下载日志已记录，仅供内部参考

证券研究报告

2026 年 09 月 22 日

华鑫证券 CHINA FORTUNE SECURITIES

谷歌发布 Gemini 3.8 Live，Step 5 Preview 推出

—计算机行业周报

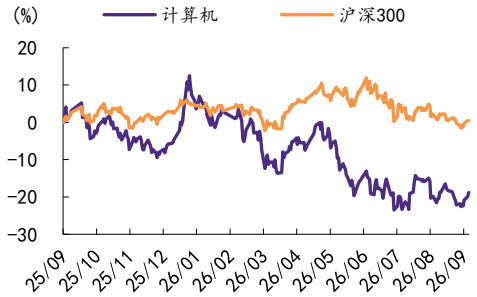
推荐(维持)

分析师：任春阳 S1050521110006
rency@cfsc.com.cn

行业相对表现

表现	1M	3M	12M
计算机(申万)	1.4	-6.5	-20.7
沪深 300	-1.6	-10.2	0.5

市场表现



资料来源：Wind，华鑫证券研究

相关研究

1、《计算机行业周报：DeepSeek 发布 V4.1Flash，GPT-Images-2.5 发布》2026-09-14

2、《计算机行业点评报告：禾赛科技（HSAI）：激光雷达出货量延续高增，机器人业务开启第二增长曲线》2026-09-07

3、《计算机行业点评报告：文远知行（WRD）：海外业务驱动收入高增，规模扩张下亏损持续收窄》2026-09-07

投资要点

算力：阿里发布云端 Agent 托管平台 Qoder，阶跃星辰推出 MoE 旗舰 Step 5 Preview

阿里推出企业级云端 Agent 托管平台 Qoder Cloud Agents 1.0，支持持久化长程运行与弹性调度，月均调用破千万次；阶跃星辰发布 MoE 旗舰 Step 5 Preview（总参数 600B、激活 27B），位居全球开源第三，实测具备高吞吐长程优化能力，API 资费对标 DeepSeek-V4-Pro 高峰期且单任务成本降至 Claude Opus 5 的 1/8，推动大模型加速由单点实验向生产级常态化落地演进。

AI 应用：QuillBot 周访问量环比+6.79%，谷歌发布 Gemini 3.8 Live

2026 年 9 月 16 日，DeepMind 正式推出了 Gemini 3.8 Live 与 Gemini 3.8 Live Extended Thinking。其核心突破在于能够在不干扰用户当前操作流程的前提下，同步完成对话、深度思考以及后台任务处理。

AI 融资动向：Crusoe 完成 39 亿美元 F 轮融资，估值攀升至 309 亿美元

AI 基础设施商 Crusoe 完成 39 亿美元 F 轮融资，估值达 309 亿美元，由 Atreides、Mubadala 与 Valor 联合领投，英伟达等参投。资金投向吉瓦级园区与 Spark 模块化工厂，公司具备电网/清洁能源调度壁垒，在手 Jane Street 130 亿美元订单并服务 OpenAI，加速算力由传统基建向工业标准化交付演进。

投资建议

近期，算力芯片的 HBM 降配动向持续受到关注。据 TrendForce 集邦咨询研究，DRAM 供不应求格局将延续至 2027 年、原厂 HBM4e 验证进度存在变量，NVIDIA 对 RubinUltra 的 HBM 配置已由 HBM4e12Hi 基准设计，改为并行评估 HBM4e8Hi、HBM412Hi、HBM48Hi 等方案，目前尚未定案，若最终调降规格，预计将从调整堆叠层数着手。值得注意的是，本轮调整呈现"代际提升、层数下降"的特征——HBM 自身仍在向 HBM4/HBM4e 演进，降的是单颗堆叠层数。TrendForce 同时指出，部分 CSP 也在考虑调降下一代自研 ASIC 的 HBM 容量；此前 NVIDIA 已因 LPDDR5X 短缺，将 VeraRubinSuperchip 模组的

行业研究

下载日志已记录，仅供内部参考

SOCAMM 容量减半；其预计 2027 年 HBM 出货位元同比增长 50-60%仍不足以满足需求，议价权将偏向供应商，AI 芯片厂商面临供给紧缩与采购成本上升的双重压力。在存储产能持续吃紧的背景下，牺牲容量、保带宽具备一定的产业和经济合理性。

本次产业变化的核心在于，HBM 正从堆叠竞赛转向带宽与容量的再平衡。海外维度，在算力芯片满产满销的背景下，保带宽方案有利于光互连环节；对存储行业而言，需求端无虞，降配有助于提升良率与出货量，且 DRAM 盈利能力强于 HBM，叠加 HBF、内存池化等技术路径，存储环节的重要性不降反增。国产维度，我们认为产业界对 HBM 配置态度的变化，对国产算力板块具有重要的跟踪价值：若海外降配方案落地，HBM 出货总颗粒数提升，部分国产 GPU 公司有望直接受益。整体来看，国产算力板块的利好将更集中地体现在与 HBM 相关公司；若国内开始寻求带宽与容量之间的再平衡，有望成为板块的重要催化。

中长期，建议关注专注于功率半导体及模拟芯片测试系统研发的联动科技（301369.SZ）、专注于半导体等高端制造业的罗博特科（300757.SZ）、新能源业务高增并供货科尔摩根等全球电机巨头的唯科科技（301196.SZ）、AI 智能文字识别与商业大数据领域巨头的合合信息（688615.SH）、深耕工业 AI 与软件并长期服务高端装备等领域头部客户的能科科技（603859.SH）。

风险提示

1）AI 底层技术迭代速度不及预期。2）政策监管及版权风险。3）AI 应用落地效果不及预期。4）推荐公司业绩不及预期风险。

重点关注公司及盈利预测

公司代码	名称	2026-09-22	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	605.88	-0.40	2.13	4.49	-1514.70	284.45	134.94	买入
301196.SZ	唯科科技	90.54	2.05	1.92	2.39	39.28	48.46	38.90	买入
301369.SZ	联动科技	137.23	0.48	1.08	2.62	285.90	127.06	52.38	买入
603859.SH	能科科技	40.40	0.92	1.05	1.32	43.91	38.48	30.61	买入
688615.SH	合合信息	93.49	3.24	4.22	5.25	28.85	22.15	17.81	买入

资料来源：Wind，华鑫证券研究

下载日志已记录，仅供内部参考

正文目录

1、 算力动态：阿里发布云端 AGENT 托管平台 QODER，阶跃星辰推出 MOE 旗舰 STEP 5 PREVIEW 5
1.1、 Tokens 跟踪 5
1.2、 算力租赁动态：租金价格整体平稳 7
1.3、 云厂动态：阿里发布企业级云端 Agent 构建及托管平台 Qoder Cloud Agents 1.0，月调用量突破千万级 9
1.4、 产业动态：阶跃星辰发布 MoE 旗舰模型 Step 5 Preview，跻身全球前三且主打 Agent 长程落地 10
2、 AI 应用动态：QUILLBOT 周访问量环比+6.79%，谷歌发布 GEMINI 3.8 LIVE 12
2.1、 周流量跟踪：QuillBot 周访问量环比+6.79% 12
2.2、 产业动态：谷歌发布 Gemini 3.8 Live，将实时视觉与多语言交互能力融入对话智能体 12
3、 AI 融资动向：CRUSOE 完成 39 亿美元 F 轮融资，估值攀升至 309 亿美元 15
4、 行情复盘 17
5、 投资建议 19
6、 风险提示 20

