

海外科技

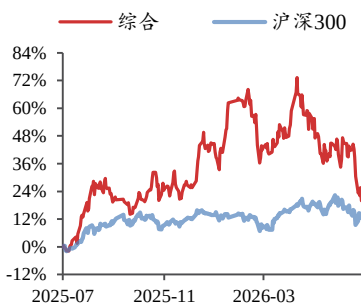
2026 年 08 月 09 日

投资评级：看好（维持）

云计算需求持续验证，DeepSeek 拟涨价强化开源模型通胀信心

——行业周报

行业走势图



数据来源：聚源

相关研究报告

《模型收敛趋势加剧，高端卡紧缺，成熟卡价格重估——港股行业深度报告》-2026.8.3

《AI 硬件迎来三重催化，ASMPTQ2 业绩表现强劲——行业周报》-2026.8.2

《教育/时代天使/美图 2026H1 业绩高增，奢牌加码女表——行业周报》-2026.8.2

初敏（分析师）

chumin@kysec.cn

证书编号：S0790522080008

杨哲（分析师）

yangzhe@kysec.cn

证书编号：S0790524100001

荀月（分析师）

xunyu@kysec.cn

证书编号：S0790524110001

● 云：需求持续验证，收入增速及利润率持续提升，Capex 上修幅度放缓

云巨头 Capex 上修幅度放缓、需求从前沿模型厂商向 B 端客户扩散叠加开源模型与闭源模型性能收敛加速 token 商品化，收入及利润率持续改善，ROI 初步验证。根据 Synergy，2026Q2 全球企业云基础设施开支总规模达 1430 亿美元，同比增长 43%；头部厂商中，亚马逊、微软、谷歌份额分别为 28%、20%、15%。中长期看算力集群仍是核心壁垒。随 AI 需求从训练逐步转向推理，供给端电力、内存等要素成本及供应偏紧，盈利端仍需依赖自研芯片及规模杠杆实现改善，新增需求及估值溢价向具备 GPU/TPU/ASIC 异构算力、模型能力、积压订单可见性及资本开支投入能力的公司扩散。受益标的谷歌、亚马逊、微软、Meta、Nebius、CoreWeave、金山云。

● 模型：国产大模型商业化信号增强，百融云积极回购彰显公司信心

(1) DeepSeek 预告整体提价，背后可能对应需求增长、推理容量和成本回收压力，重点看正式价格落地后调用量能否保持；Qwen 拟引入分成，则把商业化边界从阿里云/API 收费延伸到客户自有数据中心的商业部署，模型厂商开始尝试分享下游应用收入。(2) 百融云近期发布盈利预警，市场对九号文带来的业绩影响已有预期，该同时公司迅速启动股份回购，彰显管理层对长期业务前景充足信心。短期业绩承压主要源于 AI 战略投入，而物流标杆项目落地标志硅基员工商业模式得到跨行业验证，以百融云为代表的垂类应用依托场景及数据优势有望迎来板块修复机会。

● 互联网：市场或重新转向重新定价具备全栈 AI 能力的巨头

具备超级 APP 入口与全链路生态协同能力的厂商有望持续扩大领先优势。首选 AI 落地领先、具备全场景生态壁垒的龙头公司，开源模型能力持续提升下，AI 云需求有望持续验证，国产 AI 芯片增长迅速，推荐阿里巴巴-W、百度集团-SW，受益标的腾讯控股。

● 投资建议：

AI：云服务需求与推理算力需求持续增长，北美云计算高景气延续，看好具备全栈能力的头部平台及服务长尾客户需求的 Neocloud；国产大模型迭代带动商业化变现能力提升；AI 智能体产业进入快速增长期，受益标的谷歌、亚马逊、微软、Nebius、CoreWeave、智谱、minimax、百融云-W、迅策、德适-B、金山云。

互联网：关注 AI 商业化及应用场景拓展，随多模型能力持续提升、Agent 技术与应用生态快速成熟，国产 AI 芯片增长迅速；全球化出海有望由粗放扩张转向精细化盈利，推荐阿里巴巴-W、百度集团-SW，受益标的腾讯控股。

● **风险提示：**产能及供应链风险、监管政策变动、经济增长放缓、地缘政治风险。

目 录

1、云：需求持续验证，Capex 上修幅度放缓	4
2、服务器：基础设施建设需求带动 2026 年 AI 服务器出货量上调.....	7
3、模型及软件： DeepSeek 拟涨价强化开源模型通胀信心，百融云发掘物流行业拓展增量	8
3.1、 DeepSeek 预告涨价、Qwen 拟引入分成，国产大模型商业化信号增强	8
3.2、 百融云积极回购彰显公司信心，物流行业成为重要增量.....	9
3.2.1、 进军物流场景： AI Rollup 战略重要里程碑，标杆客户验证复制能力	9
3.2.2、 基石业务：服务用户全生命周期，硅基员工转化效率显著	11
3.2.3、 商业模式：三种收费模式满足不同客户需求.....	13
3.2.4、 行业分析：AICC 赛道广阔，AI 替代空间广阔	14
3.2.5、 技术壁垒：自研垂类大模型 + 业务数据闭环飞轮.....	15
4、周度数据更新	16
4.1、 Token 调用量保持高位，国产模型持续扩大市场影响力	16
4.2、 港股数据更新	18
5、投资建议	20
6、风险提示	20

图表目录

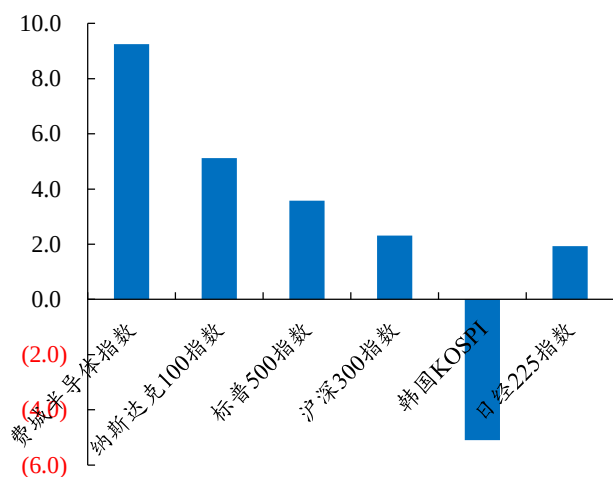
图 1： 20260803-0807 费城半导体指数涨 9.2%（%）	4
图 2： 20260803-0807Coreweave 涨 26.3%（单位：%）	4
图 3： 全球云收入增长持续加快	5
图 4： TrendForce 预计 2027 年全球 CSP 资本开支同比增速放缓	6
图 5： 2026 年 1 月 TrendForce 预计全年 AI 服务器出货量同比增速为 28.3%.....	8
图 6： 百融智能以 AICC 垂类大模型为底座，打造企业级硅基员工闭环能力	10
图 7： 硅基客服大幅提升物流工单处理效率.....	10
图 8： 物流场景硅基客服部分工作效率、成本优于碳基客服.....	11
图 9： 券商场景验证硅基员工价值，关键业务指标优于传统人工座席	12
图 10： 券商场景商业价值持续验证，硅基员工促成交易额快速提升	12
图 11： 银行场景实现规模化触达，AI 员工提升运营效率.....	13
图 12： 三大付费模式覆盖客户需求	14
图 13： AICC 赛道空间广阔，企业 AI 化改造需求持续释放.....	14
图 14： 百融云中后期可 AI 化改造的人力成本空间为 1496 亿元.....	15
图 15： 三大自研垂类大模型形成技术壁垒，驱动 AICC 能力持续迭代	15
图 16： 真实业务数据持续反哺模型迭代，形成数据闭环优势	16
图 17： OpenRouter 的 token 调用量保持高位	17
图 18： 国产模型在 ZenMux 的 Token 市场份额占比超过 60%.....	17
图 19： 本周（20260803-0807）恒生指数跌 0.8%（%）	18
图 20： 本周（20260803-0807）恒生科技涨 0.6%.....	18
图 21： 本周（20260803-0807）恒生创新药、生物科技、涨幅较大	18
图 22： 2026 年 3 月以来港股通当日买入成交净额有所波动.....	19
图 23： 恒生沪港通 AH 溢价指数或已触底.....	20

