

2026 年 08 月 04 日



华鑫证券
CHINA FORTUNE SECURITIES

MiniMax 发布 MiniMax, H3, Kimi K3 正式开源

— 计算机行业周报

推荐(维持)

投资要点

分析师：任春阳 S1050521110006

reny@cfsc.com.cn

行业相对表现

表现	1M	3M	12M
计算机(申万)	-1.5	-13.4	-11.9
沪深 300	-6.2	-5.5	11.6

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：Anthropic 发布 ClaudeOpus5，谷歌发布三款 Flash 模型》2026-07-28
- 2、《计算机行业周报：月之暗面正式发布 KimiK3，StepAOS 与 STEPXNeo 发布》2026-07-22
- 3、《计算机行业周报：OpenAI 正式发布 GPT-5.6 系列大模型，SpaceX AI 发布 Grok4.5》2026-07-15

算力：算力租赁价格平稳，MiniMax 发布通用全模态生成模型 MiniMax H3

2026 年 7 月 31 日，MiniMax 正式发布通用全模态生成模型 MiniMax H3。该模型支持对文本、图像、视频、声音组成的多模态上下文的统一理解，能够生成具备原生双声道的音视频内容，最高支持 15s 2K 分辨率输出，同时具备商用级的多场景内容生成能力，在指令遵循、文字与品牌信息呈现、V2V Motion Transfer（视频到视频动作迁移）等方面表现出色，可广泛适用于广告、品牌、电商、产品设计、UI/UX、游戏等商业场景。

AI 应用：文心一言周时长环比+13.95%，Kimi K3 正式开源

2026 年 7 月 28 日，Moonshot AI 正式开源 Kimi K3，成为全球首个真正实现落地的、达到三万亿参数级别的开源模型。此次发布的内容不仅限于模型权重、技术报告，还包括其训练的关键 Infra 技术：MoonEP、FlashKDA 和 AgentEnv。

AI 融资动向：Groundcover 完成 1 亿美元 C 轮融资，投后估值达 5 亿美元

2026 年 7 月 30 日，AI 云原生可观测性平台 Groundcover 宣布完成 1 亿美元 C 轮融资，投后估值达 5 亿美元。本轮融资由 One Peak 领投，Morgan Stanley Expansion Capital 及原有投资方 Zeev Ventures、Angular Ventures、Heavybit、Jibe 共同参投。当前，公司累计融资总额已达 1.6 亿美元，本轮融资将主要用于加速北美市场的商业化拓展，推进国际市场布局，并重点强化与云服务提供商的联合销售合作。

投资建议

2026 年 7 月 29 日，微软发布 2026 财年第四财季（截至 6 月 30 日）财报：当季实现总营收 900 亿美元，同比增长 18%；净利润 358 亿美元，同比增长 31%；摊薄后每股收益 4.81 美元，同比增长 32%。分业务看，智能云部门当季营收 393 亿美元，同比增长 32%，其中 Azure 及其他云服务营收增长 43%，较上季度提速 3 个百分点；公司首次披露 Azure2026 财年全年营收突破 1000 亿美元；商业剩余履约义务（RPO）达 6780 亿美元，同比增长 84%。生产力与业务流程部门营收 378 亿美元，同比增长 14%；Microsoft365Copilot 付费席位突破 3000 万，较三个月前披露的约 2000 万增长约 50%；Microsoft365 商用

云营收经调整后增长 16%。更多个人计算部门营收 129 亿美元，同比下降 4%，其中 Windows OEM 和设备营收下降 7%，Xbox 内容和服务营收下降 10%，搜索广告收入（不含流量获取成本）增长 10%。资本开支方面，公司单季资本开支 410 亿美元，同比增长约 70%；2026 财年全年资本开支 1453 亿美元；公司此前预计 2026 自然年资本开支约 1900 亿美元，较 2025 年增长 61%，公司解释称其中约 250 亿美元来自芯片短缺导致的内存和组件成本上升；大规模投入下公司现金流仍保持正值。