图表目录

图表 1：TOKENS 规模 LEADERBOARD（TOKEN 消耗量为统计口径） 7
图表 2：文本类模型市场份额占据示意图（REQUEST 调用量为统计口径） 7
图表 3：2026.9.13-2026.9.19 日 H100 价格（单位：美元/小时） 8
图表 4：2026.9.13-2026.9.19 日 H200 价格（单位：美元/小时） 8
图表 5：2026.9.13-2026.9.19 日 A100 价格（单位：美元/小时） 8
图表 6：2026.9.13-2026.9.19 日 B200 价格（单位：美元/小时） 9
图表 7：STEP 5 PREVIEW 与主流前沿模型多维度 BENCHMARK 评测成绩对比 10
图表 8：STEP 5 PREVIEW 与 DEEPSEEK 主流模型 API 资费对比表 11
图表 9：2026.9.11-2026.9.17 AI 相关网站流量 12
图表 10：GEMINI 3.8 LIVE EXTENDED THINKING 在 ARTIFICIAL ANALYSIS 语音质量榜的表现 13
图表 11：GEMINI 3.8 LIVE EXTENDED THINKING 在 SERVICENOW 语音智能体测试中的表现 13
图表 12：GEMINI 3.8 LIVE EXTENDED THINKING 在多个测试中的表现 14
图表 13：上周 AI 初创公司融资动态 15
图表 14：上周（2026.9.14-2026.9.18 日）指数日涨跌幅 17
图表 15：上周（2026.9.14-2026.9.18 日）AI 算力指数内部涨跌幅度排名 17
图表 16：上周（2026.9.14-2026.9.18 日）AI 应用指数内部涨跌幅度排名 18

下载日志已记录，仅供内部参考

图表 17：FICONTEC2025 年年中至今公告订单 19

图表 18：重点关注公司及盈利预测 20

下载日志已记录，仅供内部参考

1、算力动态：阿里发布云端 Agent 托管平台 Qoder，阶跃星辰推出 MoE 旗舰 Step 5 Preview

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 9 月 14 日至 9 月 20 日，周度 Token 消耗量保持平稳增长，消耗量达 129T，环比上周增加 1.57%。在 Tokens 规模 Leaderboard 前五名中，DeepSeek 的 DeepSeek V4.1 Flash 以 15.8T tokens 跃居榜首；Z-ai 的 GLM 5.3 Flash 以 14.1T tokens 位列第二；Tencent 的 Hy4 preview 以 12.5T tokens 位列第三；OpenAI 的 GPT-5.6 Luna 以 9.72T tokens 位居第四；DeepSeek 的 DeepSeek V4 Flash 0731 以 9.44T tokens 位居第五。

文本类模型的市场份额方面，deepseek 以 1.37B requests 占据 25.4%的份额，位列第一；google 以 1.00B requests 占据 18.6%的份额，位列第二；openai、z-ai、qwen 分别以 917M、505M、361M requests，对应占据 17.0%、9.4%、6.7%的市场份额。

图像生成类模型调用量小幅回落，调用量为 5.97M requests，环比上周下降约 4.78%。在 Image Leaderboard 中，Google 旗下 Nano Banana 系列重回榜首，其中 Nano Banana 2 (Gemini 3.1 Flash Image)以 1.1M requests 位居第一；Nano Banana 2 Lite 以 983K requests 位居第二；Nano Banana 以 498K requests 位列第三；OpenAI 的 GPT Image 2.5 Sunburst 以 452K requests 位居第四；Google 的 Nano Banana 2 (Gemini 3.1 Flash Image Preview)以 408K requests 位居第五。从市场份额来看，google 以 3.5M requests 占据 58.7%的份额，稳居第一；openai 以 1.36M requests 占据 22.8%，位列第二；meta、bytedance-seed、x-ai 分别以 273K、269K、191K requests，对应占据 4.6%、4.5%、3.2%的市场份额。

嵌入模型调用量显著增长，调用量为 340M requests，环比上周增加 17.65%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 145M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 59M requests 位列第二；Google 的 Gemini Embedding 2 以 25.4M requests 位列第三；BAAI 的 bge-m3 以 21.7M requests 位列第四；Google 的 Gemini Embedding 001 以 19.4M requests 位列第五。从市场份额来看，openai 以 167M requests 占据 49.1%的份额，位居第一；qwen 以 66.8M requests 占据 19.7%，位列第二；google、baai、perplexity 则分别以 52.4M、22.5M、12.3M requests，对应占据 15.4%、6.6%、3.6%的市场份额。

重排序模型调用量显著回落，调用量为 9.15M requests，环比上周下降 22.46%。在 Rerank Leaderboard 前五名中，VoyageAI 的 rerank-2.5 以 3.48M requests 位居榜首；Cohere 的 Rerank 4 Fast 以 1.76M requests 位居第二；VoyageAI 的 rerank-2.5-lite 以 1.65M requests 位列第三；Cohere 的 Rerank v3.5 以 1.35M requests 位列第四；Qwen 的 Qwen3 Reranker 8B 以 509K requests 位居第五。从市场份额来看，voyageai 以 5.12M requests 占据 56.0%的份额，位列第一；cohere 以 3.43M requests 占据 37.5%的份额，位居第二；qwen、nvidia 分别以 509K、87K requests，对应占据 5.6%、1.0%的市场份额。

视频生成类模型调用量小幅回调，调用量为 181K requests，环比上周下降约 7.18%。在

下载日志已记录，仅供内部参考

Video Leaderboard 前五名中，Google 的 Veo 3.1 Lite 以 35K requests 位居第一；ByteDance 的 Seedance 2.0 Fast 与 Seedance 2.5 均以 19K requests 并列第二；ByteDance 的 Seedance 2.0 Mini 以 12K requests 位列第四；ByteDance 的 Seedance 2.0 以 11K requests 位列第五。从市场份额来看，bytedance 以 67K requests 占据 37.2%的份额，位列第一；google 以 49K requests 占据 27.1%的份额，位居第二；x-ai、alibaba、minimax 则分别以 17K、17K、13K requests，对应占据 9.6%、9.3%、7.0%的市场份额。

语音生成类模型调用量显著回升，调用量为 2.72M requests，环比上周增加 17.24%。在 Speech Leaderboard 前五名中，Hexgrad 的 Kokoro 82M 以 1M requests 升至榜首；Google 的 Gemini 3.1 Flash TTS Preview 以 863K requests 位列第二；xAI 的 Grok Voice TTS 1.0 以 230K requests 位列第三；Fish-audio 的 S2.1 Pro Free (free)与 S2.1 Pro 分别以 107K、106K requests 位列第四和第五。从市场份额来看，hexgrad 以 1M requests 占据 38.1%的份额，跃居第一；google 以 863K requests 占据 31.7%的份额，位居第二；x-ai、fish-audio、qwen 分别以 230K、217K、153K requests，对应占据 8.5%、8.0%、5.6%的市场份额。

语音转写类模型调用量显著增长，调用量为 20.2M requests，环比上周增加 21.69%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3 与 Whisper Large V3 Turbo 分别以 7.99M、5.09M requests 稳居第一和第二位；Microsoft 的 MAI-Transcribe 2 以 2.29M requests 位列第三；OpenAI 的 GPT-4o Mini Transcribe 以 974K requests 位列第四；NVIDIA 的 Parakeet TDT 0.6B v3 以 863K requests 位列第五。从市场份额来看，openai 以 15M requests 占据 74.5%的份额，位居第一；microsoft 以 2.32M requests 占据 11.5%的份额，位列第二；qwen、nvidia、mistralai 分别以 1.04M、890K、309K requests，对应占据 5.2%、4.4%、1.5%的市场份额。

* 统计口径说明：以上数据均来源于 OpenRouter。基于官方统计口径调整，当前 Rankings 采用 Token 与 Request 两套统计维度：其中，仅 Tokens 规模 Leaderboard 仍以模型 Token 消耗量为统计口径，其余类目榜单及市场份额板块则统一采用 Request 调用量为统计口径，两者统计维度不同，不宜直接对应或进行横向比较。