表 1：北美云厂商收入同比增速及利润率持续提升.....	6
表 2：本周（20260803-0808）资金流入 MINIMAX-W、阿里巴巴-W 等标的（仅计算前十大活跃个股明细）	19

1、云：需求持续验证，Capex 上修幅度放缓

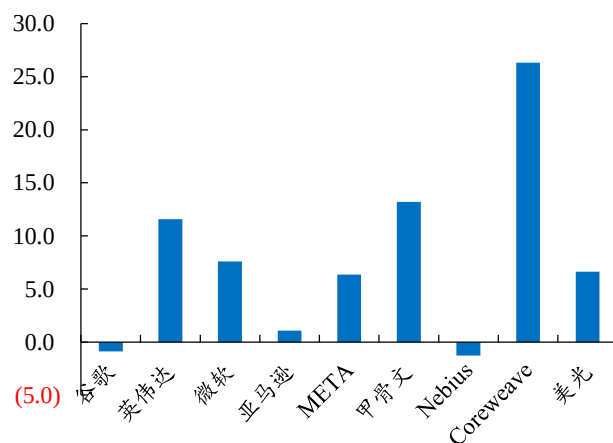
本周（20260803-0808）海外科技板块分化，云及应用层迎来兑现时刻，硬件端高位分化。费城半导体指数全周涨 9.2%，纳斯达克 100 指数全周涨 5.1%，韩国 KOSPI 全周跌 5.1%。（1）Coreweave：周涨 26.3%。（2）英伟达：周涨 11.6%。

图1：20260803-0807 费城半导体指数涨 9.2%（%）



数据来源：Wind、开源证券研究所

图2：20260803-0807Coreweave 涨 26.3%（单位：%）

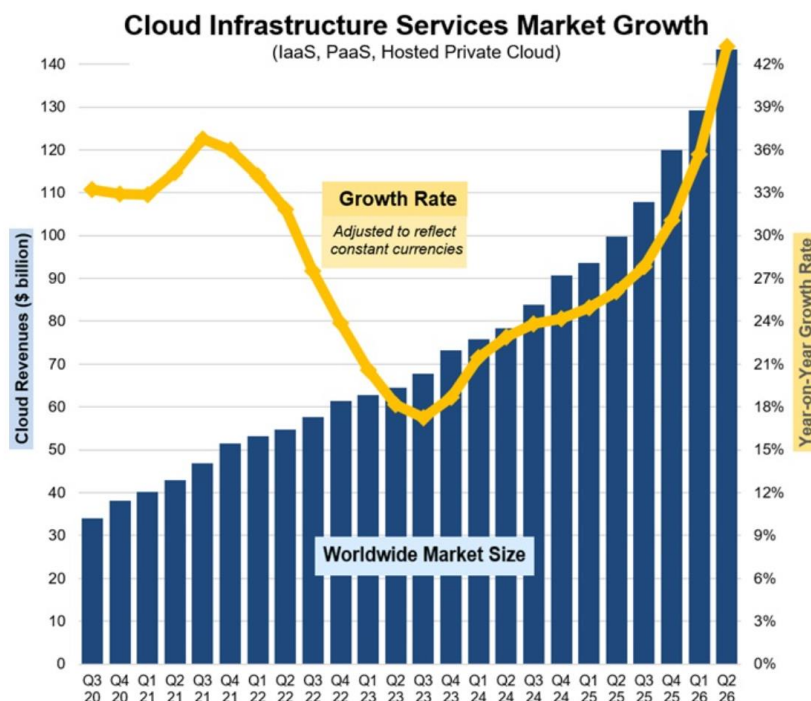


数据来源：Wind、开源证券研究所

云：巨头 Capex 上修幅度放缓、需求从前沿模型厂商向 B 端客户扩散+开源模型加速 token 商品化，收入及利润率持续改善，ROI 初步验证。

（1）需求端：根据 Synergy Research Group，2026Q2 全球企业云基础设施开支总规模达 1430 亿美元，同比增长 43%，增速持续加快。头部厂商中，亚马逊、微软、谷歌份额分别为 28%、20%、15%；二线厂商中 Coreweave、OpenAI、Oracle、Crusoe、Nebius、Anthropic 及 Nscale 增长较快。

图3：全球云收入增长持续加快



资料来源：Synergy Research Group

(2) 支出端：根据 TrendForce 数据，2026 年全球九大云端服务供应商（CSP）总资本支出预估将同比增长约 90%，其中字节跳动、腾讯、阿里巴巴、百度的资本开支合计增长 80%；2027 年预计全球九大 CSP 资本开支总额为 1.32 万亿美元，同比增长约 49%，增速放缓主要是反映高基数因素，并非 AI 投资趋缓。在 AI 推理、Agentic AI、自研 ASIC 和新一代 AI 模型持续推升算力需求的情况下，产业成长有望从原先以 GPU 扩张为主，转为包含 AI 服务器、液冷散热、先进封装、高速互连、电源和存储器等基础设施全面升级的态势。

图4：TrendForce 预计 2027 年全球 CSP 资本开支同比增速放缓

2024-2027年全球九大CSP资本支出总额预估



资料来源：TrendForce 公众号

长期看云计算核心环节仍是 AI 芯片，随着 AI 需求从训练逐步转向推理，新增需求及估值溢价向具备 GPU/TPU/ASIC 异构算力、模型能力、积压订单可见性及资本开支投入能力的公司扩散。

