仅供内部参考，请勿外传！

证券研究报告

2026 年 07 月 21 日

华鑫证券 CHINA FORTUNE SECURITIES

月之暗面正式发布 Kimi K3，Step AOS 与 STEPX Neo 发布

—计算机行业周报

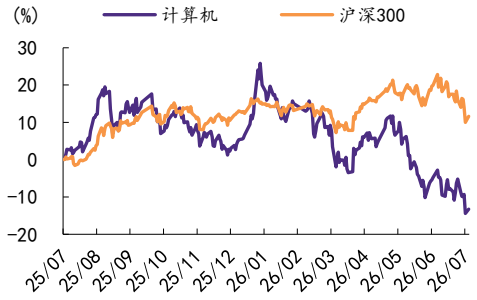
推荐(维持)

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)	-8.9	-17.5	-13.9
沪深 300	-6.9	-3.6	12.5

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：OpenAI 正式发布 GPT-5.6 系列大模型，SpaceXAI 发布 Grok4.5》2026-07-15

2、《计算机行业周报：ClaudeSonnet5 发布，DSpark 与 JetSpec 共探投机解码效率新边界》2026-07-06

3、《计算机行业周报：火山引擎发布豆包 2.1 系列大模型，ClaudeTag 开启人机协作新范式》2026-07-02

投资要点

算力：算力租赁价格平稳，月之暗面正式发布 Kimi K3

2026 年 7 月 17 日，月之暗面正式推出新一代大模型 Kimi K3。该模型总参数达 2.8 万亿，拥有 100 万 token 上下文窗口，是目前全球参数规模最大的开源模型。架构上引入 KDA 与 AttnRes 两项创新，训练效率较上一代提升约 2.5 倍。性能上，该模型在编程、知识工作等多项评测中均表现优异，并展现了 Kernel 优化、芯片设计、科研编程等多领域自主任务完成能力。定价上对标 Sonnet，价格为每百万 Token 输入：2 元（命中缓存）和 20 元（未命中缓存），每百万 Token 输出：100 元。当前，该模型已在 Kimi 网页版、Kimi App、Kimi Work 以及官方 API 上线，完整权重将于 7 月 27 日前发布。

AI 应用：Kimi 周访问量环比+29.10%，Step AOS 与 STEPX Neo 发布

2026 年 7 月 13 日，阶跃星辰展示了三款 AI 终端产品：全球首个大模型原生 AI 终端品牌 STEPX、全球首个智能体原生操作系统 Step AOS，以及首款大模型原生智能体手机 STEPX Neo，反映出行业竞争焦点正从单纯的模型性能比拼演进至“模型-软件-硬件”三位一体的全栈生态较量。

AI 融资动向：Fireworks AI 完成 15.05 亿美元 D 轮融资，投后估值达 175 亿美元

2026 年 7 月 17 日，AI 基础设施公司 Fireworks AI 宣布完成 15.05 亿美元 D 轮融资，投后估值达 175 亿美元。本轮融资由 Atreides Management、Index Ventures、TCV 联合领投，Evantic、Lightspeed Venture Partners、NVIDIA、20VC、Bessemer Venture Partners 等多家机构参投。该公司以专属智能 AI 平台作为自身核心定位，通过面向开发者的云平台产品提供开源 AI 模型微调、定制化部署及托管 GPU 集群访问服务。当前，该平台已上线超过 200 个模型，每日 Token 处理量超过 40 万亿，核心客户涵盖 Uber、Shopify 等全球头部企业，同时公司年化营收已突破 10 亿美元，较上一轮融资阶段增长超过 5 倍。

投资建议

行业研究

仅供内部参考，请勿外传！

2026 年 7 月 14 日，TowerSemiconductor（高塔半导体）宣布在日本政府支持下启动双轨战略扩产，重点扩大 300mm 硅光子、硅锗及先进封装产能，项目总投资约 40 亿美元，其中日本政府补助约 10 亿美元、公司自有投资约 30 亿美元。第一轨将荒井工厂（原 Fab6）改造为 300mm 硅光子产线及先进封装产能，并最大化鱼津 Fab7 产能输出，预计 2027Q4 实现全面量产就绪；第二轨同步启动，待协议签署完成后于鱼津 Fab7 旁新建 300mm 厂房，达产后硅光子与硅锗产能将提升数倍，主要承接 AI 与数据中心驱动的下一代光连接需求，预计 2029 年起显著增厚业绩。基于扩产计划，公司同步上调 2028 年财务目标：预计营收 36 亿美元（高于市场普遍预期的 31.7 亿美元，公司 2025 年营收为 16 亿美元），净利润 12 亿美元（公司 2025 年净利润为 2.2 亿美元），且仅依靠第一轨即可达成。此外，公司年内资本开支已明显提速：2026 年初原定资本开支 6.5 亿美元，2 月于财报说明会上宣布追加 2.7 亿美元，合计 9.2 亿美元投向硅光设备采购，计划 2026 年底完成全部设备进场认证，以支撑 2027 年硅片出货量达到 2025Q4 月均出货量的 5 倍以上。

本次双轨扩产印证了硅光晶圆的结构性短缺，上游设备环节将成为公司资本开支传导下的核心受益方向。需求端看，即使产能扩张至 5 倍以上，公司硅光供给仍然紧张：截至 2028 年的硅光总产能中超过 70%已被预订或正在预订，且有客户预付款作为保障；今年 5 月公司更与头部硅光客户签订 2027 年 13 亿美元长期供货协议，已收到 2.9 亿美元预付款。这表明短缺是结构性的、而非周期性的，产能扩张具备持续性——今年 2 月的追加投资系为缓解供给紧张而实施的阶段性产能补充，本次的扩产则为中长期战略布局，二者在规划属性上存在本质区别。从传导路径看，晶圆厂扩产的资本开支将直接转化为设备采购需求：公司 9.2 亿美元硅光/硅锗投资以设备采购为主，本次 40 亿美元双轨投资亦以产线改造和新厂建设为主。Tower 作为全球硅光代工核心产能，其大规模扩产、产能被长约锁定的事实，印证了光通信需求的高景气与持续性；晶圆厂资本开支将直接转化为设备采购需求，上游硅光晶圆测试、耦合封装等高精度设备环节有望率先且持续受益于订单放量。

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）。

仅供内部参考，请勿外传！

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	2026-07-21	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	373.10	-0.30	0.30	0.60	-1243.67	1243.67	621.83	买入
301196.SZ	唯科科技	75.33	2.53	3.34	3.98	29.77	22.55	18.93	买入
301369.SZ	联动科技	129.80	0.48	1.08	2.62	270.42	120.19	49.54	买入
603859.SH	能科科技	35.32	0.92	1.21	1.50	38.39	29.19	23.55	买入
688615.SH	合合信息	97.72	3.24	4.22	5.25	30.16	23.16	18.61	买入