本次财报印证了 AI 算力需求的结构性高景气与 CSP 资本开支向硬件产业链的强传导。需求端看，AI 的大幅增速、6780 亿美元商业剩余履约义务及 Capex 单位单季约 50% 的环比跃升表明企业级 AI 需求仍在加速；管理层在电话会上明确表示，当季商业 RPO 的全部环比增量均由前沿模型公司以外的客户承诺贡献，印证 AI 算力需求正从少数基础模型公司向更广泛的企业级市场扩散，需求结构更趋多元。从传导路径看，我们推测 2026 自然年约 1900 亿美元的资本开支将直接转化为上游采购需求；电话会披露，当季 Azure 收入超预期主要得益于 CPU 和 GPU 资源效率提升及流程改进带来的新增容量更快交付，且本季度新增容量已迅速实现收入（quickly monetized），资本投入与收入确认之间的时间差显著缩短；管理层也重申客户需求持续超过可用容量，当前资本开支本质上是为已签署合同扩充交付能力，具备明确支撑。展望层面，公司指引下一季度 Azure 按固定汇率计算收入增长约 45%，并预计财年上半年（对应 2026 自然年下半年）增长将继续加速，表明算力供给缺口仍在扩大而非收敛。我们认为，本轮资本开支上行由真实需求驱动：积压订单高增、新增容量当季变现、后续指引加速三者共同表明算力供给仍滞后于需求增长，叠加芯片短缺推高的内存与组件成本为全行业共性压力，资本开支具备扎实的基本面支撑与持续性；对国内市场而言，微软资本开支的持续性为上游硬件供应链提供了较高的订单能见度，AI 服务器、数据中心与网络设备、IDC 及电力配套等环节的配置价值凸显。

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	2026-08-04	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	410.80	-0.30	0.30	0.60	-1369.33	1369.33	684.67	买入
301196.SZ	唯科科技	62.05	2.53	3.34	3.98	24.53	18.58	15.59	买入
301369.SZ	联动科技	109.41	0.48	1.08	2.62	227.94	101.31	41.76	买入
603859.SH	能科科技	40.10	0.92	1.05	1.32	43.59	38.19	30.38	买入
688615.SH	合合信息	114.10	3.24	4.22	5.25	35.22	27.04	21.73	买入

资料来源：Wind，华鑫证券研究

正文目录

1、 算力动态：算力租赁价格平稳，MINIMAX 发布通用全模态生成模型 MINIMAX H3.....	5
1.1、 Tokens 跟踪.....	5
1.2、 数据跟踪：Seedance 2.5 正式发布，API 服务即将上线火山方舟.....	7
1.3、 产业动态：MiniMax 发布通用全模态生成模型 MiniMax H3	8
2、 AI 应用动态：文心一言周时长环比+13.95%，KIMI K3 正式开源.....	10
2.1、 周流量跟踪：文心一言周时长环比+13.95%	10
2.2、 产业动态：参数量达 2.8 万亿，Kimi K3 正式开源	10
3、 AI 融资动向：GROUNDCOVER 完成 1 亿美元 C 轮融资，投后估值达 5 亿美元.....	14
4、 行情复盘	16
5、 投资建议	18
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：ARTIFICIAL ANALYSIS 的文生视频有声榜单.....	9
图表 4：MINIMAX H3 的应用案例（游戏 UI、电影片头）.....	9
图表 5：2026.7.24-2026.7.30 AI 相关网站流量.....	10
图表 6：KIMI K3 预训练中视觉塔梯度范数对比	11
图表 7：KIMI K3 相比 KIMI K2 在整体扩展效率上实现约 2.5 倍的提升.....	11
图表 8：KIMI K3 的知识图谱.....	12
图表 9：KIMI K3 的蒸馏.....	12
图表 10：KIMI K3 在编程、代码 AGENT、前端开发、长上下文推理等场景中的表现.....	13
图表 11：上周 AI 初创公司融资动态	14
图表 12：上周（2026.7.27-2026.7.31 日）指数日涨跌幅.....	16
图表 13：上周（2026.7.27-2026.7.31 日）AI 算力指数内部涨跌幅度排名	16
图表 14：上周（2026.7.27-2026.7.31 日）AI 应用指数内部涨跌幅度排名	17
图表 15：FICONTEC2025 年年中至今公告订单.....	18
图表 16：重点关注公司及盈利预测	19

1、算力动态：算力租赁价格平稳，MiniMax 发布通用全模态生成模型 MiniMax H3

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 7 月 27 日至 8 月 2 日，周度 Token 消耗量有所下降，消耗量为 56.8T，环比上周减少 2.07%。在规模 leaderboard 前五名中，DeepSeek 的 DeepSeek V4 Flash 以 7.22T tokens 位列榜首；xAI 的 MiMo-V2.5 以 6.3T tokens 位居第二；Tencent 的 Hy3 以 4.82T tokens 位居第三；DeepSeek 的 DeepSeek V4 Pro 以 3.28T tokens 位居第四；Z. AI 旗下的 GLM 5.2 以 2.89T tokens 位居第五。