表1：北美云厂商收入同比增速及利润率持续提升

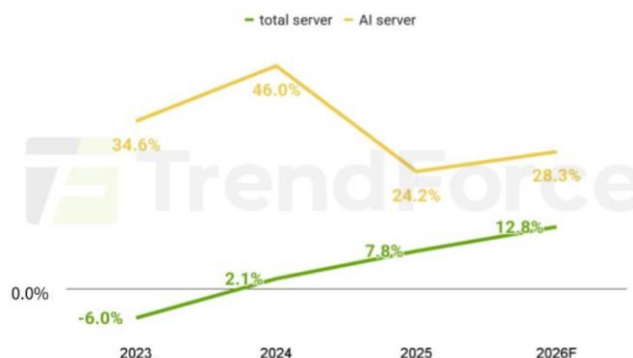
	2024Q1	2024Q2	2024Q3	2024Q4	2025Q1	2025Q2	2025Q3	2025Q4	2026Q1	2026Q2
全球云服务市场规模(亿美元)	751	782	820	860	909	1000	1070	1191	1290	1430
YOY(%)	13%	19%	21%	20%	21%	22%	28%	30%	35.0%	43%
市场份额	2024Q1	2024Q2	2024Q3	2024Q4	2025Q1	2025Q2	2025Q3	2025Q4	2026Q1	2026Q2
亚马逊（AWS）	33%	33%	33%	33%	32%	32%	32%	28%	29%	28%
微软（Azure）	21%	20%	20%	20%	23%	22%	22%	21%	21%	20%
谷歌（Google Cloud）	10%	10%	10%	11%	10%	11%	11%	14%	16%	15%
云业务收入（百万美元）	2024Q1	2024Q2	2024Q3	2024Q4	2025Q1	2025Q2	2025Q3	2025Q4	2026Q1	2026Q2
亚马逊（AWS）	25,037	26,281	27,452	28,786	29,267	30,873	33,006	35,579	37,587	42,232
YOY(%)	17.2%	18.7%	19.1%	18.9%	16.9%	17.5%	20.2%	23.6%	28.4%	36.7%
微软（Azure）	14,472	15,437	16,885	18,007	19,291	21,519	23,572	25,102	27,046	30,772
YOY(%)	35%	34%	33%	31%	33%	39%	40%	39%	40%	43%
谷歌（Google Cloud）	9,574	10,347	11,353	11,955	12,260	13,624	15,157	17,664	20,028	24,768
YOY(%)	28.4%	28.8%	35.0%	30.1%	28.1%	31.7%	34.0%	48.0%	63%	82%
云业务经营利润率(%)	2024Q1	2024Q2	2024Q3	2024Q4	2025Q1	2025Q2	2025Q3	2025Q4	2026Q1	2026Q2
亚马逊（AWS）	37.6%	35.5%	38.1%	36.9%	39.5%	32.9%	34.5%	35.0%	37.8%	39.4%
微软（Azure 及其他企业服务）	44.4%	41.4%	43.6%	42.5%	41.5%	40.6%	43.3%	42.2%	39.7%	40.6%
谷歌（Google Cloud）	9.4%	11.3%	17.1%	17.5%	17.8%	20.7%	23.7%	30.0%	33.0%	35.6%
资本开支（百万美元）	2024Q1	2024Q2	2024Q3	2024Q4	2025Q1	2025Q2	2025Q3	2025Q4	2026Q1	2026Q2
亚马逊	13,935	16,393	21,278	26,052	24,255	31,368	34,228	38,469	43,234	51,757
YOY	5%	54%	81%	91%	68%	83%	55%	62%	78%	65%
微软	14,000	19,000	20,000	22,600	21,400	24,200	34,900	37,500	31,900	41,000

YOY	79%	78%	79%	97%	53%	27%	75%	66%	49%	69%
谷歌	12,012	13,186	13,061	14,276	17,197	22,446	23,953	27,851	35,674	44,924
YOY	91%	91%	62%	30%	43%	70%	83%	75%	107%	100%
Meta	6,715	8,472	9,202	14,836	13,692	17,012	19,374	22,137	19,840	31,078
YOY	-5.2%	33.3%	36.1%	87.8%	103.9%	100.8%	110.5%	49.2%	44.9%	82.7%
资本开支计划										
亚马逊	2026 年计划资本开支上调至 2200 亿美元（前值：2000 亿美元）									
微软	2026 日历年资本开支下调至 1750 亿美元（此前指引 1900 亿美元，T 调整主要源于会计准则变化，整体支出规模不变）									
谷歌	2026 年计划资本开支上调至 1950-2050 亿美元，2027 年显著高于 2026 年									
META	2026 年计划资本开支 1300 -1450 亿美元（下限上调，前值 1250-1450 亿美元）									
自由现金流（百万美元）	2024Q1	2024Q2	2024Q3	2024Q4	2025Q1	2025Q2	2025Q3	2025Q4	2026Q1	2026Q2
亚马逊	5,054	8,888	4,693	19,584	(7,240)	1,147	1,297	15,990	(17,202)	(12,170)
YOY	271%	-83%	-76%	-7%	-243%	-87%	-72%	-18%	138%	-1161%
微软	17,917	18,195	14,180	(309)	15,644	18,447	10,157	(1,742)	14,779	14,441
YOY	8%	1%	-27%	-104%	-13%	1%	-28%	-464%	-6%	-22%
谷歌	16,836	13,454	17,637	24,837	18,953	5,301	24,461	24,551	10,116	(5,855)
YOY	-2%	-38%	-22%	215%	13%	-61%	39%	-1%	-47%	-210%
Meta	12,531	10,898	15,522	13,152	10,334	8,549	10,625	14,077	12,386	784
YOY	81%	-1%	14%	14%	-18%	-22%	-32%	7%	20%	-91%
前瞻指标（亿美元）	2024Q1	2024Q2	2024Q3	2024Q4	2025Q1	2025Q2	2025Q3	2025Q4	2026Q1	2026Q2
亚马逊云 backlog	1,510	1,566	1,610	1,770	2,050	2,230	2,420	2,450	3,640	4,960
YOY(%)	34.7%	28.2%	21.1%	24.5%	35.8%	42.4%	50.3%	38.4%	77.6%	122%
QOQ(%)	6.2%	3.7%	2.8%	9.9%	15.8%	8.8%	8.5%	1.2%	48.6%	36.3%
微软 RPO	2,350	2,690	2,590	2,980	3,150	3,680	3,920	6,250	6,270	6,780
YOY(%)	20%	20%	22%	34%	34%	37%	51%	110%	99%	84%
QOQ(%)	3%	14.5%	-3.7%	15.1%	5.7%	16.8%	6.5%	59.4%	0.3%	8.1%
谷歌云 Backlog	780	850	980	1,060	1,220	1,440	1,550	2,400	4,620	5,140
YOY(%)	26.4%	34.5%	47.8%	43.0%	56.4%	69.4%	58.2%	126.4%	278.7%	256.9%
QOQ(%)	5.3%	9.0%	15.3%	8.2%	15.1%	18.0%	7.6%	54.8%	92.5%	11.3%

资料来源：彭博、公司公告、雅虎财经、路透社、金融时报、Synergy Research Group

2、服务器：基础设施建设需求带动 2026 年 AI 服务器出货量上调

根据 TrendForce，考虑到近期超大型 CSP、Tier-2 数据中心业者对英伟达 GB/VR 等整柜式 AI 服务器方案的采购意愿提高，而且 Google（谷歌）、AWS（亚马逊云科技）新一代 ASIC 平台于 2026 年下半年陆续放量，加上中系业者扩大采用本土 AI 方案以满足云端大语言模型 AI 需求，TrendForce 集邦咨询上调 2026 年 AI 服务器出货年同比增速，从原先的 28%更新为近 31%。

图5：2026 年 1 月 TrendForce 预计全年 AI 服务器出货量同比增速为 28.3%
2023-2026年全球server、AI server出货量年增/减变化


资料来源：TrendForce 官网

TrendForce 预估 Google 2026 至 2027 年都将侧重自家 TPU 芯片发展, 2027 年出货量有望继续高速增长。AWS 2026 年以 NVIDIA GB300 作为 GPU AI 服务器主力, 自研 ASIC 也稳定出货, 预估至 2027 年产能将再扩大。Meta 2026 年主要采用 NVIDIA GB/VR、AMD (超威) Helios 整柜式方案, 自研 ASIC 布局预计至 2027 年将明显加速, 为该公司长期降低 AI 基础设施成本、提升推理效率的重要策略。

3、模型及软件： DeepSeek 拟涨价强化开源模型通胀信心， 百融云发掘物流行业拓展增量

3.1、DeepSeek 预告涨价、Qwen 拟引入分成，国产大模型商业化信号增强

8 月 6 日, DeepSeek 官方定价页近日新增提示: 计划近期整体上调 API 服务定价, 预计涨幅较大, 具体方案以正式通知为准; 当前 V4-Flash/V4-Pro 输出价格为 2/6 元/百万 tokens; 据路透社报道, 阿里计划在 Qwen3.8-Max 开放权重后, 要求大型商业用户分享部分收入, 具体比例仍在谈判, 机制参考 Kimi K3。