资料来源：Wind，华鑫证券研究

仅供内部参考，请勿外传！

正文目录

1、 算力动态：算力租赁价格平稳，月之暗面正式发布 KIMI K3 6

1.1、 Tokens 跟踪 6

1.2、 数据跟踪：阿里云发布实时语音交互模型 Qwen-Audio-3.0-Realtime，HappyOyster 1.0 开启灰度测试 8

1.3、 产业动态：月之暗面正式发布 Kimi K3 9

2、 AI 应用动态：KIMI 周访问量环比+29.10%，STEP AOS 与 STEPX NEO 发布 13

2.1、 周流量跟踪：Kimi 周访问量环比+29.10% 13

2.2、 产业动态：Step AOS 与 STEPX Neo 发布，开启人机共生的终端新纪元 13

3、 AI 融资动向：FIREWORKS AI 完成 15.05 亿美元 D 轮融资，投后估值达 175 亿美元 16

4、 行情复盘 18

5、 投资建议 20

6、 风险提示 21

图表目录

图表 1：TOKENS 规模 LEADERBOARD（TOKEN 消耗量为统计口径） 8

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） 8

图表 3：KIMI K3 整体架构示意图 10

图表 4：KIMI K3 测试结果横向对比图 10

图表 5：GPU 优化测试结果横向对比图 11

图表 6：ROOFLINE 基准测试中 MINITRITON 性能表现示意图 11

图表 7：ARTIFICIAL ANALYSIS 智能指数中模型平均任务成本对比图 12

图表 8：2026.7.11-2026.7.17 AI 相关网站流量 13

图表 9：STEP AGENTIC-NATIVE OS 14

图表 10：STEP AOS 的底层设计 14

图表 11：STEP AOS 的核心能力 15

图表 12：STEPX 的首批生态合作伙伴 15

图表 13：上周 AI 初创公司融资动态 16

图表 14：上周（2026.7.13-2026.7.17 日）指数日涨跌幅 18

图表 15：上周（2026.7.13-2026.7.17 日）AI 算力指数内部涨跌幅度排名 18

图表 16：上周（2026.7.13-2026.7.17 日）AI 应用指数内部涨跌幅度排名 19

图表 17：FICONTEC2025 年年中至今公告订单 20

仅供内部参考，请勿外传！

图表 18：重点关注公司及盈利预测 21

仅供内部参考，请勿外传！

1、算力动态：算力租赁价格平稳，月之暗面正式发布 Kimi K3

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 7 月 13 日至 7 月 19 日，周度 Token 消耗量显著上涨，消耗量为 62.3T，环比上周增加 18.44%。在 Tokens 规模 Leaderboard 前五名中，Tencent 的 Hy3 以 11.5T tokens 位居榜首；Xiaomi 的 MiMo-V2.5 以 9.21T tokens 位列第二；DeepSeek 的 DeepSeek V4 Flash 以 5.35T tokens 位列第三；MiniMax 的 MiniMax M3 以 3.77T tokens 位居第四；Z.AI 旗下的 GLM 5.2 以 3.65T tokens 位居第五。

文本类模型的市场份额，Google 以 886M requests 占据 25.5%的份额，位居第一；DeepSeek 以 714M requests 占据 20.6%，位列第二；OpenAI、Tencent、Qwen 则分别以 535M、220M、179M requests，对应占据 15.4%、6.3%、5.1%的市场份额。

图像生成类模型调用量小幅上升，调用量为 5.34M requests，环比上周增加 0.19%。在 Image Leaderboard 中，Google 的 Nano Banana 系列包揽榜单前三名，其中 Nano Banana (Gemini 2.5 Flash Image)、Nano Banana 2 (Gemini 3.1 Flash Image Preview)、Nano Banana 2 (Gemini 3.1 Flash Image)分别以 1.35M、811K、658K requests 位居第一、第二和第三；OpenAI 的 GPTImage 2 以 624K requests 位列第四；Google 的 Nano Banana Pro (Gemini 3 Pro Image Preview)以 521K requests 位列第五。从市场份额来看，Google 以 3.89M requests 占据 73.0%的份额，位居第一；OpenAI 以 864K requests 占据 16.2%，位列第二；xAI、Black Forest Labs、ByteDance Seed 则分别以 268K、125K、110K requests，对应占据 5.0%、2.4%、2.1%的市场份额。

嵌入模型调用量有所上升，调用量为 232M requests，环比上周增加 4.04%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 78.6M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 69.5M requests 位列第二；Google 的 Gemini Embedding 001 以 11.4M requests 位列第三；OpenAI 的 Text Embedding 3 Large 以 11.3M requests 位列第四；Google 的 Gemini Embedding 2 以 10.4M requests 位列第五。从市场份额来看，OpenAI 以 93.1M requests 占据 40.2%的份额，位居第一；Qwen 以 75.3M requests 占据 32.5%，位列第二；Google、Perplexity、BAAI 则分别以 24.8M、13.3M、12.7M requests，对应占据 10.7%、5.8%、5.5%的市场份额。

重排序模型调用量显著上升，调用量为 3.63M requests，环比上周增加 28.23%。在 Rerank Leaderboard 中，NVIDIA 的 Nemotron Rerank VL 1B V2 以 1.98M requests 位列榜首；Cohere 的 Rerank v3.5、Rerank 4 Pro、Rerank 4 Fast 以 623K、330K、244K requests 位列第二、第三和第四。从市场份额来看，NVIDIA 以 1.98M requests 占据 62.3%的份额，位居第一；Cohere 以 1.2M requests 占据 37.7%的份额，位列第二。

视频生成类模型调用量有所增长，调用量为 142K requests，环比上周增加 10.08%。在 Video Leaderboard 前五名中，Google 的 Veo 3.1 Fast 以 43K requests 位列第一；ByteDance 的 Seedance 2.0 以 23K requests 位居第二；Google 的 Veo 3.1 Lite、Veo 3.1 分别以 20K、19K requests 位列第三、第四；xAI 的 Grok Imagine Video 以 9.15K requests 位居第五。从市场份额来看，Google 以 81K requests 占据 57.2%的份额，位居第一；ByteDance 以 36K requests 占据 25.4%，位列第二；xAI、KwaiVGI、Alibaba 则分别以 9K、8K、7K requests，