文本类模型的市场份额，Google 以 967M requests 占据 26.1% 的份额，位居第一；DeepSeek 以 923M requests 占据 24.9%，位列第二；OpenAI、Qwen、Anthropic 则分别以 656M、190M、148M requests，对应占据 17.7%、5.1%、4.0% 的市场份额。

图像生成类模型调用量大幅上升，调用量为 6.78M requests，环比上周增加 21.94%。在 Image Leaderboard 中，Google 的 Nano Banana (Gemini 2.5 Flash Image)、Nano Banana 2 (Gemini 3.1 Flash Image) 分别以 1.67M、1.61M requests 位居第一、第二；OpenAI 的 GPTImage 2 以 627K requests 位列第三；Google 的 Nano Banana 2 (Gemini 3.1 Flash Image Preview)、Nano Banana Pro (Gemini 3 Pro Image) 分别以 626K、533K requests 位居第四和第五。从市场份额来看，Google 以 5.27M requests 占据 77.8% 的份额，位居第一；OpenAI 以 843K requests 占据 12.4%，位列第二；xAI、Black Forest Labs、ByteDance Seed 则分别以 294K、165K、136K requests，对应占据 4.3%、2.4%、2.0% 的市场份额。

嵌入模型调用量有所上升，调用量为 287M requests，环比上周增加 4.74%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 101M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 61.7M requests 位列第二；Perplexity 的 Embed V1 0.6B 以 28.9M requests 位列第三；BAAI 的 bge-m3 以 16.4M requests 位列第四；Google 的 Gemini Embedding 001 以 16.4M requests 位列第五。从市场份额来看，OpenAI 以 119M requests 占据 41.3% 的份额，位居第一；Qwen 以 68.1M requests 占据 23.7%，位列第二；Google、Perplexity、BAAI 则分别以 35.6M、32.6M、18.4M requests，对应占据 12.4%、11.4%、6.4% 的市场份额。

重排序模型调用量大幅上升，调用量为 3.95M requests，环比上周增加 21.17%。在 Rerank Leaderboard 中，NVIDIA 的 Llama Nemotron Rerank VL 1B V2 以 2.32M requests 位列榜首；Cohere 的 Rerank v3.5、Rerank 4 Fast、Rerank 4 Pro 以 774K、562K、263K requests 位列第二、第三和第四；VoyageAI 的 rerank-2.5-lite 以 20K requests 位列第五。从市场份额来看，NVIDIA 以 2.32M requests 占据 58.6% 的份额，位居第一；Cohere 以 1.6M requests 占据 40.5% 的份额，位列第二；VoyageAI 以 36K requests 占据 0.9% 的份额。

视频生成类模型调用量大幅上升，调用量为 159K requests，环比上周增加 15.22%。在 Video Leaderboard 前五名中，Google 的 Veo 3.1 Fast、Veo 3.1 Lite 分别以 51K、27K requests 位列第一、第二；ByteDance 的 Seedance 2.0、Seedance 2.0 Fast 分别以 19K、15K requests 位居第三、四；Google 的 Veo 3.1 以 13K requests 位居第五。从市场份额来看，Google 以 90K requests 占据 56.8% 的份额，位居第一；ByteDance 以 41K requests 占

据 25.9%，位列第二；xAI、KwaiVGI、Alibaba 则分别以 11K、8K、6K requests，对应占据 6.6%、5.0%、3.6%的市场份额。

语音生成类模型调用量有所上涨，调用量为 1.64M requests，环比上周增加 7.19%。在 Speech Leaderboard 前五名中，Google 的 Gemini 3.1 Flash TTS Preview 以 887K requests 位居第一；Hexgrad 的 Kokoro 82M 以 421K requests 位列第二；xAI 的 Grok Voice TTS 1.0 以 157K requests 位列第三；Microsoft 的 MAI-Voice-2 以 27K requests 位列第四；Deepgram 的 Aura-2 以 26K requests 位列第五。从市场份额来看，Google 以 887K requests 占据 54.2% 的份额，位居第一；Hexgrad 以 421K requests 占据 25.7%，位列第二；xAI、fish-audio、Microsoft 则分别以 157K、42K、37K requests，对应占据 9.6%、2.6%、2.3%的市场份额。

语音转写类模型调用量有所上涨，调用量为 9.06M requests，环比上周增加 8.89%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3 以 5.42M requests 位列第一；OpenAI 的 Whisper Large V3 Turbo 以 2.05M requests 位列第二；OpenAI 的 GPT-4o Mini Transcribe 以 532K requests 位列第三；Qwen 的 Qwen3 ASR Flash 以 254K requests 位列第四；OpenAI 的 Whisper 1 以 208K requests 位列第五。从市场份额来看，OpenAI 以 8.31M requests 占据 91.7% 的份额，位居第一；Qwen 以 254K requests 占据 2.8%，位列第二；Mistral AI、Microsoft、xAI 则分别以 188K、127K、53K requests，对应占据 2.1%、1.4%、0.6%的市场份额。

