10.05-2026.10.09）友邦保险、腾讯控股与阿里巴巴-W 流入资金居前

股票代码	股票名称	成交净买入（亿港元）	港股通持股数占总股本比例
1299.HK	友邦保险	23.8	4.1%
0700.HK	腾讯控股	21.9	11.9%
9988.HK	阿里巴巴-W	4.7	10.0%
0883.HK	中国海洋石油	2.4	22.4%
6166.HK	剑桥科技	2.4	64.6%
0100.HK	MINIMAX-W	2.2	10.5%
3696.HK	英矽智能	1.8	27.9%
2513.HK	智谱	1.4	14.1%
2269.HK	药明生物	-1.0	25.2%
1347.HK	华虹宏力	-3.5	16.0%
1548.HK	金斯瑞生物科技	-4.2	22.3%
1888.HK	建滔积层板	-4.2	12.5%
1810.HK	小米集团-W	-4.9	19.7%
0981.HK	中芯国际	-6.7	28.2%
6869.HK	长飞光纤光缆	-9.9	49.3%

数据来源：Wind、开源证券研究所（注：10月5日、10月6日、10月7日港股通受假期影响关闭）

本周恒生沪港通 AH 溢价指数为 126.06，环比有所下降，AH 股溢价触底回弹，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12：恒生沪港通 AH 溢价指数或已触底



数据来源：Wind、开源证券研究所

6、模型与算力价格数据更新

6.1、模型信息更新

9 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十余款新模型，产品定位与定价梯队分化明显。本周（2026.10.05-2026.10.09）新上线模型包括开源模型 Beam 与 Claude Haiku 5.5；待正式发布模型有 Claude Fable 5.2。

表3：9 月以来新上线模型与待发布模型

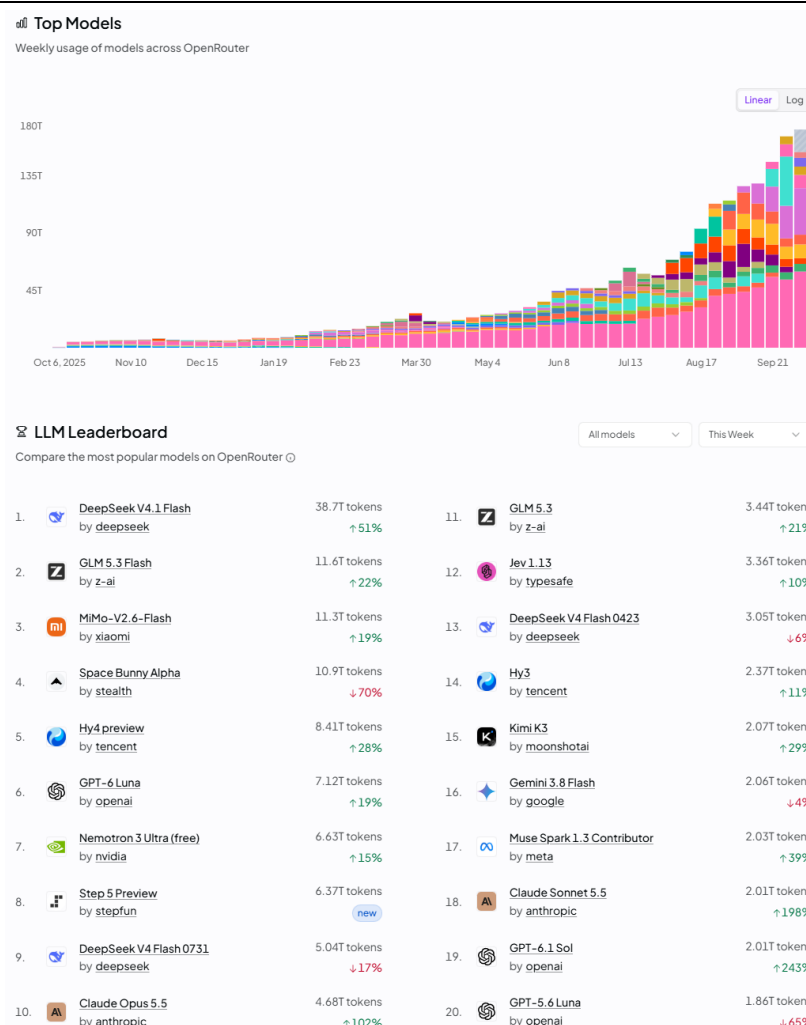
已发布模型	发布时间	厂商	输入价格（/M）	缓存命中输入价格（/M）	输出价格（/M）	上下文窗口
Claude Fable 5.1	2026-09-01	Anthropic	\$10.00	\$0.25	\$50.00	1M
Claude Mythos 5.1	2026-09-01	Anthropic	\$10.00	\$0.25	\$50.00	1M
Gemini 3.8 Flash	2026-09-02	Google	\$0.75	\$0.08	\$3.75	1M
Muse Spark 1.3	2026-09-02	Meta	\$1.25	\$0.15	\$4.25	1M
GPT-6 Astra	2026-09-03	OpenAI	\$10.00	\$1.00	\$50.00	1.05M
DeepSeek-V4.1-Flash	2026-09-10	DeepSeek	¥2.00（高峰）	¥0.04	¥8.00	1M
			¥1.00（空闲）	¥0.02	¥4.00	1M
Kimi K2.8 Preview	2026-09-11	月之暗面	目前只在 Kimi Code 与 Kimi Work 上线			1M
GLM-5.3-FlashX	2026-09-18	智谱	¥2.00	¥0.57	¥7.00	1M
Step 5 Preview	2026-09-20	阶跃星辰	¥7.00	¥0.35	¥20.00	1M
MiMo-V2.6-Pro	2026/9/22	小米	¥3.00	¥0.03	¥6.00	1M
MiMo-V2.6-Flash	2026/9/22	小米	¥1.00	¥0.02	¥2.00	1M
Grok 4.7	2026/9/22	SpaceXAI	\$2.00	\$0.50	\$6.00	500K
GPT-6 Sol	2026/9/23	OpenAI	\$2.00	\$0.20	\$10.00	1.05M
GPT-6 Luna	2026/9/23	OpenAI	\$0.10	\$0.01	\$0.50	1.05M
Claude Opus 5.5	2026/9/23	Anthropic	\$4.00	\$0.20	\$20.00	1M
MiniMax M3.1 Flash Preview	2026/9/27	MiniMax	目前只在 MiniMax Code 上线			1M
Claude Sonnet 5.5	2026/9/28	Anthropic	\$2.00	\$0.20	\$10.00	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
GPT-6.1 Sol	2026/9/29	OpenAI	\$2.00	\$0.10	\$10.00	1.05M
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	2026/10/8	Anthropic	\$0.10	\$0.01	\$0.50	1M
待正式发布模型			厂商			
Claude Fable 5.2			Anthropic			