我们认为, 两个事件共同推动行业叙事从低价获客走向商业验证。DeepSeek 主动预告整体提价, 背后可能对应需求增长、推理容量和成本回收压力, 重点看正式价格落地后调用量能否保持; Qwen 拟引入分成, 则把开源模型商业化边界从 API 收费延伸到客户自有数据中心的商业部署, 模型厂商开始尝试分享下游应用收入。

需求与供给方面, DeepSeek 选择预告整体提价, 说明价格调整已进入商业化测试阶段。即使 V4-Flash 定位偏高性价比, 仍进入涨价通道, 说明模型服务的价格空间取决于使用价值、供给约束和推理成本; 正式涨价后的调用量、客户留存和成本变化, 是判断需求弹性与利润兑现的关键。

投资含义上, Qwen 分成协议若落地, 将形成“云/API 收费+开放权重授权+下游收入分成”的新变现路径。阿里官方已披露, AI 相关产品收入占外部收入 30%, 百

炼平台客户数同比增长 8 倍，商业化已有收入侧验证。短期需等待 Qwen3.8-Max 正式许可证、收入门槛与分成比例；DeepSeek 则观察正式调价、调用量和推理成本变化。我们倾向于认为，国内模型竞争正在从单纯价格竞争转向算力效率、服务稳定性和商业客户价值的比拼，涨价公告本身仍不能替代利润兑现。

3.2、百融云积极回购彰显公司信心，物流行业成为重要增量

公司近期发布盈利预警，市场对九号文带来的业绩影响已有预期，同时公司迅速启动股份回购，彰显管理层对长期业务前景充足信心。短期业绩承压主要源于 AI 战略投入，而物流标杆项目落地标志硅基员工商业模式得到跨行业验证。以百融云为代表的垂类应用依托场景及数据优势有望迎来板块修复机会。

3.2.1、进军物流场景：AI Rollup 战略重要里程碑，标杆客户验证复制能力

公司以 AICC 垂类 SOTA 大模型为技术底座，向企业客户输出企业级 AI Agent（硅基员工）落地，并积累真实企业数据反哺模型研发，形成“模型—Agent—真实企业数据”的闭环飞轮。AI Rollup 是公司的核心战略，其内涵是通过 AICC 技术对传统人力密集型客服/营销/运营模式进行 AI 化改造，实现“硅基员工”对“碳基员工”的替代与赋能。战略执行分为三步：

第一步：打磨 AICC 垂类 SOTA 大模型。公司专注于智能联络场景的通话、对话、话术能力，自研三大领域专属大模型——BR-Voice 语音多模态大模型、BR-LLM-Proactive 主动引导式对话大模型、BR-Vision-Doc 文档理解大模型，在语音识别、对话理解、文档解析等维度达到行业领先水平。

第二步：将大模型能力输出为硅基员工。公司将 AICC 垂类大模型能力产品化为可直接上岗的硅基员工，直接替代传统的按工时或按座席收费模式。

第三步：跨行业输出硅基员工能力。公司以金融行业为起点，逐步将 AICC 能力向券商、物流、运营商、保险、教育、医疗等全行业输出，对客户传统人力密集型客服/营销/运营模式进行 AI 化改造，实现规模化收入增长。

图6：百融智能以 AICC 垂类大模型为底座，打造企业级硅基员工闭环能力

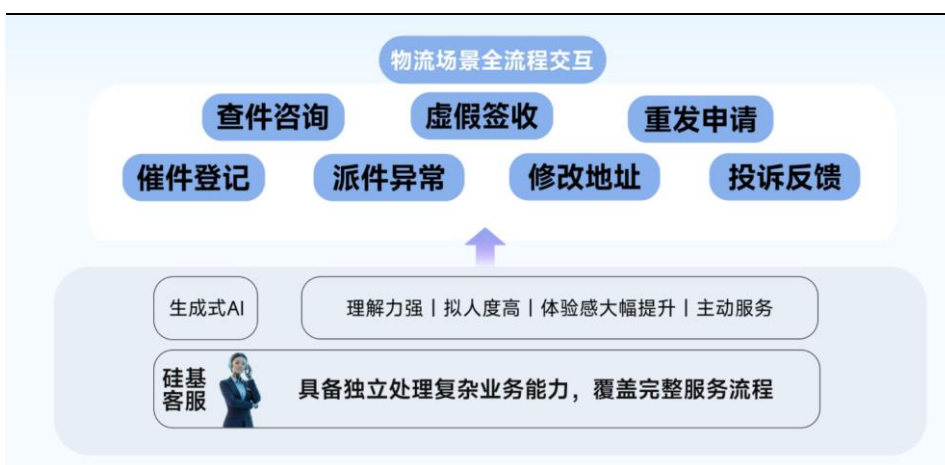


资料来源：百融智能公司材料

其中，物流场景是公司 AI Rollup 战略的重要里程碑。2026 年 7 月，公司在公告中披露物流行业标杆客户实质性落地，这是 AI Rollup 战略第一个里程碑式验证，证明 AICC 能力不局限于金融行业，具备跨行业复制能力。

物流场景的硅基客服可以覆盖全流程交互，包括：查件咨询、虚假签收、重发申请、修改地址、投诉反馈、催件登记、派件异常、服务满意度采集等。这些场景都是物流行业高频的客服需求，人力成本高，AI 替代空间大。

图7：硅基客服大幅提升物流工单处理效率



资料来源：百融智能公司材料

效率对比方面（公司内部统计与估算口径）：平均处理话务时长，硅基员工仅需 2 分钟，是碳基员工 4 分钟的一半；月处理通话峰值，硅基员工可达 1.8 万通，远高于碳基员工的 0.5 万通；满意度方面，硅基员工已接近碳基员工 98% 的水平。硅基客服的平均工资远低于碳基员工（碳基员工全口径 8-10 万元/年），成本优势显著。

图8：物流场景硅基客服部分工作效率、成本优于碳基客服



资料来源：百融智能公司材料

3.2.2、基石业务：服务用户全生命周期，硅基员工转化效率显著

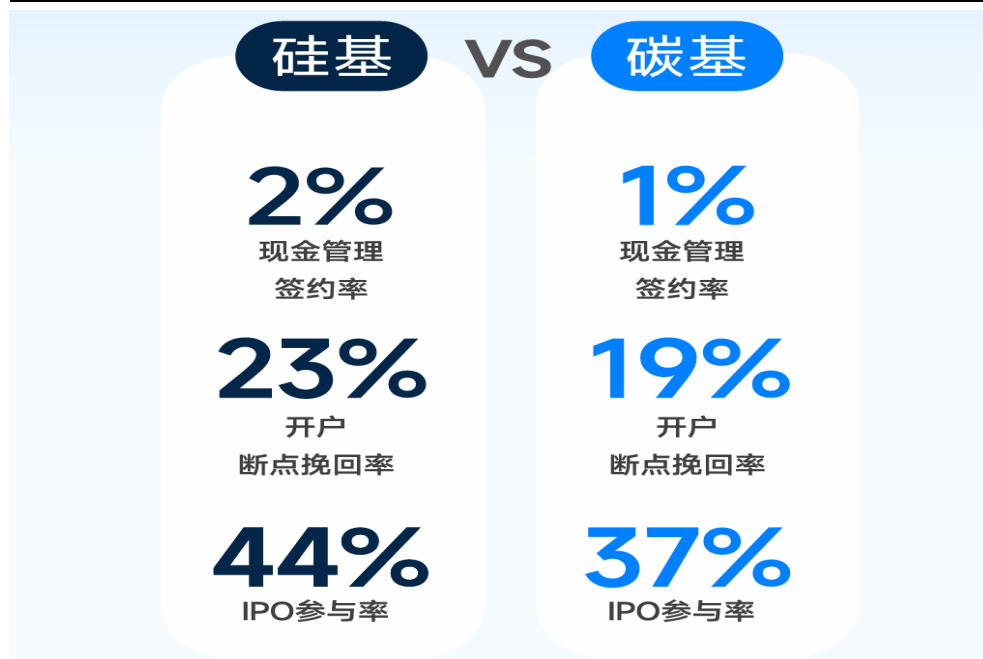
（1）券商场景：硅基员工全域运营转化效果显著

券商场景是公司核心业务场景。公司的硅基客服可以服务用户全生命周期，从获客、激活、互动、留存到推荐，覆盖券商客户运营的各个环节。具体包括：精准获客、开户引导、提升交易、挽留流失、MGM 裂变等。