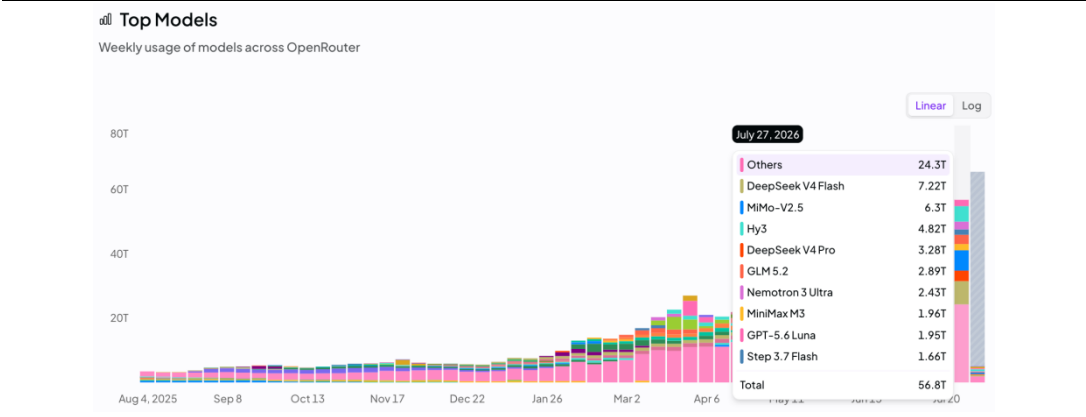
* 统计口径说明：以上数据均来源于 OpenRouter。基于官方统计口径调整，当前 Rankings 采用 Token 与 Request 两套统计维度：其中，仅 Tokens 规模 Leaderboard 仍以模型 Token 消耗量为统计口径，其余类目榜单及市场份额板块则统一采用 Request 调用量为统计口径，两者统计维度不同，不宜直接对应或进行横向比较。

7 月 31 日，OpenAI 宣布下调 GPT-5.6 系列模型 API 价格。其中，GPT-5.6 Luna 价格下调 80%，每百万 Token 输入、输出价格分别调整为 0.20 美元和 1.2 美元；GPT-5.6 Terra 价格下调 20%，降价后每百万 Token 输入 2 美元、每百万 Token 输出 12 美元；旗舰模型 GPT-5.6 Sol 定价保持不变，同时新增 Fast 模式，在保持模型能力不变的情况下，最高可实现 2.5 倍推理速度。

7 月 31 日，DeepSeek V4-Flash 正式版上线公测。该模型采用 MoE 架构，总参数规模达 284B，支持上下文长度最长达 100 万 token，其模型结构、尺寸均与此前发布的 preview 版本保持一致。价格上，该模型输入命中缓存 0.2 元/百万 token，未命中输入 1 元/百万 tokens，输出 2 元/百万 tokens。

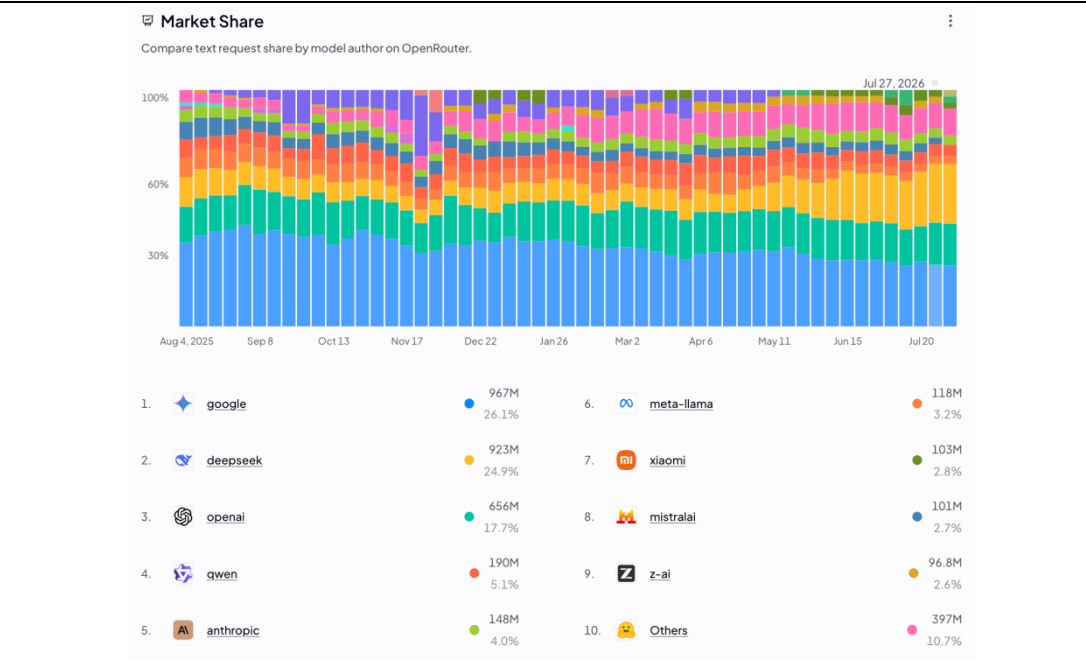
近期，大模型 API 市场呈现明显两极分化。Kimi K3 每百万 Token 输出定价 100 元，较上代涨价 2 至 3 倍，站上国产模型价格高位；DeepSeek V4-Pro 常规输入价 3 元、输出价 6 元，缓存命中后输入价进一步降至 0.25 元，成为性价比代表。表面报价分化背后，缓存命中率不透明、模型性能减配、Token 计量单位不统一等问题普遍存在，服务商从粗放价格战转向自主锚定价值，定价权博弈刚启幕。