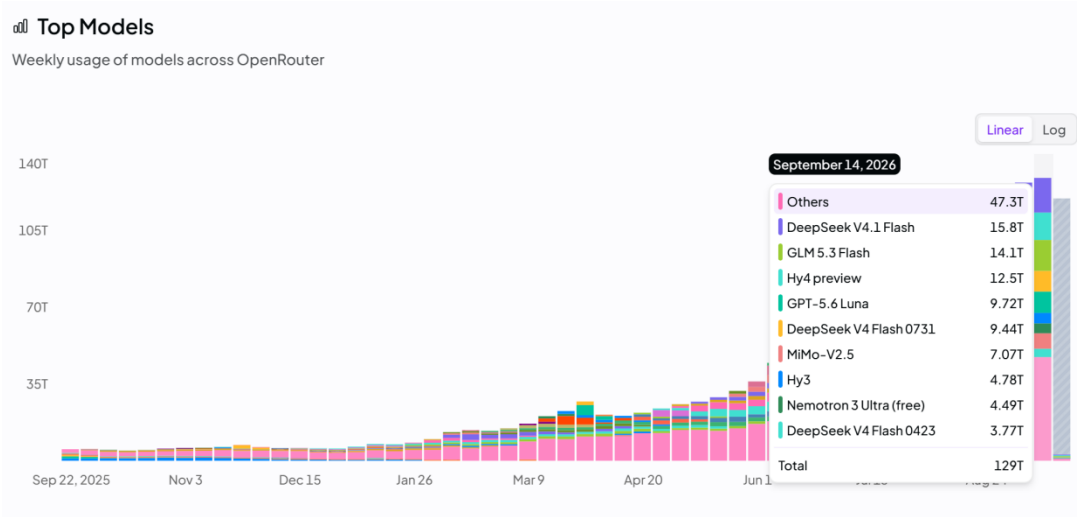
9 月 17 日全国首个面向海洋产业的专属智能要素供给平台“海洋 Token 工厂”在青岛正式发布落地，该平台依托数据、算力、算法三要素构建，已建成 130 多类涉海数据库、超 800P 海洋算力资源池并汇聚 50 余款专有与通用大模型，以“水电厂”式 Token 计量计费模式赋能深海与远洋航运生产。

9 月 20 日国内独立生态 Token 供应商硅基流动（SiliconFlow）宣布先后完成 B+轮二期与 C 轮融资，投资方包括中网投、国新基金及中国移动链长基金等，年内累计融资额近 29 亿元，资金将重点用于推理引擎、异构算力调度研发与 Token 供应平台建设，加速全球市场拓展。

9 月 20 日词元（Token）东盟生态合作推介会在南宁举行，现场集中发布广西自贸试验区数据出境负面清单及跨境流动服务平台等三项服务矩阵成果，中国移动、中国联通与多家跨境数字企业围绕词元出海与 AI 跨境落地签署合作协议，协同打通面向东盟的 Token 产业生态合作链路。

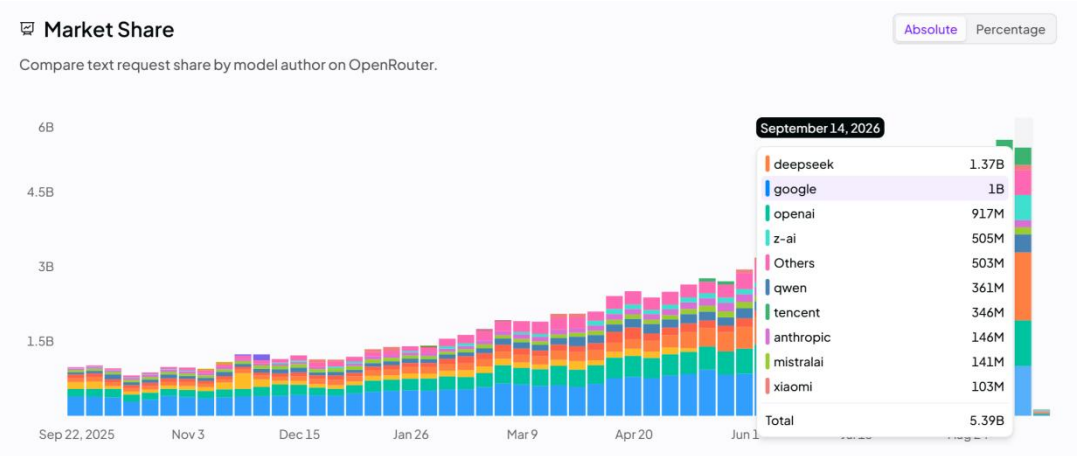
下载日志已记录，仅供内部参考

图表 1：Tokens 规模 Leaderboard（Token 消耗量为统计口径）



资料来源：OpenRouter，华鑫证券研究

图表 2：文本类模型市场份额占据示意图（Request 调用量为统计口径）



资料来源：OpenRouter，华鑫证券研究

1.2、 算力租赁动态：租金价格整体平稳

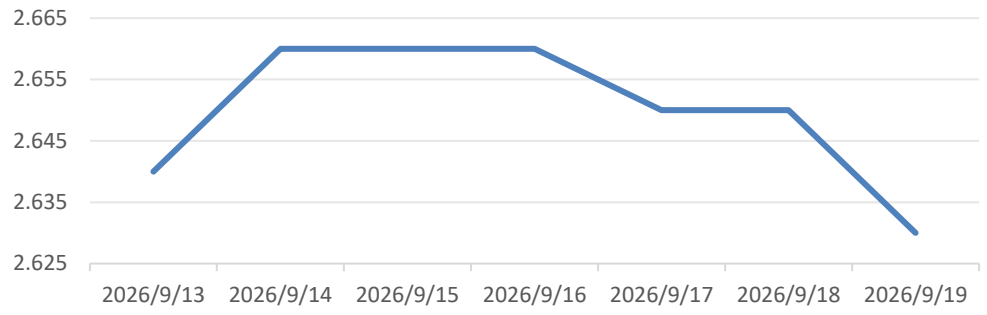
上周（2026.9.13-2026.9.19 日），算力租赁市场整体表现平稳。H100 指数/H200 指数/A100 指数/B200 指数日租金（美元/小时）最大值分别为 2.66/3.32/1.59/5.73，H100 指数/H200 指数/A100 指数/B200 指数日租金（美元/小时）最小值分别为 2.63/3.28/1.58/5.70，未出现明显波动。

* 计算方法说明：以上数据均来源于 silicon data。每条 GPU 租赁记录首先会被转换为标准化的每 GPU 每小时费率——即单价等于总租赁价格除以（GPU 数量乘以租赁时长）。预处理后的数据会经过多阶段过滤，以剔除上报错误，同时不会忽略真实的结构性价格变动。由于单纯汇总价格会随每日配置组合的变化而波动，因此我们采用基差调整（basis adjustment）将每条记录标准化为基准等价合约，在此过程中会综合考虑租赁类型、地理

下载日志已记录，仅供内部参考

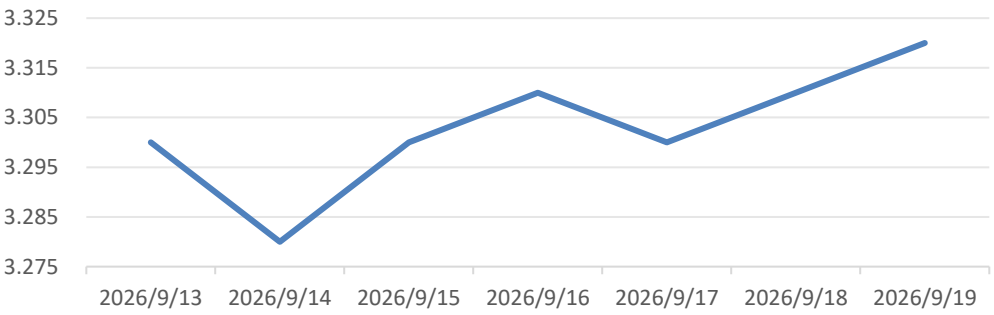
位置、CPU 平台以及 GPU 型号等因素。经基差调整后的价格会在提供商层面进行汇总，最终的指数值是各提供商数据的加权平均值。