券商场景的硅基员工已经在多个业务指标上超越了碳基员工。根据券商 AB 测试结果（公司内部统计与估算口径）：现金管理签约率方面，硅基员工达到 2%，高于碳基员工的 1%。开户断点挽回率方面，硅基员工为 23%，高于碳基员工的 19%。IPO 参与率方面，硅基员工达到 44%，比碳基员工的 37% 高出 7 个百分点。

硅基员工在效率上全面超越碳基员工的核心原因在于：第一，硅基员工可以 7×24 小时不间断服务，不受时间限制；第二，硅基员工的话术标准化程度高，服务质量稳定，不受情绪影响；第三，硅基员工可以基于用户画像进行精准推荐，转化率更高。

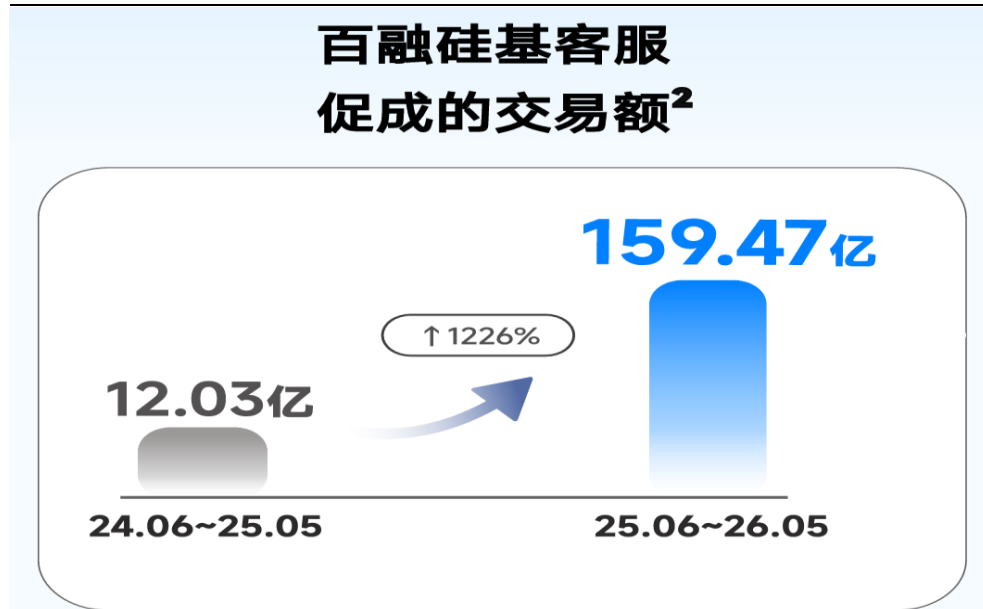
图9：券商场景验证硅基员工价值，关键业务指标优于传统人工座席



资料来源：百融智能公司材料

促成交易规模方面，券商场景硅基客服促成交易额达到 159.47 亿元（2025 年 6 月至 2026 年 5 月期间），同比增长 1226%，增长极为迅猛。该数据基于业务反馈的内部统计与估算口径，反映了硅基员工对券商客户的实际业务贡献。

图10：券商场景商业价值持续验证，硅基员工促成交易额快速提升



资料来源：百融智能公司材料，促成交易额的测算口径为：实际接通数 × 接通转化率（真实项目数据）× 人均入金金额（真实项目数据）× 人均年换手率（行业水平）

硅基员工数量方面，截至 2026 年 5 月底，公司已签约及合同流程中的券商达 14

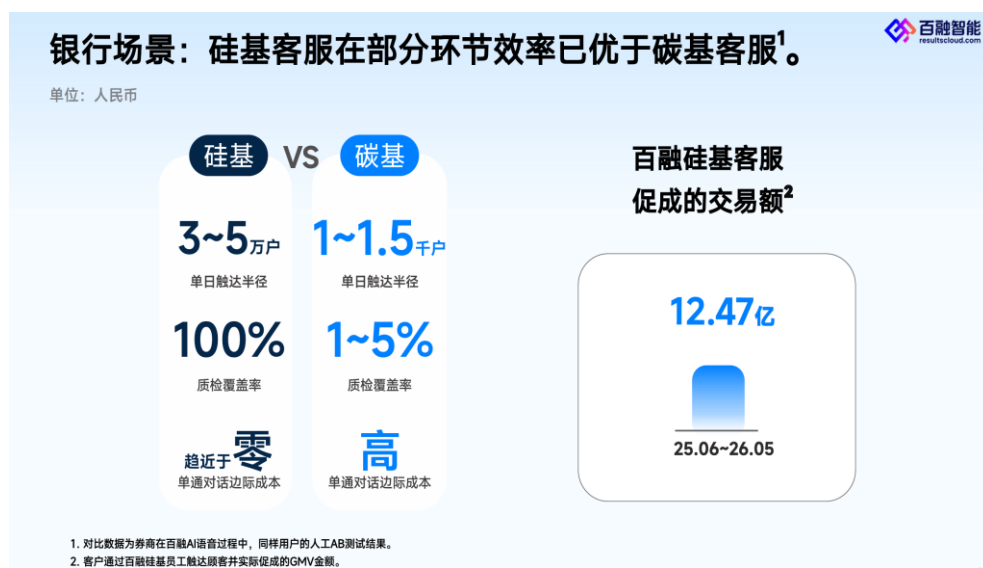
家。在某头部券商上线仅1个月，硅基客服数量就从30个增长至100个，增幅达233%。目前已跑通30个业务场景，包括休眠激活、断点挽回、入金激励等。硅基员工数量的快速增长，反映了客户对AICC技术的高度认可和旺盛需求。

(2) 银行场景：规模化触达与智能质检提升运营效率

银行场景也是公司业务场景的重要方向。银行场景的硅基员工在规模化触达和质检方面具有显著优势。2025年6月至2026年5月期间，银行场景硅基客服促成交易额为12.47亿元。但该业务受监管政策影响较大，收入整体承压。

单日触达半径方面，硅基员工可达3~5万户，而碳基员工仅为1~1.5千户，硅基员工的触达能力是碳基员工的数十倍。质检覆盖率方面，硅基员工可以实现100%全量质检，而碳基员工仅能覆盖1~5%。单通对话边际成本方面，硅基员工趋近于零，而碳基员工成本较高。

图11：银行场景实现规模化触达，AI员工提升运营效率



资料来源：百融智能公司材料

3.2.3、商业模式：三种收费模式满足不同客户需求

公司的AICC业务提供三种商业模式，满足不同客户的需求：

第一种：AI岗位外包（硅基员工）。客户将特定岗位的工作外包给公司的硅基员工，按碳基员工岗位收费，但硅基员工的工资更低，帮助客户降低成本。这种模式适用于标准化程度高、替代性强的岗位。