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：Seedance 2.5 正式发布，API 服务即将上线火山方舟

2026 年 7 月 31 日，字节跳动正式发布新一代视频创作模型 Seedance 2.5。该模型延续了统一多模态音视频联合生成架构，在生成时长、多模态参考和精准编辑三个方面实现了显著突破，同时逐步嵌入真实生产流程，在工业制造、汽车、具身智能、智能驾驶等多个实体产业中实现落地应用。

核心能力方面，Seedance 2.5 实现了单次生成 30 秒高质量视频的能力，较前代产品实现翻倍增长，并可通过多轮延长至数分钟，使模型能够生成具有连贯叙事逻辑的完整内容，显著拓展了 AI 视频生成在叙事性场景中的应用边界。同时，在多模态参考能力方面，该模

型单次最多可输入 30 张图片、10 段视频和 10 段音频作为参考，能够同时理解并稳定还原复杂场景中的多个人物、场景与声音，大幅提升了模型对复杂创作需求的理解与执行能力。不仅如此，该模型支持通过时间戳控制特定片段的剧情、视角，生成后还能定向修改角色、动作或剧情，赋予创作者更高的后期调整自由度，同时强化了绿幕编辑和运镜编辑能力，可在保持主体不变的情况下替换场景或调整运镜方式，进一步降低了视频后期制作的技术门槛与时间成本。

应用方面，Seedance 2.5 的能力已从内容创作延伸至更广泛的产业场景。在内容创作领域，模型可用于影视制作、广告营销等创意行业；在教育领域，可将静态课文转化为动态视频，丰富教学内容的呈现形式；在工业制造、具身智能和自动驾驶等产业领域，模型可用于生成合成视频数据以训练机器人感知能力、进行工业仿真，或模拟暴雨、浓雾等极端天气场景，为自动驾驶系统提供多样化的训练样本。

在产业合作方面，徐工集团、小鹏汽车、智元机器人、灵初智能、微分智飞、穹彻智能、Xspark AI 等七家企业已达成合作意向，将率先接入 Seedance 2.5，标志着该模型在 B 端产业场景的落地已迈出实质性步伐。

定价方面，火山引擎公布了 Seedance 2.5 的 API 定价：含视频输入 42 元/百万 Tokens，不含视频输入 70 元/百万 Tokens。相较 Seedance 2.0，新版模型在能力大幅升级的同时，定价也相应上调。当前，新版本模型已陆续上线即梦 AI、豆包专业版等平台，API 服务也将近期上线火山方舟。

1.3、产业动态：MiniMax 发布通用全模态生成模型 MiniMax H3

2026 年 7 月 31 日，MiniMax 正式发布通用全模态生成模型 MiniMax H3。该模型支持对文本、图像、视频、声音组成的多模态上下文的统一理解，能够生成具备原生双声道的音视频内容，最高支持 15s 2K 分辨率输出，同时具备商用级的多场景内容生成能力，在指令遵循、文字与品牌信息呈现、V2V Motion Transfer（视频到视频动作迁移）等方面表现出色，可广泛适用于广告、品牌、电商、产品设计、UI/UX、游戏等商业场景。