图表 3：2026.9.13-2026.9.19 日 H100 价格（单位：美元/小时）



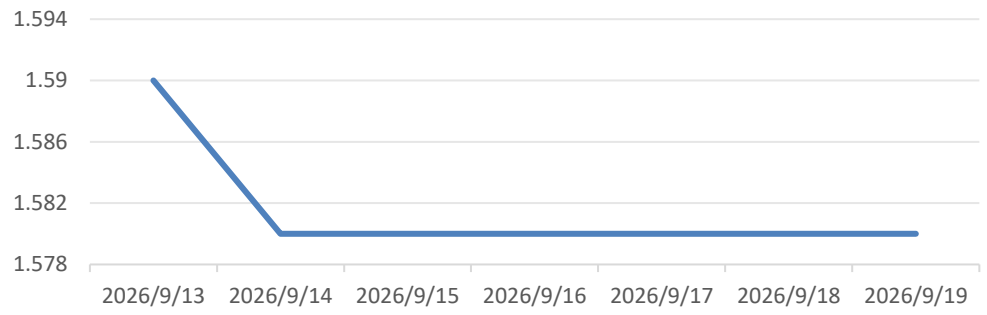
资料来源：silicon data，华鑫证券研究

图表 4：2026.9.13-2026.9.19 日 H200 价格（单位：美元/小时）



资料来源：silicon data，华鑫证券研究

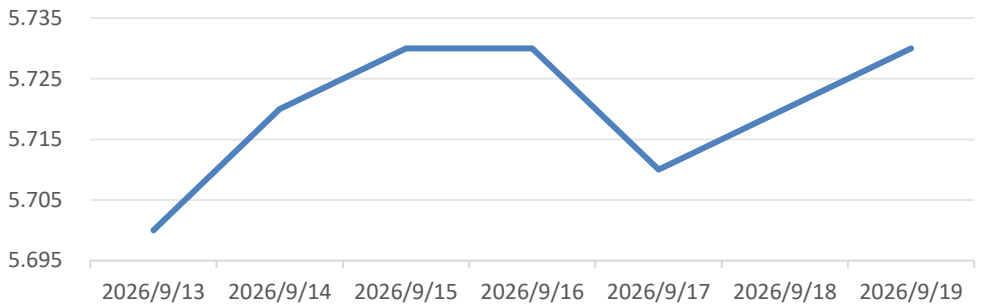
图表 5：2026.9.13-2026.9.19 日 A100 价格（单位：美元/小时）



资料来源：silicon data，华鑫证券研究

下载日志已记录，仅供内部参考

图表 6：2026.9.13-2026.9.19 日 B200 价格（单位：美元/小时）



资料来源：silicon data，华鑫证券研究

1.3、 云厂动态：阿里发布企业级云端 Agent 构建及托管平台 Qoder Cloud Agents 1.0，月调用量突破千万级

2026 年 9 月 16 日，阿里正式发布企业级云端 Agent 构建及托管平台 Qoder Cloud Agents 1.0。今年以来，托管形态的 Agent 服务成为行业热门方向，Anthropic 与 OpenAI 先后推出 Claude Managed Agents 和 Agents API，提供稳定、可控、可审计的 Agent 基础设施；阿里于今年 5 月推出 Qoder Cloud Agents，基于 Qoder 家族统一智能引擎打造，以 Agent as a Service 为核心提供托管运行时，使用户无需从头搭建底层运行系统，10 分钟内即可将业务系统 Agent 接入云端并在验证通过后切入生产环境，实现 7×24 小时稳定运行。

系统架构与企业级特性方面，Qoder Cloud Agents 1.0 提供持久化工作环境与云端基础设施，支持复杂任务连续运行数天及高并发执行，具备文件处理、代码运行与中间结果保存能力，将任务由单一进程生命周期升级为可调度、可恢复、可观测与可扩展的云服务。在交互与接入形态上，平台通过 API 与多语言 SDK 开放底层运行及运维能力，支持主流 AI 编程工具通过 Skill 和插件调用，并支持 Realtime Voice 实时语音以及国内外多种 IM 渠道交互。面向企业治理需求，平台支持通过模板统一配置能力，按用户隔离资源并管理权限、凭证与用量；在大规模并发调度上，通过批量任务、云端调度及多地域资源池承接突发需求，执行环境随任务量动态按需创建与空闲回收，以降低峰值资源预留与等待损耗。

在业务应用示范方面，阿里披露了平台的落地探索案例：以某大型平台 AI 客服团队为例，此前该场景主要依靠人工抽检约 1%的会话，面临全局覆盖不足、质检口径不一及规则调整需修改代码发版等痛点；在试点接入 Qoder Cloud Agents 后，该团队尝试将质检拆解为意图识别、问题归因、复检与总结等节点，并通过夜间闲时批处理机制进行自动回放验证，实现了质检全量覆盖与规则迭代周期的压缩。官方数据显示，Qoder Cloud Agents 已在金融服务、企业软件、零售、电商等领域对接数百家企业客户，平台月均调用次数达到千万级。

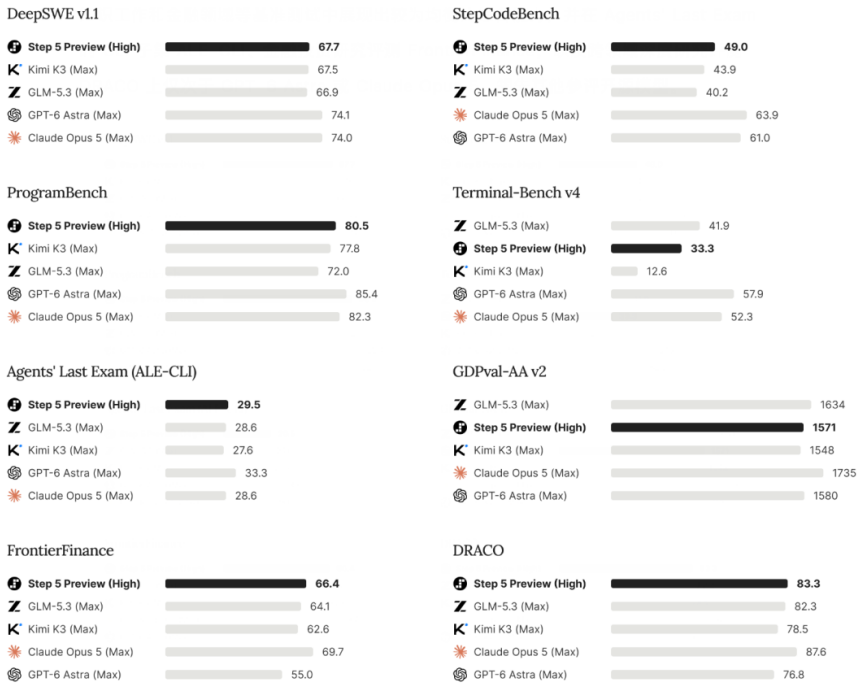
下载日志已记录，仅供内部参考

1.4、产业动态：阶跃星辰发布 MoE 旗舰模型 Step 5 Preview，跻身全球前三且主打 Agent 长程落地

2026 年 9 月 20 日，阶跃星辰正式推出新一代 MoE 旗舰大模型 Step 5 Preview，并宣布全量开放 API 及 Studio 平台在线体验，预计将于 10 月 15 日正式开源。该模型专为真实世界 Agentic 任务打造，总参数量达 600B，激活参数仅 27B，支持 100 万 Token 上下文窗口，原生支持文本与视觉混合输入。在全球 AI 权威榜单 Artificial Analysis Intelligence Index 中，Step 5 Preview 获得 44 分，位居全球开源模型第三位，仅次于 Qwen3.8 Max（0902）与 GLM-5.3（max）；输出性能方面，其解码速度达到 100 token/秒，在 Artificial Analysis 榜单中排名第 6 位。