资料来源：各模型厂商官网、IT之家、新智元、开源证券研究所

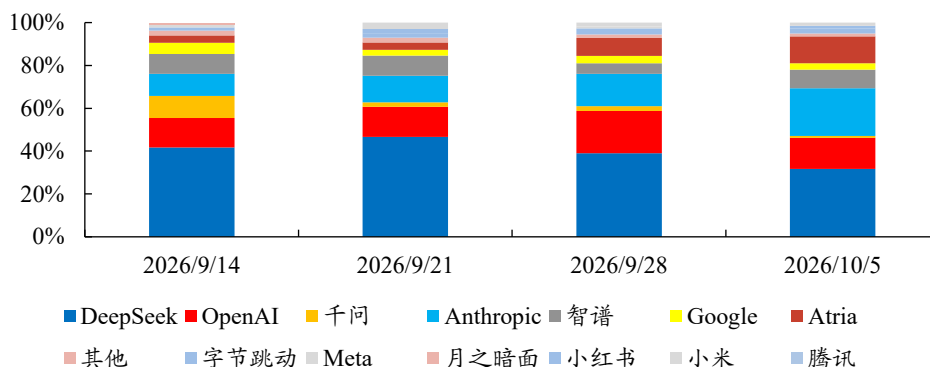
本周（2026.10.05-2026.10.09）OpenRouter 的 token 调用量持续提升，其中国产模型 DeepSeek V4.1 Flash、GLM 5.3 Flash 与 MiMo-V2.6-Flash 位居前列。截至 10 月 05 日，在 ZenMux 统计的 Token 市场份额中，国产 AI 占比接近 50%，DeepSeek 市场份额保持第一，占比为 31.7%。依托开源生态及出色的性价比优势，国产模型持续扩大市场影响力。

图13: OpenRouter 的 token 调用量持续提升



资料来源：OpenRouter 官网

图14: 国产 AI 在 ZenMarket 市场份额占比接近 50%



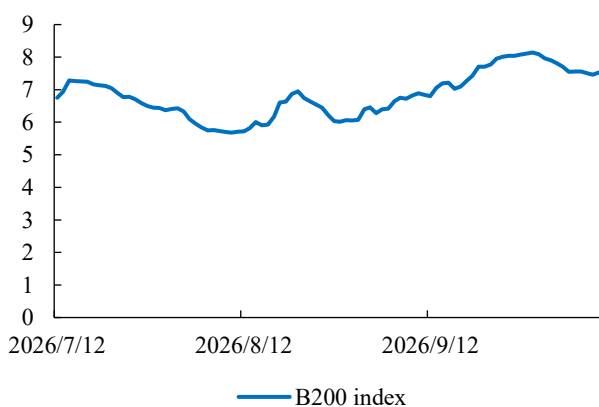
数据来源: ZenMux、开源证券研究所

6.2、算力租赁价格更新

GPU 租赁价格整体高位运行，不同性能 GPU 价格走势有所分化。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 租赁价格走势坚挺，而 H100 与 A100 租赁价格走势则偏弱。