新模型的核心亮点在于其统一的多模态理解与生成能力。真实创作往往需要融合不同模态的复杂信息，该模型支持同时引入图片、声音、视频等多模态信息源，能够自动提取不同素材中的关键特征，如镜头运动、人物身份、声音特征等，再根据自然语言描述组合成新视频，实现跨模态素材的自由组合与特征融合。在输出侧，该模型最高支持 15 秒、2K 分辨率的音视频生成，音频默认为原生双声道，为视听内容创作提供了完整的输出能力。

技术路径上，团队在 H3 之前先后研发了 Hailuo 01 和 Hailuo 02 两代视频模型，其中，前者完成从 0 到 1 的系统搭建，后者重点提升架构效率、数据质量和模型规模。过往的生成模型通常将图片生成、声音生成、视频生成拆分为数十个独立任务，虽然方便产品定义，却限制了用户自由组合素材的能力。H3 的研究则将重点转向了任务统一和泛化，在预训练阶段就混合了图像、视频、音频以及多种参考和编辑数据，将参考、保留和修改关系全部通过自然语言描述，而非限制在少数预定义任务中，使模型借助语言理解开放式创作意图，从而突破固定任务范式的局限。未来演进方面，MiniMax 表示后续 H 系列将进一步融合 M 系列模型的能力，并持续扩大模型规模、提升画面分辨率和细节表现，推动视频生成技术在更多高要求场景中的落地。

测评结果显示，该模型在 Artificial Analysis 的文生视频有声榜单中排名第二，Elo 得分 1242 分，显著高于 Dreamina Seedance 2.0 720p 的 1225 分，略低于 Gemini Omni Flash 的 1245 分，展现出较强的竞争力。

图表 3：Artificial Analysis 的文生视频有声榜单

■ Text to Video Leaderboard (With Audio) Artificial Analysis

Added to the leaderboard in the last month:
MiniMax H3, Wan2.7-260612

Category: All Current models All models With Audio No Audio All Open weights Global Leaderboard Personal Leaderboard

Rank	Range	Creator	Model	Elo	95% CI	Samples	Released	API Pricing
1	1-2	Google	Gemini Omni Flash	1,245	-9/9	6,588	May 2026	\$6.00 /min
2	1-2	MiniMax	MiniMax H3	1,242	-11/11	5,878	Jul 2026	Coming soon
3	3	ByteDance Seed	Dreamina Seedance 2.0 720p	1,225	-7/7	13,334	Mar 2026	\$9.07 /min
4	4	Alibaba	Wan2.7-260612	1,163	-8/8	6,958	Jun 2026	\$9.00 /min
5	5	Alibaba-ATH	HappyHorse-13	1,158	-8/8	7,461	Jun 2026	\$9.90 /min

资料来源：智东西，华鑫证券研究

技术选择方面，Contextual Omni Representation（上下文全模态表示）作为实现多模态上下文理解的关键，用语言作为可泛化的连结同时描述目标视频、参考素材及其关系，而模型则通过定制模型和全模态理解管线，从而做到将大部分素材从约 10 万 Token 的推理压缩为平均约 4K Token，大幅降低上下文长度；H3-VAE 则通过提高视觉信息压缩率，带来 4 倍的序列长度收益，显著降低训练和推理成本，并为 2K 分辨率输出提供技术支持；架构采用更通用的 H3-Omni Transformer，同时采用异构训练架构并优化样本间负载均衡，将模型端到端训练吞吐提升近 30%；通过 In-context Regeneration（上下文再生），模型基模结合原始多模态上下文重新生成高分辨率画面，无需独立超分辨率模块，基模已有能力得以复现，更好还原小文字和局部细节。

在定价方面，官方称新模型有能力提供行业内最优的性价比：在 2K 分辨率下每秒价格不到主流模型的三分之一，768P 分辨率下不到二分之一，具体价格暂未公布。此外，MiniMax 表示将开放 H3 的模型权重，以推动开源生态建设和国产芯片适配，进一步降低开发者的使用门槛。

在应用场景方面，除基础案例电影片头、游戏 UI、动态海报以及广告电商等场景都具备较好的生成效果。其中，游戏 UI 场景，角色从装备选择，到配置确认、界面加载，整个交互过程衔接流畅，反馈自然；同时在广告展示中，不同镜头之间的光影、服装和配色保持一致，镜片反光和金属材质表现较为突出，已初步具备商业广告的基本质感。