基准测试与性能表现方面，根据阶跃公开数据，Step 5 Preview 在 Agent 长周期 CLI 任务测试、金融投资研究评测 FrontierFinance，以及跨领域深度研究评测 DRACO 等测试中全面优于 Kimi K3 与 GLM-5.3，综合表现略逊色于 GPT-6 Astra 或 Claude Opus 5。在金融领域，模型在内部构建的 FinStepBench 三项测试中表现突出，信息检索、估值建模及完整投研流程得分较高，FrontierFinance 得分仅次于 Claude Opus 5。在真实软件工程场景下，阶跃内部构建了覆盖 553 个独立代码仓库、9 类任务、20 个应用领域和 33 种编程语言的测试集 StepCodeBench，Step 5 Preview 在任务成功率与跨场景稳定性上取得 49 分的综合得分，超过 Kimi K3 和 GLM-5.3。此外，在获得开发文档和授权后，模型支持连续工作超 3 小时，调用相机、COM 端口、截图并模拟鼠标输入，自主阅读 API 文档将 ESP32-S3 开发板改造成支持蓝牙按键与语音输入的 Vibe Coding 键盘。

图表 7：Step 5 Preview 与主流前沿模型多维度 Benchmark 评测成绩对比



资料来源：智东西，华鑫证券研究

下载日志已记录，仅供内部参考

长程任务与业务落地方面，Step 5 Preview 在 24 小时内自主优化 MLA GPU 内核，历经约 22 小时将单张 NVIDIA H100 上的峰值性能提升至 508 TFLOPS，超越 Claude Opus 5 的 493 TFLOPS；在自动化后训练优化中，24 小时内将 Qwen3-30B-A3B 在 AIME24 测试集上的准确率由 53.3%提升至 60%，与 Claude Opus 5 持平且消耗更少 Token；在未经专项优化下，于《Pokémon Red》游戏中持续完成超 3000 轮交互（累计近 600 万 Tokens），验证了跨周期长目标规划与资源管理能力。在多模态业务实测中，模型支持文档与博客一键生成结构化 PPT、基于 Three.js 生成带交互控件的 3D 场景网页，以及在长达 7 轮的茶饮创业模拟中动态根据低成本偏好推进决策并直出交互式复盘产品。

商业化定价与调用成本方面，Step 5 Preview 在 API 服务中每百万 Token 输入价格为 7 元（缓存未命中）、缓存命中价格为 0.35 元，输出价格为 20 元，定价水平对标 DeepSeek-V4-Pro-0813 的高峰时段。在综合使用成本上，阶跃官方数据显示 Step 5 Preview 的单任务成本为 Claude Opus 5（max）的 1/8。随着模型规模与任务复杂度增长，大模型竞争正从原始算力堆叠转向算力转化率之争，通过高稀疏架构以更少激活计算承接复杂长程任务已成为适配 Agent 时代产业落地的核心命题。

图表 8：Step 5 Preview 与 DeepSeek 主流模型 API 资费对比表

阶跃新模型与DeepSeek模型价格对比（单位：元）

模型	百万tokens输入（缓存命中）		百万tokens输入（缓存未命中）		百万tokens输出	
	空闲时段	高峰时段	空闲时段	高峰时段	空闲时段	高峰时段
Step 5 Preview	**0.35**		**7**		**20**	
DeepSeek-V4.1-Flash	0.02	0.04	1	2	4	8
DeepSeek-V4-Pro-0813	0.15	0.30	4.5	9	13.5	27

资料来源：智东西，华鑫证券研究

下载日志已记录，仅供内部参考

2、AI 应用动态：QuillBot 周访问量环比+6.79%，谷歌发布 Gemini 3.8 Live

2.1、周流量跟踪：QuillBot 周访问量环比+6.79%

本期（2026.9.11-2026.9.17）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1379.0M）、Bing（835.2M）和 Gemini（615.7M），访问量环比增速第一为 QuillBot（6.79%）；平均停留时长前三位分别为 Character.AI（00:15:39）、Discord（00:10:51）和 NotionAI（00:08:52）;平均停留时长环比增速第一为 NotionAI（0.76%）。

图表 9：2026.9.11-2026.9.17 AI 相关网站流量

应用	应用类型	归属公司	周平均访问量（M）	访问量环比	平均停留时长	时长环比
ChatGPT	聊天机器人	OpenAI	1379.0	-0.14%	6:46	0.25%
Bing	搜索	微软	835.2	-0.33%	7:04	-0.47%
Gemini	聊天机器人	谷歌	615.7	0.08%	6:39	-1.48%
Canva	在线设计	Canva	221	2.46%	5:48	-1.14%
Github	代码托管	微软	149.2	0.20%	6:27	0.00%
Discord	游戏社区	微软	143.6	-0.28%	10:51	-0.61%
Character.AI	聊天机器人	Character.AI	35.17	-0.71%	15:39	-0.21%
DeepI	翻译工具	DeepL	22.78	0.75%	2:25	-1.36%
Perplexity	AI 搜索	Perplexity	21.61	0.23%	4:21	0.00%
QuillBot	释义工具	QuillBot	8.587	6.79%	2:48	0.00%
Kimi	聊天机器人	Moonshot AI	6.029	-12.09%	5:50	-1.96%
NotionAI	文本/笔记	Notion	3.17	3.52%	8:52	0.76%
文心一言	聊天机器人	百度	0.00	6.25%	0:22	-8.33%

资料来源：similarweb,华鑫证券研究

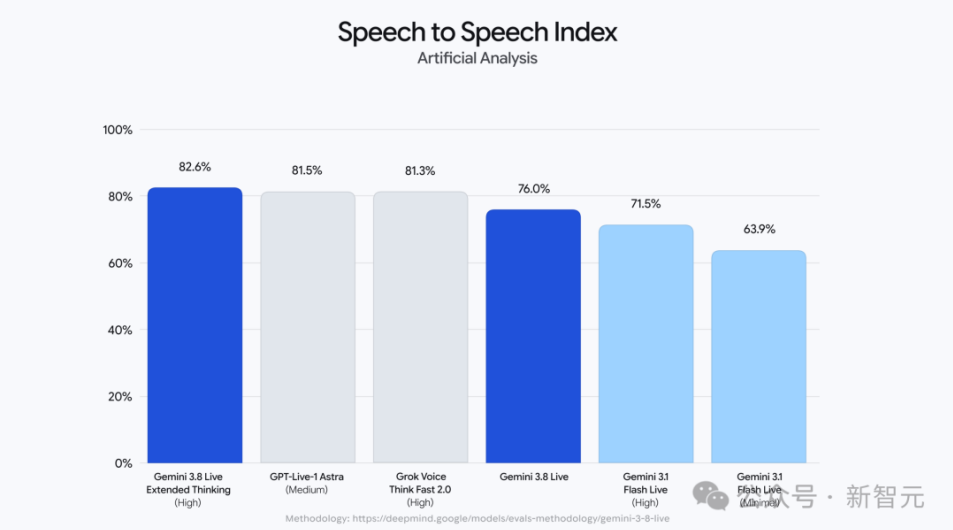
2.2、产业动态：谷歌发布 Gemini 3.8 Live，将实时视觉与多语言交互能力融入对话智能体