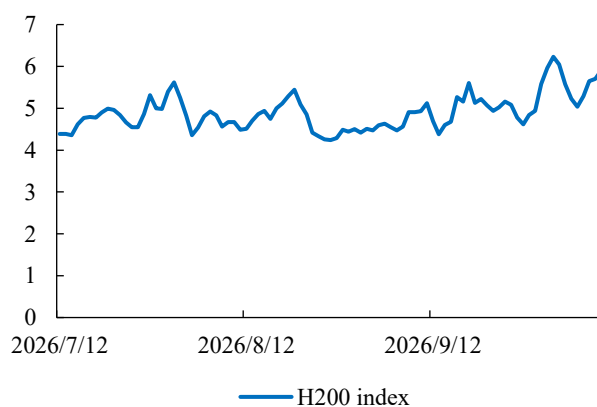
截至 10 月 10 日，B200 租赁价格为 7.53 \$/小时；H200 租赁价格为 5.91\$/小时；H100 租赁价格为 2.67\$/小时；A100 租赁价格为 0.95 \$/小时。

图：B200 租赁价格高位运行（单位：\$/小时）



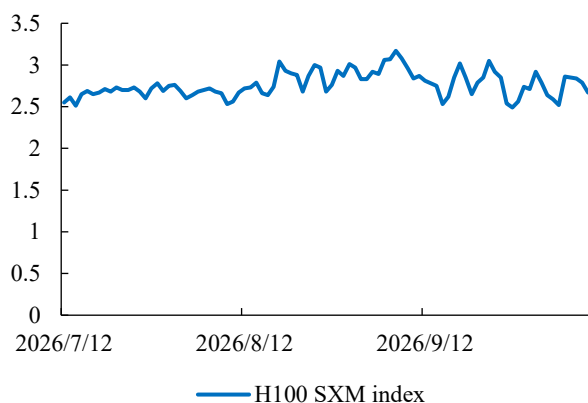
数据来源: Ornn、开源证券研究所

图：H200 租赁价格高位运行（单位：\$/小时）



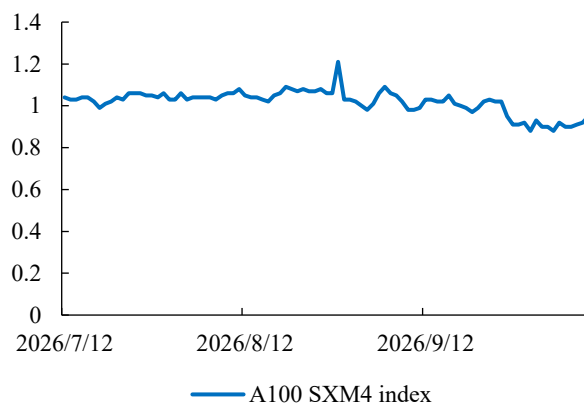
数据来源: Ornn、开源证券研究所

图：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：A100 租赁价格有所下降（单位：\$/小时）



数据来源：Ornn、开源证券研究所

7、投资建议

AI 硬件：我们认为此次 AMD 收购 WorldLabs 有望强化 AMD 在世界模型领域的研发能力，推动 InstinctGPU 与 ROCm 软硬件协同优化，并拓展机器人仿真及物理 AI 市场，带动 EPYC 与嵌入式芯片需求，提升长期竞争力。Marvell 的指引上修隐含 1.6T 升级及 Scale-Up 光化有望推动硅光晶圆和高速电芯片代工需求增长，利好 Tower 半导体；光引擎渗透率提升将带动 CW 激光器需求放量，受益标的：Lumentum、Coherent。

服务器：服务器 OEM 行业进入 AI 算力+传统 IT 换机双轮驱动的上行周期，需求从单一超大规模云厂扩散至新云/企业/主权算力，产品形态从单机硬件向整机柜液冷一体化方案升级；行业盈利模型重构，具备客户结构、全栈方案、供应链调度、高密度交付能力的头部 OEM 厂商业绩确定性更强，受益标的：联想、Dell、HPE。

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 应用支撑算力建设，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、Nebius、CoreWeave。

8、风险提示

AI 资本开支增速放缓风险：管理层 FY28 前瞻属于前瞻性判断，若客户融资、利用率或回报率低于预期，可能造成订单调整或交付延后。

CPU 竞争格局变化风险：Vera 放量既验证 CPU 需求，也可能加剧 Arm 与 x86 竞争；AMD 等传统 CPU 厂商的实际受益程度取决于产品竞争力和客户份额。

CPO 产业化不及预期风险：CPO 仍涉及光源可靠性、硅光良率、散热、封装、客户认证和可维护性，Spectrum-6 支持 CPO 不等同于 CPO 版本已形成大规模收入。

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。

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