图表 4：MiniMax H3 的应用案例（游戏 UI、电影片头）



资料来源：MiniMax 稀宇科技，华鑫证券研究

2、AI 应用动态：文心一言周时长环比+13.95%，Kimi K3 正式开源

2.1、周流量跟踪：文心一言周时长环比+13.95%

本期（2026.7.24-2026.7.30）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1256.0M）、Bing（873.8M）和 Gemini（600.5M），访问量环比增速第一为 NotionAI（4.05%）；平均停留时长前三位分别为 Character.AI（00:16:33）、Discord（00:11:13）和 NotionAI（00:09:01）；平均停留时长环比增速第一为文心一言（13.95%）。

图表 5：2026.7.24-2026.7.30 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	1256.0	1.62%	6:36	1.02%
Bing	搜索	微软	873.8	1.68%	7:18	0.46%
Gemini	聊天机器人	谷歌	600.5	0.10%	7:01	-0.24%
Canva	在线设计	Canva	185.2	0.54%	5:58	-0.28%
Discord	游戏社区	微软	154.6	2.59%	11:13	-0.30%
Github	代码托管	微软	147.3	1.03%	6:24	0.00%
Character.AI	聊天机器人	Character.AI	37.04	2.69%	16:33	1.12%
DeepL	翻译工具	DeepL	23.48	-1.10%	2:29	0.68%
Perplexity	AI 搜索	Perplexity	22.9	-1.04%	4:17	-0.39%
Kimi	聊天机器人	Moonshot AI	17.48	-14.44%	7:12	-3.14%
QuillBot	释义工具	QuillBot	7.03	0.41%	2:53	0.58%
NotionAI	文本/笔记	Notion	2.72	4.05%	9:01	5.46%
文心一言	聊天机器人	百度	0.119531	-55.32%	2:27	13.95%

资料来源：similarweb, 华鑫证券研究

2.2、产业动态：参数量达 2.8 万亿，Kimi K3 正式开源

2026 年 7 月 28 日，Moonshot AI 正式开源 Kimi K3，成为全球首个真正实现落地的、达到三万亿参数级别的开源模型。此次发布的内容不仅限于模型权重、技术报告，还包括其训练的关键 Infra 技术：MoonEP、FlashKDA 和 AgentEnv。

Kimi K3 是一个原生多模态的专家模型，总参数量高达 2.88 万亿，每个 Token 激活参数量为 1040 亿，在保持巨量参数的同时，极大地提升了计算效率。该模型原生支持高达 100 万 Token 的超长上下文窗口，并具备强大的推理期思考缩放能力，支持从 Low、High 到 Max 多种 Thinking Effort 级别。相比前代 K2，K3 的整体扩展效率提升了约 2.5 倍。其架构的核心创新体现在三个方面：首先，它采用了 KDA 与 Gated MLA 结合的混合注意力机制，通过将 Kimi Delta Attention (KDA) 和 Gated MLA 以 3: 1 的比例搭配，既高效处理了超长序列，又保留了精准的全局信息检索能力。其次，引入的注意力残差技术使得深层网络中的信息流转更为顺畅，减少了信息损耗。最后，其超稀疏的 MoE 架构包含 896 个专家，并配合 SiTU-GLU 激活函数和分位数负载均衡技术，在保证训练稳定性的同时，实现了近乎完美的算力分配。

图表 6: Kimi K3 预训练中视觉塔梯度范数对比

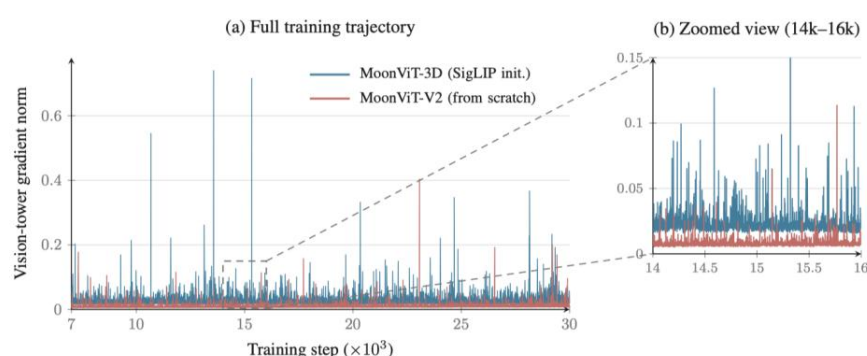


Figure 6: Vision-tower gradient norms in our pre-training ablations. Compared with the SigLIP-initialized MoonViT-3D, the from-scratch MoonViT-V2 maintains lower gradient norms with fewer spikes, indicating more stable optimization.