仅供内部参考，请勿外传！

对应占据 6.4%、5.3%、5.3%的市场份额。

语音生成类模型调用量大幅回落，调用量为 976K requests，环比上周减少 24.34%。在 Speech Leaderboard 前五名中，Google 的 Gemini 3.1 Flash TTS Preview 以 419K requests 位居第一；Hexgrad 的 Kokoro 82M 以 344K requests 位列第二；xAI 的 Grok Voice TTS 1.0 以 152K requests 位列第三；Microsoft 的 MAI-Voice-2 以 44K requests 位列第四；Deepgram 的 Aura-2 以 7.14K requests 位列第五。从市场份额来看，Google 以 419K requests 占据 43.0%的份额，位居第一；Hexgrad 以 344K requests 占据 35.3%，位列第二；xAI、Microsoft、Deepgram 则分别以 152K、44K、7K requests，对应占据 15.6%、4.5%、0.7%的市场份额。

语音转写类模型调用量有所上升，调用量为 3.98M requests，环比上周增加 13.39%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3 以 1.43M requests 位列第一；OpenAI 的 Whisper Large V3 Turbo 以 991K requests 位列第二；OpenAI 的 GPT-4o Mini Transcribe 以 505K requests 位列第三；NVIDIA 的 Parakeet TDT 0.6B v3 以 314K requests 位列第四；Qwen 的 Qwen3 ASR Flash 以 229K requests 位列第五。从市场份额来看，OpenAI 以 3.15M requests 占据 79.1%的份额，位居第一；NVIDIA 以 314K requests 占据 7.9%，位列第二；Qwen、Mistral AI、Microsoft 则分别以 229K、178K、60K requests，对应占据 5.7%、4.5%、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。

7 月 18 日，首个城市级绿色词元（Token）工厂项目正式签约落地常州。该项目由常州市数据局、常州高新区、国家电投电投综能、山铁数科、阿里云智能集团、网宿科技、深度动力七方共同合作，旨在构建完善的算电协同生态体系。

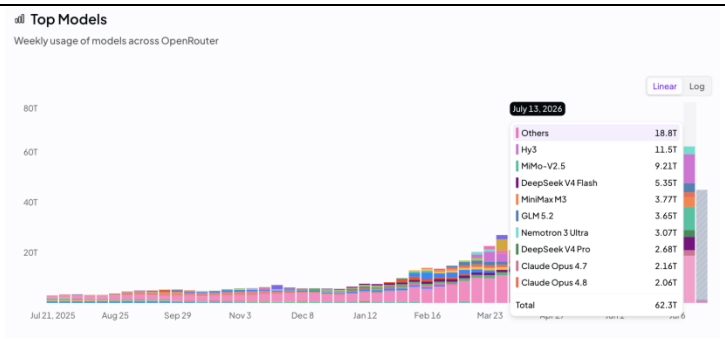
7 月 18 日，云天励飞董事长兼 CEO 陈宁在世界人工智能大会（WAIC2026）上指出，AI 大推理时代已经到来，算力竞争正从单颗芯片性能转向 Token 生成效率。与 AI 训练时代不同，推理时代面向大量不同模型与应用，核心需求是覆盖多场景、实现高吞吐与低成本。为此，云天励飞提出 GPNPU 架构并计划推出三款云端推理芯片（DeepVerse100P、DeepVerse100D、DeepVerse100L），同时联合产业链发起“1001 计划”，推动 Token 生成效率与成本协同优化。

7 月 19 日，华为副总裁曹冲于世界人工智能大会（WAIC2026）上发表演讲并指出金融行业生产级智能体在 Token 供给与消费环节面临三大核心挑战，其中包括实现金融知识及业务逻辑与智能体的深度融合、保障全链路中的数据安全与业务合规，以及兼顾成本优化，探索混合架构，构建企业级高可靠、可持续架构。

7 月 19 日，中国工程院院士、清华大学计算机科学与技术系教授郑纬民于 Agentic 时代 AI 基础设施创新发展论坛上指出，算力规模的快速扩张不等同于高效的 Token 产出能力，芯片并非智能体时代单一的稀缺性资源，能够稳定、低成本、高等级地产出 Token 的系统能力也同等重要。当前，Token 正从技术指标演变为产业核心生产要素，而推理系统的发展方向也已从单机优化转向分布式、缓存化、异构化与服务化。

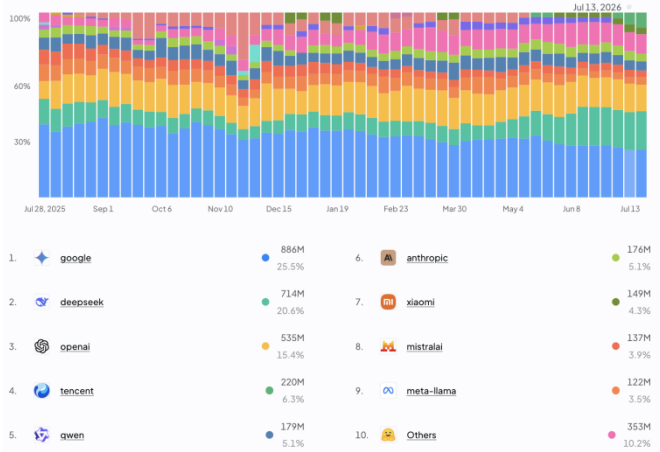
仅供内部参考，请勿外传！

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、数据跟踪：阿里云发布实时语音交互模型 Qwen-Audio-3.0-Realtime，HappyOyster 1.0 开启灰度测试

2026 年 7 月 16 日，阿里云发布实时语音交互模型 Qwen-Audio-3.0-Realtime，相较于上一代模型，该模型在推理能力、Agent 工具调用、共情对话和双工交互四个方面进行了重点升级，同时可针对时延敏感场景直接生成回复，实现毫秒级响应。

推理能力与 Agent 能力方面，3.0-Realtime 原生支持工具调用，能够根据对话语境自行调用外部工具，并将结果融入对话记忆以支持多轮追问，同时基于 Function Call 标准协议，支持 MCP、API 及知识库的灵活接入。

共情对话方面，新版本模型在语音指令遵循基准 VStyle 上取得较高得分，并且能够根据对话语境动态调整语气、节奏、音调与情感，具体表现为，在辩论场景中能够组织反驳，在情感陪伴中能通过笑声、叹息等副语言信号实现共情回应等。

交互效率方面，该模型内置双工控制子模型，具备多模态感知能力，能够通过分析音频信号、语义内容与说话人声纹特征判断交互方式，做到有效抵抗嘈杂环境中的误打断，在多人讨论中锁定主对话对象，在多说话人切换时根据线索自然过渡，实现边说边听的高效双工交互。此外，API 还预留了 audio_prompt 字段，支持用户自行上传音频样本以锁定