第二种：APO（AI业务外包）。公司承担客户的整个业务流程，按结果付费，帮助客户提升收入。这种模式与客户价值深度绑定，客户满意度高，粘性强。

第三种：调用API构建专属Agent。客户调用公司的API接口，构建自己的专属Agent，按Token量收费，帮助客户降本提效。这种模式适用于有一定技术能力、希望自主构建AI应用的客户。

图12：三大付费模式覆盖客户需求



资料来源：百融智能公司材料

3.2.4、行业分析：AICC 赛道广阔，AI 替代空间广阔

AICC 赛道具备广阔的市场空间。核心原因在于：**第一**，客服/联络中心是人力密集型行业，人力成本高企，AI 替代的经济性显著；**第二**，语音和对话是大模型技术最先成熟的应用场景之一，技术可行性高；**第三**，AI 客服可以实现 7×24 小时不间断服务，服务质量稳定，不受情绪影响，用户体验有保障。

图13：AICC 赛道空间广阔，企业 AI 化改造需求持续释放



资料来源：百融智能公司材料

中国 Contact Center 人力成本广阔，企业主对 AI 化改造需求。据百融云测算中期可 AI 化改造的人力成本空间为 1496 亿元。随着 AICC 技术不断成熟、硅基员工能力持续提升，AI 对人力座席的替代率有望逐步提高，市场空间将进一步打开。

图14：百融云中期可 AI 化改造的人力成本空间为 1496 亿元



资料来源：百融智能公司材料

3.2.5、技术壁垒：自研垂类大模型 + 业务数据闭环飞轮

以百融云为例，通过分析其竞争壁垒，可帮助我们理解软件企业的商业逻辑和成长性。我们认为软件企业依托其垂类的场景优势、数据积累及深耕行业产生的 know-how，有望借助 AI 实现持续增长。

公司自研 BR-Voice、BR-LLM-Proactive 和 BR-Vision-Doc 三大领域专属大模型，分别覆盖语音、多轮对话和文档理解，构成 AICC 智能联络的技术底座。相关模型已在语音识别、主动引导式对话、文档解析等任务中验证效果，并实现商业化落地。

图15：三大自研垂类大模型形成技术壁垒，驱动 AICC 能力持续迭代

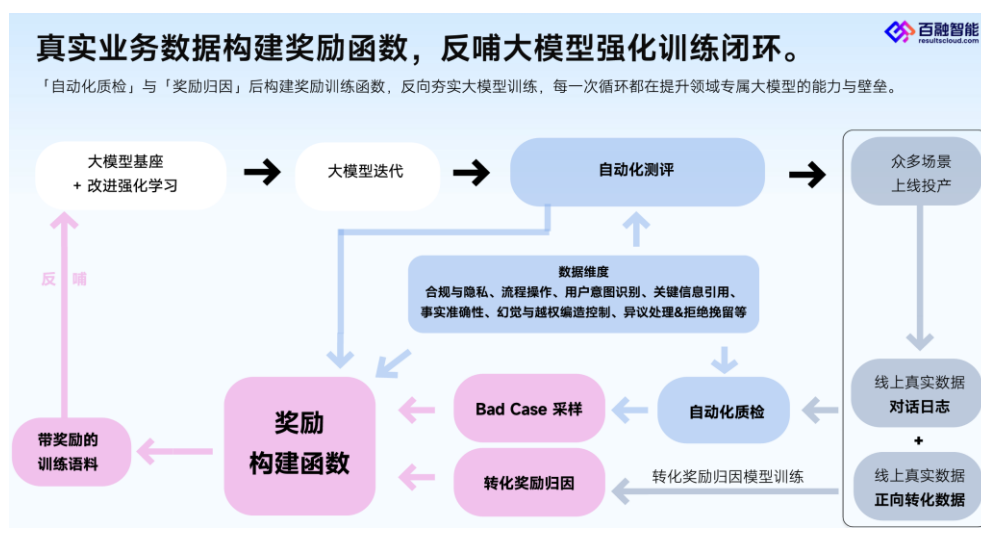


资料来源：百融智能公司材料

此外，公司构建了“大模型基座→上线投产→线上真实数据→奖励训练→反哺大模型迭代”的数据闭环飞轮。真实业务中的对话日志、转化结果和质检反馈持续反哺模型优化，形成从模型能力到客户结果、再到数据资产的正向循环。

具体而言，公司的大模型基座上线投产后，在众多场景中进行真实业务落地，积累了大量的线上真实数据，包括对话日志和正向转化数据。然后通过自动化质检和转化奖励归因进行奖励训练，将真实业务数据反哺到大模型的迭代优化中。

图16：真实业务数据持续反哺模型迭代，形成数据闭环优势



资料来源：百融智能公司材料

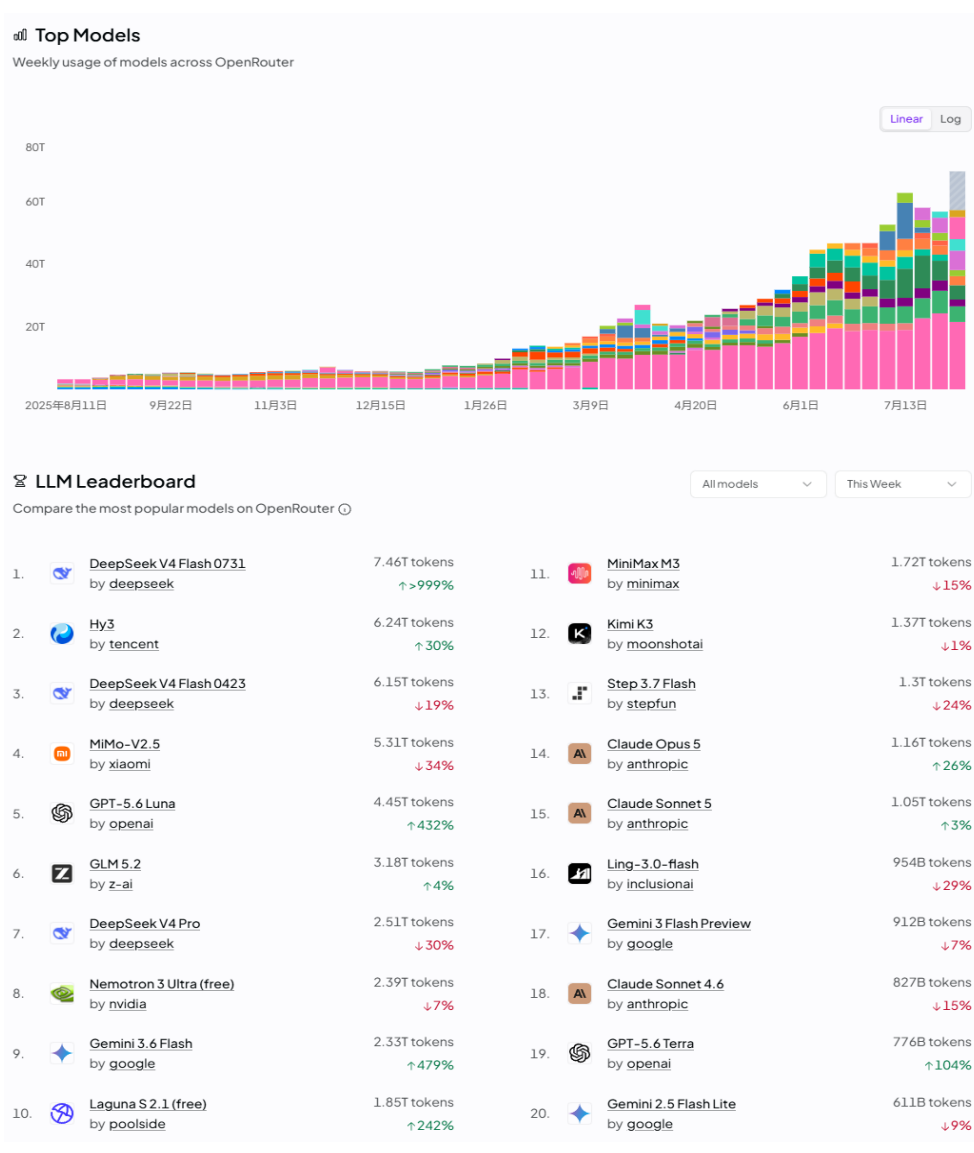
数据壁垒方面，公司拥有 10.5 亿+合规个人数据和 1.5 亿+合规企业数据，以及持续增加的真实业务行为反馈数据。所有数据的采集与使用均符合相关法律法规。海量合规数据为大模型训练提供了坚实基础，也构成了公司的核心竞争壁垒之一。