2026 年 9 月 16 日，DeepMind 正式推出了 Gemini 3.8 Live 与 Gemini 3.8 Live Extended Thinking。其核心突破在于能够在不干扰用户当前操作流程的前提下，同步完成对话、深度思考以及后台任务处理。

下载日志已记录，仅供内部参考

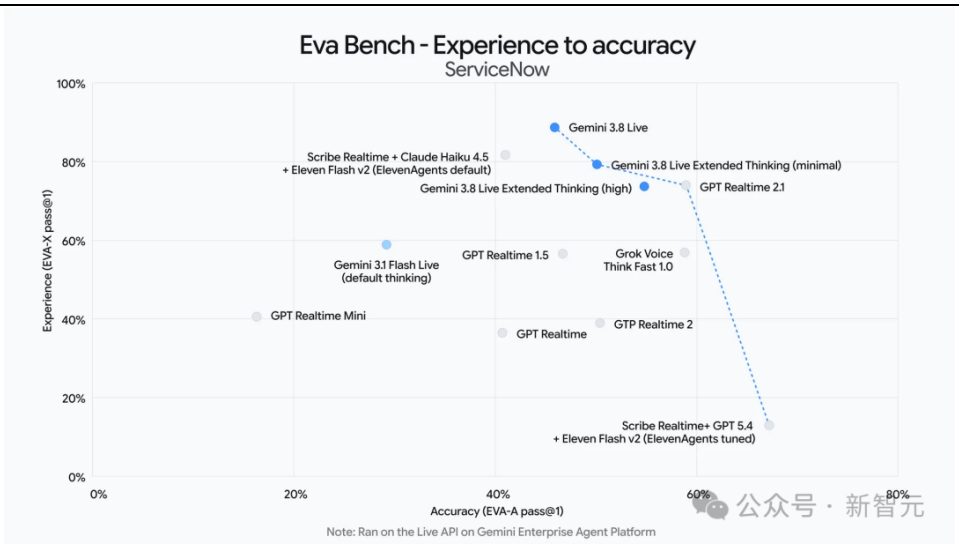
在技术性能的评估上，Gemini 3.8 Live Extended Thinking 在 Artificial Analysis 的语音质量排行榜中，以 82.6 分的成绩位列总榜第一，展现出卓越的语音交互能力。而在 ServiceNow 进行的语音智能体测试中，该模型也解决了以往同类产品面临的痛点——过去的模型往往难以兼顾回答的准确性与对话的自然流畅度。Gemini 3.8 Live 与 Gemini 3.8 Live Extended Thinking 模型首次在“准确”与“流畅”两个维度上实现了同步提升。

图表 10：Gemini 3.8 Live Extended Thinking 在 Artificial Analysis 语音质量榜的表现



资料来源：新智元，华鑫证券研究

图表 11：Gemini 3.8 Live Extended Thinking 在 ServiceNow 语音智能体测试中的表现



资料来源：新智元，华鑫证券研究

从产品定位来看，这两款模型分工明确。Gemini 3.8 Live 主要面向规模化与成本效率，将对话智能、流畅度与视觉理解能力整合在一起，适合大规模部署；而 Extended Thinking 则专注于高复杂度任务，具备更强的智能水平与多步推理能力。谷歌将这两者定位为可投入生产环境的语音智能体基础组件。在交互体验上，Extended Thinking 解决了传统语音助手在遇到复杂问题时出现的“沉默等待”问题。当需要多步推理时，模型会先用一句“Let me check that…”来承接用户，并在推理过程中实时同步进度，避免了用户因等待过久而重复提问或打断对话。此外，模型还支持在后台处理工具调用与 API 请求，可以先应下用户

下载日志已记录，仅供内部参考

请求并继续聊天。

在多语言与视觉交互方面，Gemini 3.8 系列展现了强大的自适应能力。模型支持多达 97 种语言，能够在对话中途自动检测并切换语种，无需用户提前设置，实现了无缝的同声传译体验。同时，近实时的视觉输入能力使得摄像头捕捉到的画面几乎同步进入对话，用户无需再经历截图、上传、等待描述的繁琐步骤。在谷歌的演示中，Extended Thinking 能够作为编程家教，实时观察屏幕上的代码，当用户指着一行代码询问报错原因时，模型能够准确识别所指位置，将过去的三步操作简化为一句自然的对话。

在具体的性能指标上，除了语音质量的 82.6 分，Extended Thinking 在任务执行能力上也表现出色。在 τ-Voice 测试中，其得分为 68.6%，小幅领先于 GPT-Live-1Astra 的 67.9%；在 Sierra 的 τ-Voice-banking 中，得分为 35.1%，领先于 GPT-Live-1Astra 的 32.0%；在 Big Bench Audio 测试中则达到了 97.7%的高分。而标准版 Gemini 3.8 Live 则更侧重于成本控制，在真人盲听投票中排名第二，且谷歌表示其价格具有竞争力。

图表 12：Gemini 3.8 Live Extended Thinking 在多个测试中的表现



资料来源：新智元，华鑫证券研究

下载日志已记录，仅供内部参考

3、AI 融资动向：Crusoe 完成 39 亿美元 F 轮融资，估值攀升至 309 亿美元

AI 基础设施服务商 Crusoe 宣布完成 39 亿美元 F 轮融资，企业估值从 10 个月前的 100 亿美元跳涨至 309 亿美元。本轮融资由 Atreides Management、Mubadala Capital 与 Valor Equity Partners 联合领投，参投方涵盖英伟达、GIC、卡塔尔投资局及 Founders Fund 等多家主权基金、科技巨头与基建资本；同时公司引入 Cloudflare 前 CFO、特斯拉董事 JB·斯特劳贝尔等三名新董事补全董事会，在为 IPO 铺路的同时强化能源、财务与基建决策能力。资金供给与战略方向方面，本轮巨额资金将同时投入吉瓦级大型算力园区与可移动 Spark 模块化 AI 工厂的建设，聚焦破解 AI 算力扩张中面临的电力供应瓶颈、数年级漫长建设周期以及社区邻避阻力等行业痛点。

业务模式与产品演进方面，Crusoe 成立于 2018 年，联合创始人蔡斯·洛克米勒具备 MIT 数理、斯坦福 AI 背景及量化交易从业经历，联合创始人 Cully Cavness 深耕能源经济领域。公司早期以集装箱式算力设施消化油气田放空燃烧的废弃天然气，起点为加密货币挖矿业务；在红利褪去后，团队依托沉淀的能源调度与模块化建造经验全面剥离挖矿业务，转向 AI 基础设施服务。针对传统超大型数据中心征地耗时、配套协调周期长及周边居民抵制等现实挑战，Crusoe 自研了 Spark 模块化 AI 工厂，单单元体积接近集装箱、具备约 1 兆瓦算力，全部在自有工厂完成冷却、消防、GPU 机架及监控系统的预制集成，通过卡车运输抵达现场接电即可投入运行，将部署周期压缩至数月级别；该模块化单元适配光伏、储能、电网及废弃油气等多元供电，既能单独作为边缘推理节点，亦可灵活集群堆叠，与服务 OpenAI 的得州阿比林吉瓦级训练大园区形成“巨型园区承担训练、Spark 模块化单元承担推理与边缘计算”的业务互补格局，推动数据中心由传统建筑工程向标准化工业制造品演进。商业化落地方面，Crusoe 手握 Jane Street 高达 130 亿美元的长期云服务订单，客户群体涵盖 OpenAI、Meta 与微软，已打通出租机房空间、租赁 GPU 硬件以及直接售卖推理算力服务三条核心商业化变现路径。