仅供内部参考，请勿外传！

特定说话人声纹。

在训练框架方面，该模型采用 On-Policy Distillation 框架，将文本大模型的推理能力蒸馏至语音模型，并引入多教师蒸馏策略，涵盖口语化表达、基础问答、工具调用、音频理解等方向的单独训练，以确保模型在各能力维度上的均衡发展。

此外，模型提供了推理更强的 Plus 版本和速度更快的 Flash 版本，供用户根据实际需求灵活选择。在技术表现上，Plus 版本在语音问答基准 VoiceBench 上，获得标准 prompt 提问下 92.5 的得分，以及口语化 prompt 提问下 90.5 的得分，二者之间仅 2 分的差距；Flash 版本在多轮音频对话基准 AudioMultiChallenge 中，标准和口语化 prompt 提问下得分分别为 43.6 和 38.1，差距为 5.5 分。

与此同时，7 月 19 日，阿里云百炼正式上线开放式世界模型产品 HappyOyster 1.0。该模型从海量自然视频中学习物理世界的状态转移规律，能够主动推演从动作到反馈的因果链，并保持人物与环境的长程一致性，支持用户在任意节点介入并改变数字世界走向。

产品提供世界探索（Adventure）和实时导演（Directing）两大模式，其中，世界探索模式支持 1 分钟连续实时交互，实时导演模式则支持暂停、回溯与分支叙事，同时支持生成 3 分钟以上的 480p 或 720p 实时画面，用户仅需输入一句话或一张图即可生成可互动、可探索的 AI 数字世界。

目前，该产品已提供 Android、iOS、Web 三端 SDK，同时在阿里云百炼开启灰测。

1.3、 产业动态：月之暗面正式发布 Kimi K3

2026 年 7 月 17 日，月之暗面正式推出新一代大模型 Kimi K3。该模型面向长程编程、知识工作和推理等前沿智能场景进行设计，整体基于 MoE 架构采用 Stable LatentMoE 框架，拥有 100 万 token 上下文窗口，总参数达 2.8 万亿，是目前全球参数规模最大的开源模型。

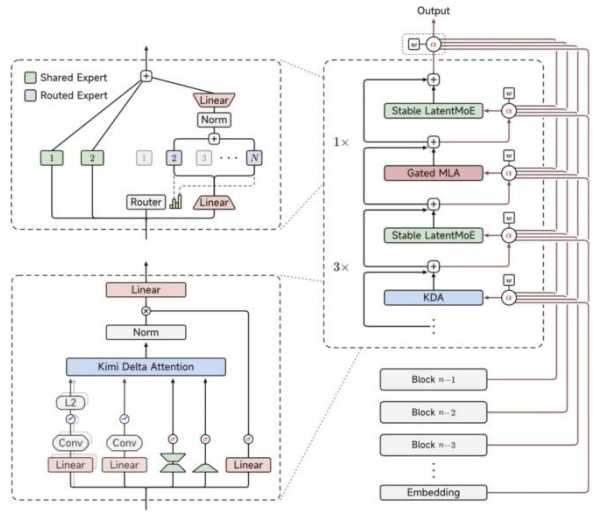
架构方面，结合 Stable LatentMoE 框架后，MoE 架构的稀疏度得以进一步提升，即模型可以在 896 个专家中高效激活 16 个以严格控制计算量。同时，该模型引入两项关键架构创新：Kimi Delta Attention（KDA）和 Attention Residuals（AttnRes）。其中，KDA 为混合线性注意力机制，可将 KV 缓存使用量减少 75%，并将 100 万 token 上下文下解码吞吐量最高提升 6 倍，AttnRes 则让模型能有选择地跨层检索信息，避免深层模型遗忘底层学习的内容，二者共同构成模型的架构骨架，使模型能够扩展到万亿参数以上的规模。以上结构性的改进共同作用于 K3，使之整体训练效率较上一代 K2 提升约 2.5 倍。

性能表现方面，K3 整体性能虽仍落后于 Claude Fable 5 和 GPT-5.6 Sol，但已稳定超过其他所有模型。数据显示，在编程评测中，K3 在 Terminal Bench 2.1 基准上获得 88.3 的得分，超过 Fable 5 的 84.6，接近于 Sol 的 88.8；在 FrontierSWE 测评中，该模型得分 81.2，仅次于 Fable 5 的 86.6，远高于 Sol 的 71.3；在 Program Bench 基准上，该模型获得 77.8 的评分，超过 Sol 的 77.6 以及 Fable 5 的 76.8；在 SWE Marathon 测评中，该模型得分 42.0，领先于 Opus-4.8 的 40.0、GPT-5.6 Sol 的 39.0，以及 Fable 5 的 35.0。

与此同时，在知识工作评测中，该模型以 1668 的得分在 GDPval-AA v2 中仅次于 Fable 5 的 1760 和 Sol 的 1748；在 AA-Briefcase 测评中得分 1548，仅次于 Fable 5，排名第二；在 BrowseComp 中得分 91.2，同时超过 Sol 和 Fable 5。此外，在视觉智能评测中，该模型在 CharXiv 中得分 91.3，仅次于 Fable 5，并在 Zerobench 中得分 41.0，与 GPT-5.5 持平。