4、周度数据更新

4.1、Token 调用量保持高位，国产模型持续扩大市场影响力

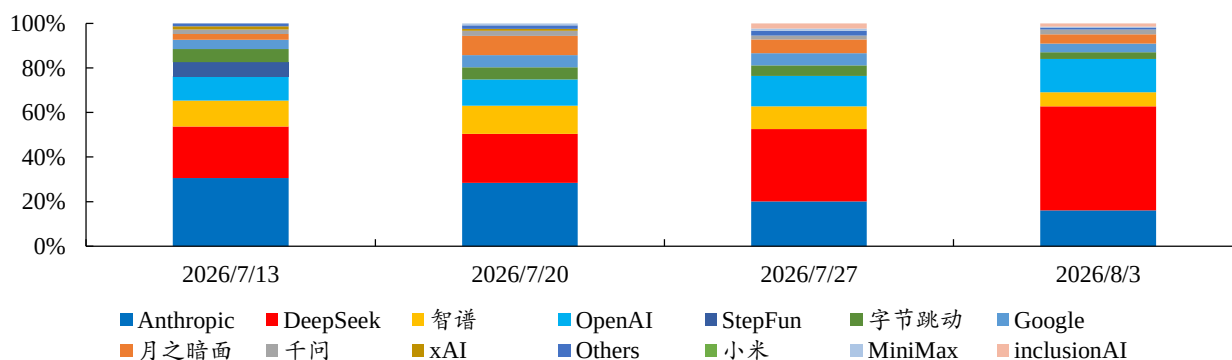
本周（20260803-0807）OpenRouter 的 token 调用量保持高位，其中国产模型 DeepSeek V4 Flash、Hy3、MiMo-V2.5 调用量位居前三。在 ZenMux 统计的 Token 市场份额中，国产 AI 占比超过 60%，DeepSeek 市场份额继续保持第一，且占比提升至 46.7%。依托开源生态及出色的性价比优势，国产模型持续扩大市场影响力。