图表 13：上周 AI 初创公司融资动态

应用	应用类型	领投方	融资轮	融资额	目前累计融资额	目前估值
Crusoe	云计算提供商及数据中心开发商	Atreides Management、Mubadala Capital、Valor Equity Partners	F 轮	39 亿美元	未披露	309 亿美元
Temporal	开源分布式工作流平台	Lightspeed & Wellington	E 轮	5.5 亿美元	12 亿美元	125.5 亿美元

下载日志已记录，仅供内部参考

地瓜机器人	具身智能与机器人软硬件芯片计算平台	未来资产	C 轮	4 亿美元	超 7.7 亿美元	未披露

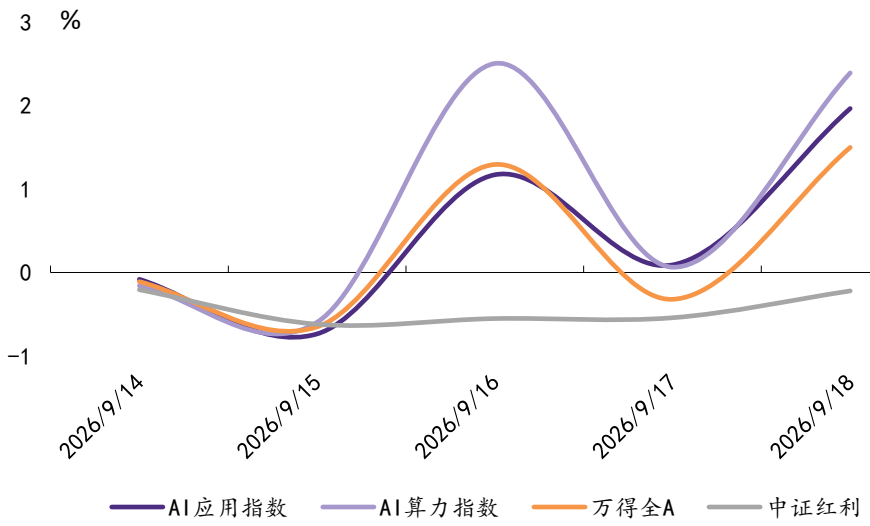
资料来源：具身涌现，Saasverse，Geekwire，深圳新闻网，经济观察网，华鑫证券研究

下载日志已记录，仅供内部参考

4、 行情复盘

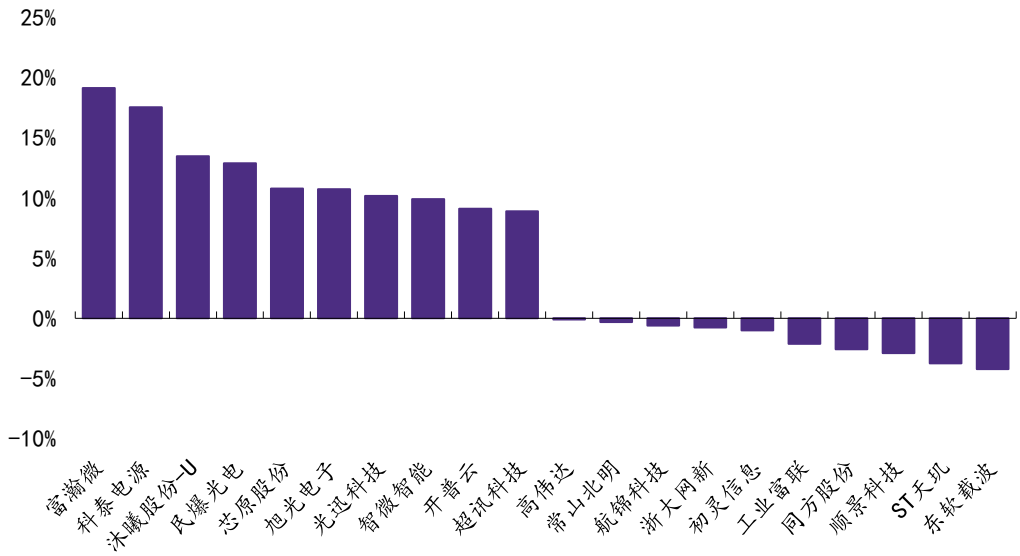
上周（2026.9.14-2026.9.18 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 1.96%/2.5%/1.5%/-0.21%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-0.74%/-0.6%/-0.65%/-0.62%。AI 算力指数内部，富瀚微以 19.12%录得上周最大涨幅，东软载波以-4.21%录得上周最大跌幅。AI 应用指数内部，中岩大地以 28.53%录得上周最大涨幅，众信旅游以-6.41%录得上周最大跌幅。

图表 14：上周（2026.9.14-2026.9.18 日）指数日涨跌幅



资料来源：wind，华鑫证券研究

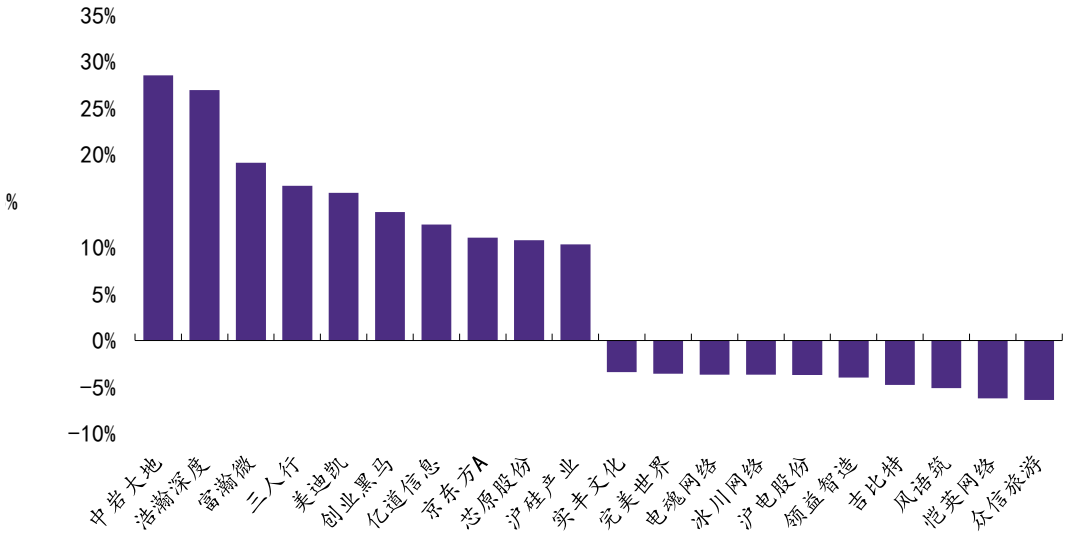
图表 15：上周（2026.9.14-2026.9.18 日）AI 算力指数内部涨跌幅度排名



资料来源：wind，华鑫证券研究

下载日志已记录，仅供内部参考

图表 16：上周（2026.9.14-2026.9.18 日）AI 应用指数内部涨跌幅度排名



资料来源：wind,华鑫证券研究

下载日志已记录，仅供内部参考

5、 投资建议

近期，算力芯片的 HBM 降配动向持续受到关注。据 TrendForce 集邦咨询研究，DRAM 供不应求格局将延续至 2027 年、原厂 HBM4e 验证进度存在变量，NVIDIA 对 RubinUltra 的 HBM 配置已由 HBM4e12Hi 基准设计，改为并行评估 HBM4e8Hi、HBM412Hi、HBM48Hi 等方案，目前尚未定案，若最终调降规格，预计将从调整堆叠层数着手。值得注意的是，本轮调整呈现"代际提升、层数下降"的特征——HBM 自身仍在向 HBM4/HBM4e 演进，降的是单颗堆叠层数。TrendForce 同时指出，部分 CSP 也在考虑调降下一代自研 ASIC 的 HBM 容量；此前 NVIDIA 已因 LPDDR5X 短缺，将 VeraRubinSuperchip 模组的 SOCAMM 容量减半；其预计 2027 年 HBM 出货位元同比增长 50－60%仍不足以满足需求，议价权将偏向供应商，AI 芯片厂商面临供给紧缩与采购成本上升的双重压力。在存储产能持续吃紧的背景下，牺牲容量、保带宽具备一定的产业和经济合理性。