仅供内部参考，请勿外传！

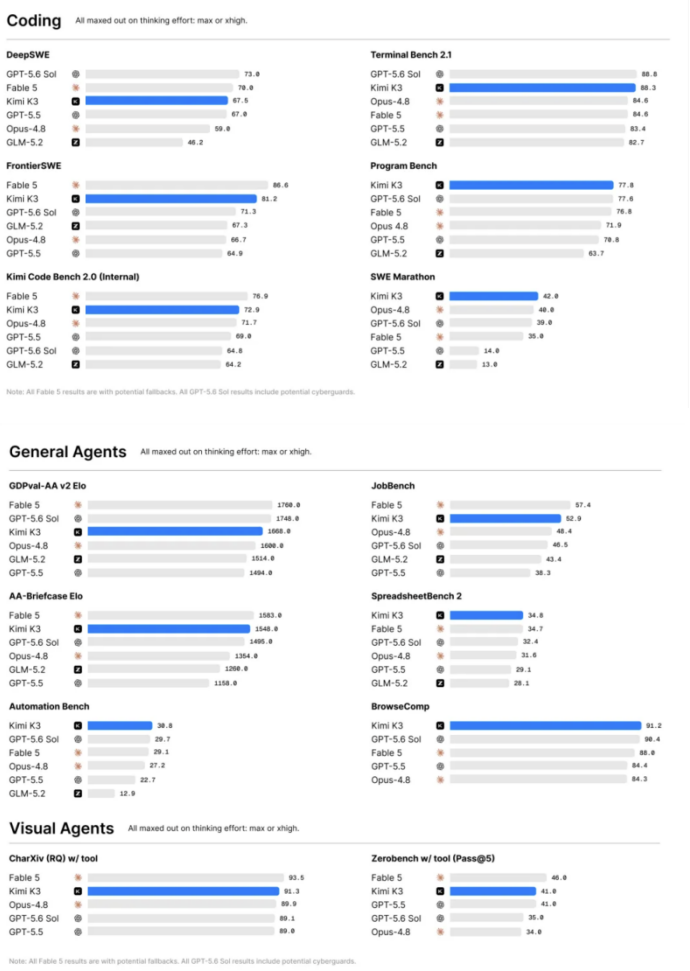
图表 3：Kimi K3 整体架构示意图



Kimi K3 architecture: the Stable LatentMoE and KDA modules (left), the AttnRes operation α (top right), and the Block Attention Residuals backbone (right).

资料来源：月之暗面，华鑫证券研究

图表 4：Kimi K3 测试结果横向对比图



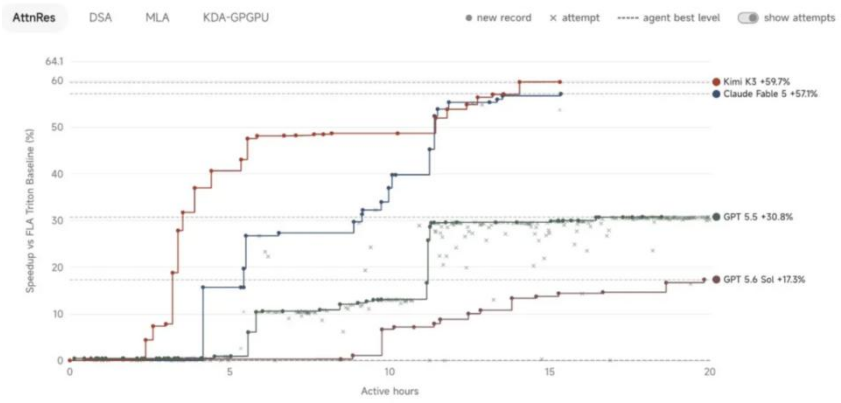
资料来源：月之暗面，华鑫证券研究

仅供内部参考，请勿外传！

在能力演示方面，K3 展现了多领域自主任务能力。在 Kernel 优化任务中，模型在 GPU 沙箱中独立分析、重写和验证内核，测验结果显示，在最大思考强度下，K3 表现接近 Fable 5，并明显领先 Opus-4.8、GPT-5.6 Sol 和 GPT-5.5。在 GPU 编译器开发任务中，K3 独立完成了 MiniTriton 编译器从零到一的构建，并且通过了端到端验证，Roofline 基准测试结果显示，其开发的编译器性能达到或超越 Triton 和 torch.compile。

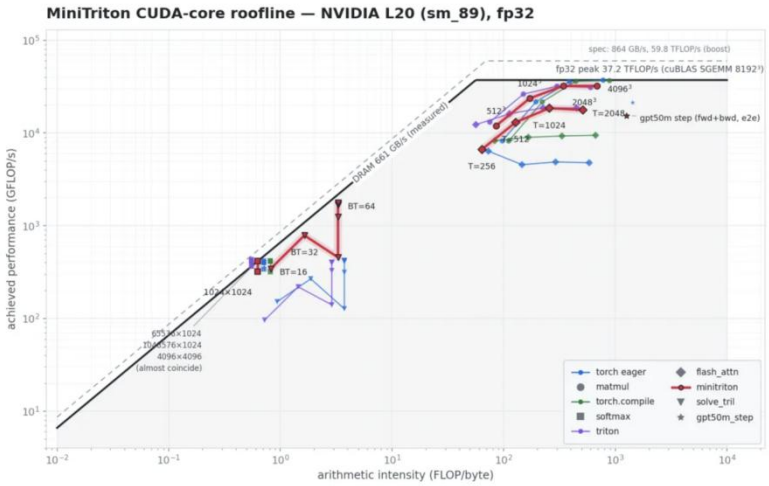
此外，在芯片设计领域，K3 能够基于开源 EDA 工具在 48 小时内独立完成一颗 4mm²芯片的设计、优化和验证；在科研编程方面，该模型可以打通科学文献与可执行代码，自主完成复杂计算研究流程的实现、验证和分析，用约两小时时间完成资深研究人员一到两周才能完成的工作；在行业研究领域，该模型能够完成推理芯片行业研究报告，历经 120 多轮递归自我改进，覆盖 42 年历史、处理超 11000 页资料；在视频剪辑方面，该模型能用 56 段原始素材自动完成选片、动作匹配剪辑、逐帧卡点与音频处理，完成通常需专业剪辑师 1 至 2 天的工作量。

图表 5：GPU 优化测试结果横向对比图



资料来源：月之暗面，华鑫证券研究

图表 6：Roofline 基准测试中 MiniTriton 性能表现示意图



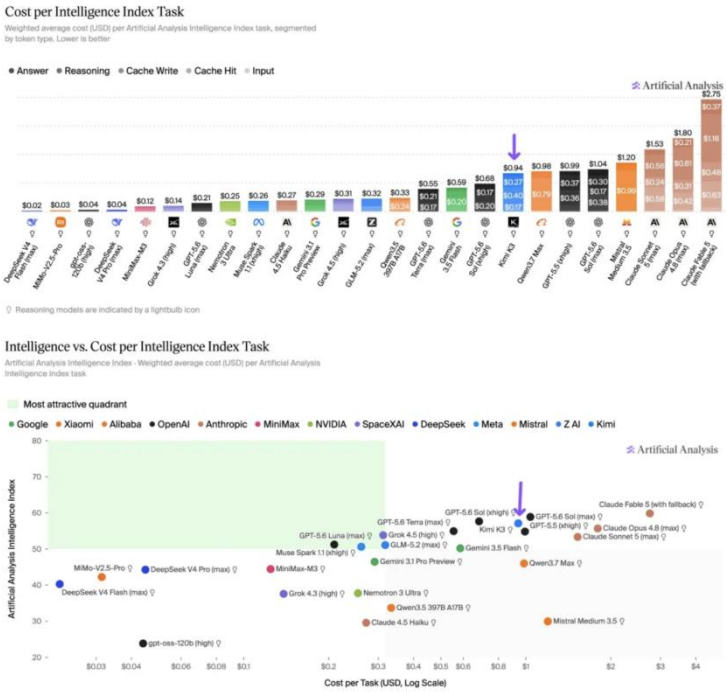
资料来源：月之暗面，华鑫证券研究

仅供内部参考，请勿外传！

定价方面，K3 定价对标 Sonnet 模型，输入价格为 2 元/百万 Token（命中缓存）和 20 元/百万 Token（未命中缓存），输出价格为 100 元/百万 Token。借助 Mooncake 分离式推理架构，官方 API 编程场景缓存率超过 90%，实际输入价格仅为标准价格的四分之一。根据 Artificial Analysis 数据，K3 在智能指数上的平均任务成本为 0.94 美元，接近 GPT-5.6 Sol 的 1.04 美元，约为 Opus 4.8 的一半。