图17: OpenRouter 的 token 调用量保持高位



资料来源: OpenRouter 官网

图18: 国产模型在 ZenMux 的 Token 市场份额占比超过 60%

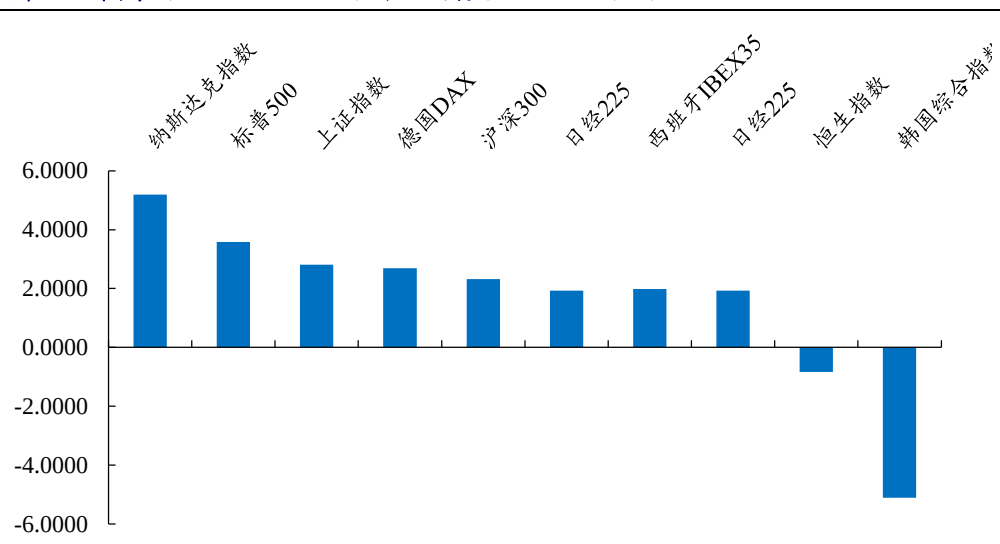


数据来源: ZenMux、开源证券研究所

4.2、港股数据更新

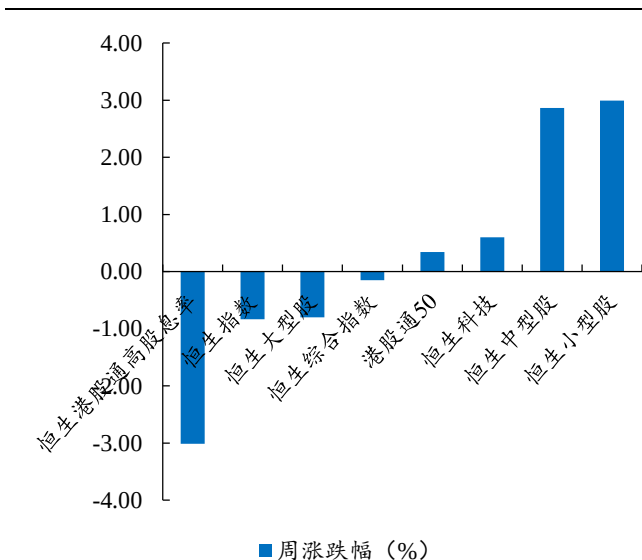
本周恒生指数跌 0.8%，从成份上看，创新药、生物科技涨幅较大。

图19：本周（20260803-0807）恒生指数跌 0.8%（%）



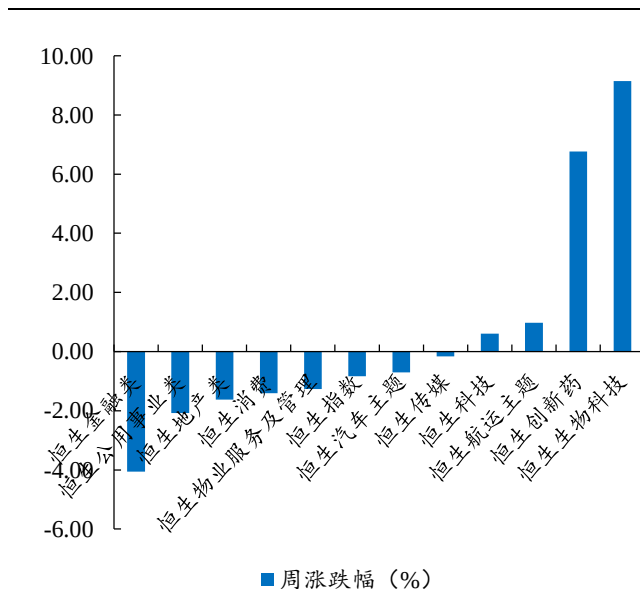
数据来源：Wind、开源证券研究所

图20：本周（20260803-0807）恒生科技涨 0.6%



数据来源：Wind、开源证券研究所

图21：本周（20260803-0807）恒生创新药、生物科技、涨幅较大

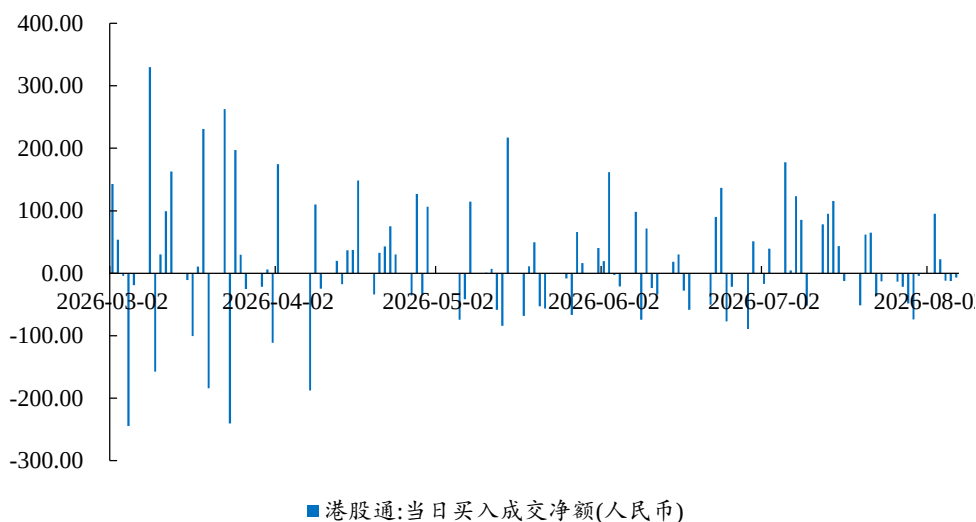


数据来源：Wind、开源证券研究所

本周港股通成交净额为流入 85.9 亿元，环比由流出转为流入，从头部活跃个股交易层面来看，MINIMAX-W (+49.2 亿港元)、阿里巴巴-W (+26.8 亿港元)、建滔

积层板（+24.7 亿港元）资金净流入居前，资金流出腾讯控股（-35.2 亿港元）、美团-W（-7.3 亿港元）、中国海洋石油（-6.2 亿港元）。

图22：2026 年 3 月以来港股通当日买入成交净额有所波动



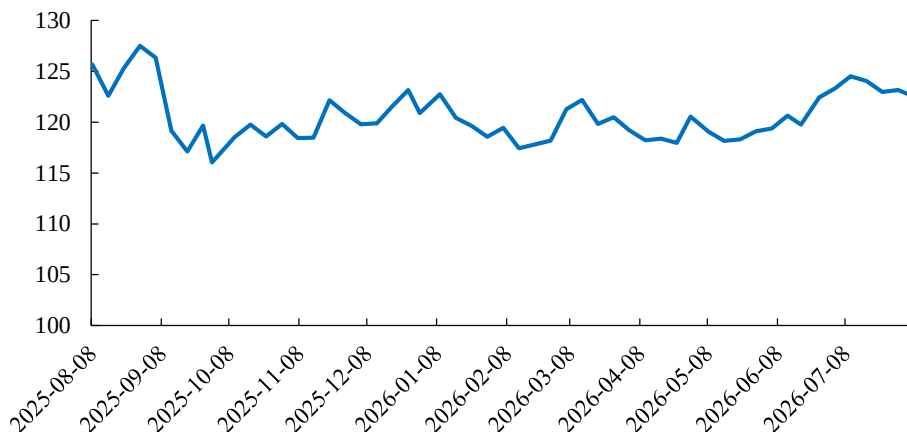
数据来源：Wind、开源证券研究所

表2：本周（20260803-0808）资金流入 MINIMAX-W、阿里巴巴-W 等标的（仅计算前十大活跃个股明细）

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(%)
0100.HK	MINIMAX-W	49.2	0.0
9988.HK	阿里巴巴-W	26.8	10.4
1888.HK	建滔积层板	24.7	13.0
1347.HK	华虹宏力	17.7	22.3
6869.HK	长飞光纤光缆	12.5	42.3
2513.HK	智谱	12.3	10.7
2269.HK	药明生物	7.8	24.0
0981.HK	中芯国际	4.7	29.2
3986.HK	兆易创新	4.0	48.8
1810.HK	小米集团-W	3.0	19.3
0148.HK	建滔集团	0.5	24.1
2476.HK	胜宏科技	(0.1)	22.9
9992.HK	泡泡玛特	(1.9)	24.7
0883.HK	中国海洋石油	(6.2)	22.3
3690.HK	美团-W	(7.3)	19.5
0700.HK	腾讯控股	(35.2)	11.7

数据来源：Wind、开源证券研究所

本周（20260803-0808）恒生沪港通 AH 溢价指数为 122.52，环比略微下降，AH 股溢价或已触底，考虑到该指数成分中金融行业权重超 50%，能源行业权重超 15%，背后或更多反映对红利概念有需求的资金特征，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23：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

5、投资建议

AI：云服务需求与推理算力需求持续增长，云计算高景气延续，国产大模型流量增长带动商业化变现能力提升；AI 智能体产业进入快速增长期，央国企 IT 支出回暖与 AI 国产化双轮驱动行业景气度上行，AI 算力基础设施建设仍处于加速通道，但市场对于巨额资本开支的回报周期存疑。上游硬件因订单确定性强、业绩可验证性高，中长期景气逻辑并未破坏；下游云服务与应用端则因资本开支强度远超收入增速，短期盈利承压引发估值收缩。建议关注充分受益长尾客户需求及提价周期的高弹性 Neocloud，微软 Azure 增速加速，受益标的谷歌、亚马逊、微软、Nebius、CoreWeave、智谱、minimax、百融云-W、迅策、德适-B、金山云。

互联网：关注 AI 商业化及应用场景拓展，随多模型能力持续提升、Agent 技术与应用生态快速成熟，token 消耗量有望延续高速增长，推理需求有望带动云计算行业摆脱低价内卷，供给端 DRAM 价格与产能紧约束，CPU、服务器等成本同步上行，共同推动云计算进入以 AI 算力溢价为核心的新阶段；国产 AI 芯片增长迅速；全球化出海有望由粗放扩张转向精细化盈利，推荐阿里巴巴-W、拼多多，百度集团-SW，受益标的腾讯控股。

6、风险提示

产能及供应链风险：若产能爬坡速度较慢，供给端受限，创收弹性无法充分释放，则相关行业产业链业绩或受到影响。

监管政策变动：科技政策及垂类行业相关政策的变动，或将影响公司增长景气度及战略决策，进一步影响公司的增长节奏。

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。

地缘政治风险：地缘政治变动或伴随国际政策调整，或对科技行业供应链等相关企业带来限制。

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

负责准备本报告以及撰写本报告的所有研究分析师或工作人员在此保证，本研究报告中关于任何发行商或证券所发表的观点均如实反映分析人员的个人观点。负责准备本报告的分析师获取报酬的评判因素包括研究的质量和准确性、客户的反馈、竞争性因素以及开源证券股份有限公司的整体收益。所有研究分析师或工作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”）的机构或个人客户（以下简称“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，如接收人并非开源证券客户，请及时退回并删除。

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。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。

本报告可能附带其它网站的地址或超级链接，对于可能涉及的开源证券网站以外的地址或超级链接，开源证券不对其内容负责。本报告提供这些地址或超级链接的目的纯粹是为了客户使用方便，链接网站的内容不构成本报告的任何部分，客户需自行承担浏览这些网站的费用或风险。

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。除非另有书面显示，否则本报告中的所有材料的版权均属本公司。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