本次产业变化的核心在于，HBM 正从堆叠竞赛转向带宽与容量的再平衡。海外维度，在算力芯片满产满销的背景下，保带宽方案有利于光互连环节；对存储行业而言，需求端无虞，降配有助于提升良率与出货量，且 DRAM 盈利能力强于 HBM，叠加 HBF、内存池化等技术路径，存储环节的重要性不降反增。国产维度，我们认为产业界对 HBM 配置态度的变化，对国产算力板块具有重要的跟踪价值：若海外降配方案落地，HBM 出货总颗粒数提升，部分国产 GPU 公司有望直接受益。整体来看，国产算力板块的利好将更集中地体现于 GPU 与 HBM 相关公司；若国内开始寻求带宽与容量之间的再平衡，有望成为板块的重要催化。

中长期，建议关注专注于功率半导体及模拟芯片测试系统研发的联动科技（301369.SZ）、专注于半导体等高端制造业的罗博特科（300757.SZ）、新能源业务高增并供货科尔摩根等全球电机巨头的唯科科技（301196.SZ）、AI 智能文字识别与商业大数据领域巨头的合合信息（688615.SH）、深耕工业 AI 与软件并长期服务高端装备等领域头部客户的能科科技（603859.SH）。

图表 17：ficonTEC2025 年年中至今公告订单

签约日期	客户/描述	业务类型	金额	折合人民币
2025/6/20	美国某头部公司 A 及其子公司	光电子封测设备	约 1,710 万欧元	约 1.36 亿元
2025/7/11	美国某头部公司 B 及其子公司	光电子封测设备	约 1,418 万美元	约 0.98 亿元
2025/9/3	瑞士某头部公司 C 的子公司	全自动硅光子封装整线设备或服务	约 946.50 万欧元	约 0.75 亿元
2025/10/21	武汉驿路通科技股份有限公司	光纤预制及组装线相关自动化设备	约 900 万美元	约 0.62 亿元
2026/1/6	瑞士某头部公司 C 的子公司	第二条全自动 OCS（光交换机）封装整线设备及服务	约 770.00 万欧元	约 0.61 亿元
2025/9/24-2026/1/26	以色列的纳斯达克上市的头部公司 E	单面晶圆测试设备及服务	约 921.60 万美元	约 0.64 亿元

下载日志已记录，仅供内部参考

2026/3/13	暂未披露	双面晶圆测试设备及服务	约 608.09 万欧元	约 0.48 亿元
2026/3/19-2026/3/25	纳斯达克上市的公司 F 及其子公司	耦合设备及服务（可用于可插拔硅光高速光模块封装制程核心环节的量产）	约 6 亿元人民币	约 6 亿元
2026/4/1	纳斯达克上市的公司 F	耦合设备及服务（可用于可插拔硅光高速光模块封装制程核心环节的量产）	约 3,570 万美元	约 2.46 亿元
2026/4/8-2026/5/1	纽约证券交易所上市的公司 B 的子公司	耦合设备及相关服务	约 2680 万美元	约 1.83 亿元
2026/4/8-2026/5/1	纳斯达克上市的公司 F	视觉检测设备、高精度激光 bar 条封装设备及相关服务	约 3226 万美元	约 2.20 亿
2026/7/23	公司 H 的子公司	光学元器件量产的自动化制造设备及服务	约 1674 万欧元	约 1.29 亿
	总金额			约 19.22 亿元

资料来源：Wind，公司公告，华鑫证券研究

图表 18：重点关注公司及盈利预测

公司代码	名称	2026-09-22	EPS			PE			投资评级
		股价	2025		2027E	2025	2026E	2027E	
300757.SZ	罗博特科	605.88	-0.40	2.13	4.49	-1514.70	284.45	134.94	买入
301196.SZ	唯科科技	90.54	2.05	1.92	2.39	39.28	48.46	38.90	买入
301369.SZ	联动科技	137.23	0.48	1.08	2.62	285.90	127.06	52.38	买入
603859.SH	能科科技	40.40	0.92	1.05	1.32	43.91	38.48	30.61	买入
688615.SH	合合信息	93.49	3.24	4.22	5.25	28.85	22.15	17.81	买入

资料来源：Wind，华鑫证券研究

6、 风险提示

1）AI 底层技术迭代速度不及预期。2）政策监管及版权风险。3）AI 应用落地效果不及预期。4）推荐公司业绩不及预期风险。

下载日志已记录，仅供内部参考

中小盘&北交所组介绍

任春阳：华东师范大学经济学硕士，6 年证券行业经验，2021 年 11 月加盟华鑫证券研究所，从事计算机与中小盘行业上市公司研究

周文龙：澳大利亚莫纳什大学金融硕士

何春玉：金融学士、理学硕士，2023 年 8 月加盟华鑫证券研究所。

证券分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

证券投资评级说明

股票投资评级说明：

	投资建议	预测个股相对同期证券市场代表性指数涨幅
1	买入	>20%
2	增持	10%—20%
3	中性	-10%—10%
4	卖出	<-10%

行业投资评级说明：

	投资建议	行业指数相对同期证券市场代表性指数涨幅
1	推荐	>10%
2	中性	-10%—10%
3	回避	<-10%

以报告日后的 12 个月内，预测个股或行业指数相对于相关证券市场主要指数的涨跌幅为标准。

相关证券市场代表性指数说明：A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以道琼斯指数为基准。

下载日志已记录，仅供内部参考

免责条款

华鑫证券有限责任公司（以下简称“华鑫证券”）具有中国证监会核准的证券投资咨询业务资格。本报告由华鑫证券制作，仅供华鑫证券的客户使用。本公司不会因接收人收到本报告而视其为客户。

本报告中的信息均来源于公开资料，华鑫证券研究部门及相关研究人员力求准确可靠，但对这些信息的准确性及完整性不作任何保证。我们已力求报告内容客观、公正，但报告中的信息与所表达的观点不构成所述证券买卖的出价或询价的依据，该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。投资者应当对本报告中的信息和意见进行独立评估，并应同时结合各自的投资目的、财务状况和特定需求，必要时就财务、法律、商业、税收等方面咨询专业顾问的意见。对依据或者使用本报告所造成的一切后果，华鑫证券及/或其关联人员均不承担任何法律责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露。

本报告中的资料、意见、预测均只反映报告初次发布时的判断，可能会随时调整。该等意见、评估及预测无需通知即可随时更改。在不同时期，华鑫证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。华鑫证券没有将此意见及建议向报告所有接收者进行更新的义务。

本报告版权仅为华鑫证券所有，未经华鑫证券书面授权，任何机构和个人不得以任何形式刊载、翻版、复制、发布、转发或引用本报告的任何部分。若华鑫证券以外的机构向其客户发放本报告，则由该机构独自为此发送行为负责，华鑫证券对此等行为不承担任何责任。本报告同时不构成华鑫证券向发送本报告的机构之客户提供的投资建议。如未经华鑫证券授权，私自转载或者转发本报告，所引起的一切后果及法律责任由私自转载或转发者承担。华鑫证券将保留随时追究其法律责任的权利。请投资者慎重使用未经授权刊载或者转发的华鑫证券研究报告。