与此同时，官方公布了该模型的三条已知局限：模型对历史思考内容较为敏感，若 Agent 框架未按要求回传全部历史思考内容可能引发上下文干扰；模型在长程高难任务中遇到模糊意图时可能替用户做出非预期决定；与 Claude Fable 5 和 GPT-5.6 Sol 相比，用户体验仍有提升空间。当前，该模型已可在 Kimi 网页版、Kimi App、Kimi Work 及官方 API 中使用，默认思考强度为 max 模式，完整权重将于 7 月 27 日前正式发布。

图表 7：Artificial Analysis 智能指数中模型平均任务成本对比图



资料来源：Founder Park，华鑫证券研究

仅供内部参考，请勿外传！

2、 AI 应用动态：Kimi 周访问量环比+29.10%，Step AOS 与 STEPX Neo 发布

2.1、 周流量跟踪：Kimi 周访问量环比+29.10%

本期（2026.7.11-2026.7.17）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1223.0M）、Bing（842.5M）和 Gemini（611.2M），访问量环比增速第一为 Kimi（29.10%）；平均停留时长前三位分别为 Character.AI（00:16:19）、Discord（00:11:14）和 Kimi（00:07:51）;平均停留时长环比增速第一为 NotionAI（13.18%）。

图表 8：2026.7.11-2026.7.17 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	1223.0	1.75%	6:28	1.57%
Bing	搜索	微软	842.5	1.02%	7:15	0.23%
Gemini	聊天机器人	谷歌	611.2	0.21%	7:04	-0.24%
Canva	在线设计	Canva	185.1	0.93%	5:59	0.28%
Discord	游戏社区	微软	149.6	0.13%	11:14	0.15%
Github	代码托管	微软	141.4	-0.56%	6:22	0.53%
Character.AI	聊天机器人	Character.AI	37.83	-1.56%	16:19	1.45%
Perplexity	AI 搜索	Perplexity	24.11	3.17%	4:19	-1.52%
DeepI	翻译工具	DeepL	23.82	-0.08%	2:29	0.68%
Kimi	聊天机器人	Moonshot AI	10.93	29.10%	7:51	-0.84%
QuillBot	释义工具	QuillBot	7.21	2.97%	2:53	0.00%
NotionAI	文本/笔记	Notion	2.653	2.04%	7:35	13.18%
文心一言	聊天机器人	百度	0.31	5.88%	2:19	-13.13%

资料来源：similarweb,华鑫证券研究

2.2、 产业动态：Step AOS 与 STEPX Neo 发布，开启人机共生的终端新纪元

2026 年 7 月 13 日，阶跃星辰展示了三款 AI 终端产品：全球首个大模型原生 AI 终端品牌 STEPX、全球首个智能体原生操作系统 Step AOS，以及首款大模型原生智能体手机 STEPX Neo，反映出行业竞争焦点正从单纯的模型性能比拼演进至“模型-软件-硬件”三位一体的全栈生态较量。

仅供内部参考，请勿外传！

Step AOS 的核心突破在于颠覆了以“人”为操作主体的系统设计前提。传统系统为人类交互而构建，智能体若要在此类环境中运行，只能模拟人类手指进行点击滑动，这便在记忆的连贯性、决策的时效性以及行动的可信度与权限管理上，遭遇了难以逾越的架构性瓶颈。

因此，阶跃星辰推出了“Step 模型矩阵×Step Agentic-native OS”。Step AOS 是“智能体原生”的操作系统，其核心理念可概括为“AI is the OS”。该系统依托于阶跃星辰强大的模型矩阵，该矩阵涵盖旗舰 Pro、标准 Flash、端侧 Edge 模型。音频模型 Step Audio 2.5 系列位列全球前三，视觉语言模型 Step VL 达到行业 SOTA，涵盖音频、图像、视频、终端交互全模态等领域。特别是其最新的 Step Edge 端侧基座模型，在 29 项权威基准评测中位列同类端侧模型第一，能够独立处理轻量级任务，并在复杂场景下与云端模型高效协同，兼顾了响应速度与数据隐私。

图表 9：Step Agentic-native OS



资料来源：新智元，华鑫证券研究

在架构层面，Step AOS 构建了一个三层全新的技术框架。其底层是一个面向智能体重构的资源供给底座，统一调度端侧与云端的计算资源，打破传统应用沙箱的数据隔离，建立跨应用的统一语义数据层。同时，系统将各类应用与服务拆解为数千个最小能力单元，并兼容 MCP、A2A、CLI 等协议，支持渐进式装载与统一 API，第三方工具即插即用。这一底层设计使其天然具备跨手机、PC 及 IoT 设备的部署能力。

图表 10：Step AOS 的底层设计



资料来源：新智元，华鑫证券研究

在此坚实基础上，系统的中层具备智能体所需的核心能力：记忆、决策与执行，以及安全。其中，个人智能体“阶跃 Amoo”拥有独特的双域记忆结构，既能深度记忆用户的个人特征与历史，也能记录自身的能力进化，通过高效的“记-理-忆”链路，在 PersonalMem、LongMemEval 等评测中达到 SOTA 水平，且召回速度极快，日常问答记忆召回最快仅需 15 毫

仅供内部参考，请勿外传！

秒。在决策层面，系统采用“端云多脑”机制，端侧模型能确保百毫秒级响应高频简单任务，复杂推理则自动升级至云端处理。这个机制使得其在 Android Daily 上，GUI 任务成功率排名国内第一；在 Claw Eval 上，Claw 任务成功率排名国内第一；在 Step AOS-E2E-BMK 上，终端任务执行成功率排名国内第一。在安全方面，阶跃星辰联合上海人工智能实验室制定了“可见、可信、可控、可逆”的四重底线原则，制定了《新一代智能体系统安全技术白皮书》，为智能体的行动能力建立了可信的边界。

图表 11：Step AOS 的核心能力



资料来源：新智元，华鑫证券研究

Step AOS 的顶层设计是一套意图驱动的自然交互，将底层变革转化为一种全新的“结果交互”范式。用户不再需要通过繁琐的步骤下达指令，而是只需表达意图，系统便能利用多模态输入和情境感知，自动完成理解、规划与执行的全链路。目前，STEPX 已与携程、支付宝、滴滴、美团、WPS、剪映等一批生态伙伴达成深度合作，涵盖出行、生活服务、办公、创作等多个领域，旨在构建一个开放、共生的智能生态。

图表 12：STEPX 的首批生态合作伙伴



资料来源：新智元，华鑫证券研究

仅供内部参考，请勿外传！

3、AI 融资动向：Fireworks AI 完成 15.05 亿美元 D 轮融资，投后估值达 175 亿美元

2026 年 7 月 17 日，AI 基础设施公司 Fireworks AI 宣布完成 15.05 亿美元 D 轮融资，投后估值达 175 亿美元。本轮融资由 Atreides Management、Index Ventures、TCV 三家机构联合领投，20VC、Bessemer Venture Partners、Insight Partners、Lone Pine Capital 等多家机构参投，现有投资方 Evantic、Lightspeed Venture Partners 及 NVIDIA 持续跟投。当前，公司年化营收（ARR）突破 10 亿美元，与上一轮融资阶段相比，增长超过 5 倍。本轮所募集资金将用于扩大工程团队规模、拓展全球算力基础设施，以及进一步深化与 Microsoft、NVIDIA 等云计算战略合作伙伴的合作布局。

该公司将专属智能 AI 平台作为自身核心定位，致力于帮助企业利用自身独有的私有数据和行业知识，构建属于自己的 AI 模型。其核心产品为一个面向开发者的云平台，支持对开源 AI 模型进行微调和定制化部署，同时提供托管 GPU 集群访问服务，并按用量计费。平台内嵌 AI 智能体，并具备自动化完成模型训练的全流程操作能力，开发者仅需明确微调目标并上传训练数据集，智能体便可以自动处理底层技术细节。当前，平台已上线超过 200 个模型，覆盖文本生成、图像生成及多模态等多个领域，同时承诺针对重要开源模型的新版本发布，平台将在其发布后的数小时内完成上线支持，以保障用户快速获取最新开源成果，用于专属模型构建。

技术方面，其 AI 智能体具备自动搜索最优超参数组合的能力，从而最大化模型输出质量，并在必要时通过 DPO 文件对用户提供的训练数据集进行扩充。针对训练加速，该平台支持四种并行训练技术，针对不同模型类型提供差异性优化，开发团队可根据需求选择四种技术并行运行或单独选用其一。推理服务方面，该平台提供无服务器（serverless）环境、Deployments 服务两种部署选项，其中，在无服务器环境中，用户无需管理底层基础设施，而 Deployments 服务则提供专属 GPU 集群，相较而言具备更优的性能表现。此外，Deployments 服务还支持更加丰富的自定义配置，其中包括弹性伸缩策略的调整，以及通过量化技术对模型进行压缩，从而有效降低基础设施成本。

业务表现方面，平台每日服务的 Token 处理量已从 15 万亿跃升至超过 40 万亿，实现了近三倍的增长，其中高达 95%的 Token 处理量来自于经过专属化定制的模型。其核心客户包括 Uber、Shopify、Doximity、Revolut 等全球头部企业，这些经过定制化的专属模型在企业核心任务上的表现可媲美甚至超越通用闭源模型，同时在响应速度和运营成本方面具备明显优势。

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
Fireworks AI	AI 基础设施	Atreides Management、Index Ventures、TCV	D 轮	15.05 亿美元	超 18 亿美元	175 亿美元

仅供内部参考，请勿外传！

Chai Discovery	AI 药物发现	Index Ventures	C 轮	4 亿美元	6.3 亿美元	38 亿美元
Oak	AI 原生身份控制	Accel、CRV、Greylock Partners	种子轮	6000 万美元	6000 万美元	未披露

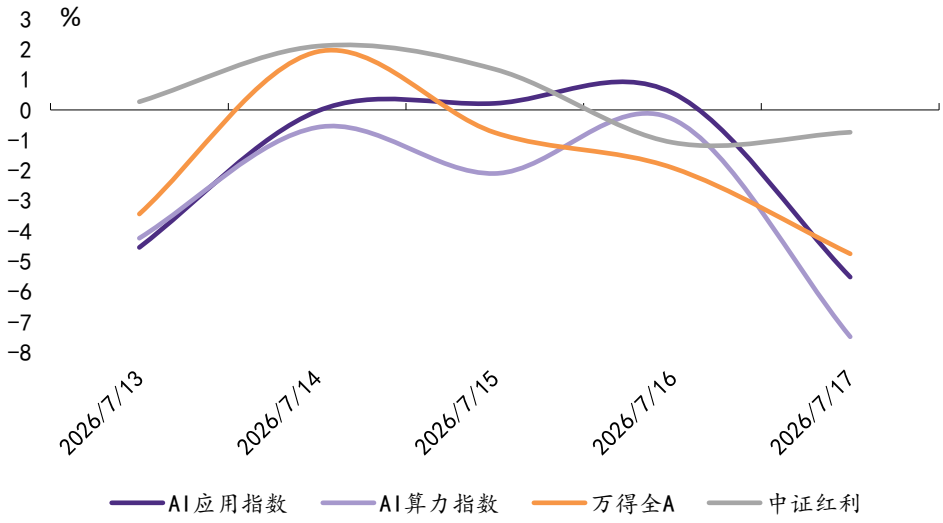
资料来源：wind，Saasverse，中财网，华鑫证券研究

仅供内部参考，请勿外传！

4、 行情复盘

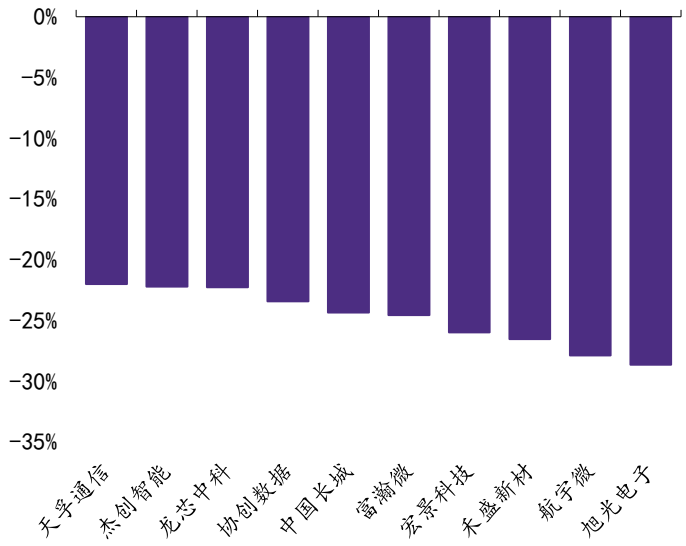
上周（2026.7.13-2026.7.17 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 0.56%/-0.3%/1.92%/2.1%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-5.52%/-7.49%/-4.75%/-1.07%。AI 算力指数内部，旭光电子以-28.67%录得上周最大跌幅。AI 应用指数内部，中岩大地以 37.65%录得上周最大涨幅，深科技以-32.64%录得上周最大跌幅。

图表 14：上周（2026.7.13-2026.7.17 日）指数日涨跌幅



资料来源：wind，华鑫证券研究

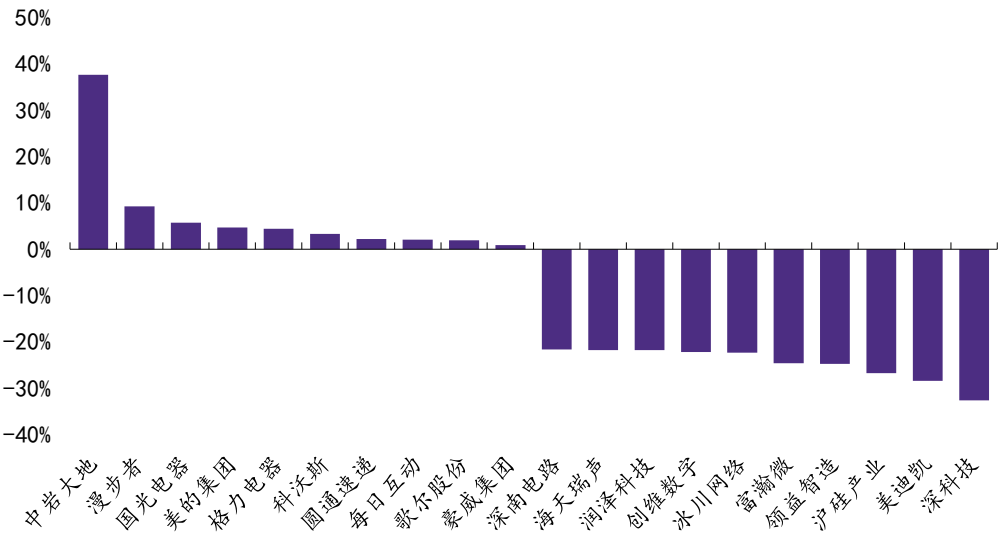
图表 15：上周（2026.7.13-2026.7.17 日）AI 算力指数内部涨跌幅度排名



资料来源：wind，华鑫证券研究

仅供内部参考，请勿外传！

图表 16：上周（2026.7.13-2026.7.17 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

仅供内部参考，请勿外传！

5、 投资建议

2026 年 7 月 14 日，TowerSemiconductor（高塔半导体）宣布在日本政府支持下启动双轨战略扩产，重点扩大 300mm 硅光子、硅锗及先进封装产能，项目总投资约 40 亿美元，其中日本政府补助约 10 亿美元、公司自有投资约 30 亿美元。第一轨将荒井工厂（原 Fab6）改造为 300mm 硅光子产线及先进封装产能，并最大化鱼津 Fab7 产能输出，预计 2027Q4 实现全面量产就绪；第二轨同步启动，待协议签署完成后于鱼津 Fab7 旁新建 300mm 厂房，达产后硅光子与硅锗产能将提升数倍，主要承接 AI 与数据中心驱动的下一代光连接需求，预计 2029 年起显著增厚业绩。基于扩产计划，公司同步上调 2028 年财务目标：预计营收 36 亿美元（高于市场普遍预期的 31.7 亿美元，公司 2025 年营收为 16 亿美元），净利润 12 亿美元（公司 2025 年净利润为 2.2 亿美元），且仅依靠第一轨即可达成。此外，公司年内资本开支已明显提速：2026 年初原定资本开支 6.5 亿美元，2 月于财报说明会上宣布追加 2.7 亿美元，合计 9.2 亿美元投向硅光设备采购，计划 2026 年底完成全部设备进场认证，以支撑 2027 年硅片出货量达到 2025Q4 月均出货量的 5 倍以上。

本次双轨扩产印证了硅光晶圆的结构性短缺，上游设备环节将成为公司资本开支传导下的核心受益方向。需求端看，即使产能扩张至 5 倍以上，公司硅光供给仍然紧张：截至 2028 年的硅光总产能中超过 70%已被预订或正在预订，且有客户预付款作为保障；今年 5 月公司更与头部硅光客户签订 2027 年 13 亿美元长期供货协议，已收到 2.9 亿美元预付款。这表明短缺是结构性的、而非周期性的，产能扩张具备持续性——今年 2 月的追加投资系为缓解供给紧张而实施的阶段性产能补充，本次的扩产则为中长期战略布局，二者在规划属性上存在本质区别。从传导路径看，晶圆厂扩产的资本开支将直接转化为设备采购需求：公司 9.2 亿美元硅光/硅锗投资以设备采购为主，本次 40 亿美元双轨投资亦以产线改造和新厂建设为主。Tower 作为全球硅光代工核心产能，其大规模扩产、产能被长约锁定的事实，印证了光通信需求的高景气与持续性；晶圆厂资本开支将直接转化为设备采购需求，上游硅光晶圆测试、耦合封装等高精度设备环节有望率先且持续受益于订单放量。

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

仅供内部参考，请勿外传！

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
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
	总金额			约 17.93 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	2026-07-21	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	373.10	-0.30	0.30	0.60	-1243.67	1243.67	621.83	买入
301196.SZ	唯科科技	75.33	2.53	3.34	3.98	29.77	22.55	18.93	买入
301369.SZ	联动科技	129.80	0.48	1.08	2.62	270.42	120.19	49.54	买入
603859.SH	能科科技	35.32	0.92	1.21	1.50	38.39	29.19	23.55	买入
688615.SH	合合信息	97.72	3.24	4.22	5.25	30.16	23.16	18.61	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。

仅供内部参考，请勿外传！

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。

仅供内部参考，请勿外传！

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。