资料来源：新智元，华鑫证券研究

图表 7: Kimi K3 相比 Kimi K2 在整体扩展效率上实现约 2.5 倍的提升

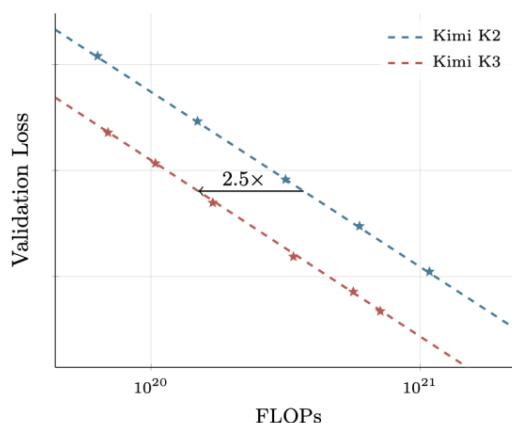


Figure 7: Fitted scaling-law curves for Kimi K2 and Kimi K3. Kimi K3 achieves 2.5x gain in scaling efficiency over Kimi K2.

资料来源：新智元，华鑫证券研究

Kimi K3 的强大性能不仅源于其创新的架构，更得益于为其百万级 Token 超长视距的智能体能力量身定制的后训练/RL 方案。其训练环境是一个基于微型虚拟机的复杂沙盒系统 AgentENV，能够模拟真实世界中的各种任务场景，让模型在模拟的协作环境中进行长周期、多步骤的复杂任务处理。研发团队还构建了基于知识图谱的任务合成系统，通过自动抓取

和生成涵盖多领域的高难度验证任务来不断提升模型能力。此外，Kimi K3 交叉训练了通用、智能体、代码等领域的专家模型，并在不同思考级别上进行强化学习，最终利用多教师策略蒸馏技术，将多种顶尖能力融合到一个统一的模型中。在基础设施层面，其 MoonEP 专家并行方案实现了完全动态且平衡的算力分配，而 KDA 融合 Prefix Cache 则通过底层引擎的重写，使得模型在处理包含数十万字相同前缀的请求时，响应速度得到指数级提升。

图表 8: Kimi K3 的知识图谱

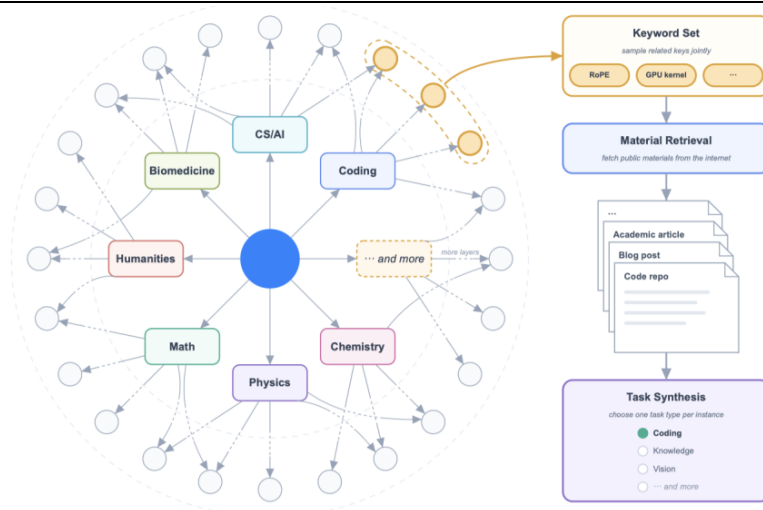


Figure 9: Overview of knowledge-graph-guided task synthesis. The hierarchically organized knowledge graph represents concepts at multiple levels, ranging from broad domains to fine-grained concepts. Related nodes are sampled to form a keyword set that guides the retrieval of publicly available source materials. For each synthesis instance, the system selects a task type and uses the retrieved materials to synthesize a corresponding task.

资料来源：新智元，华鑫证券研究

图表 9: Kimi K3 多层思考的蒸馏

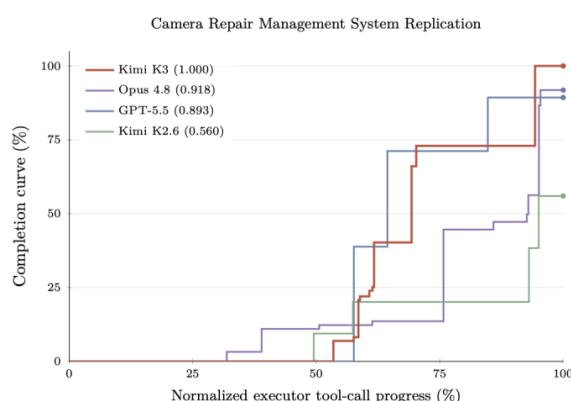
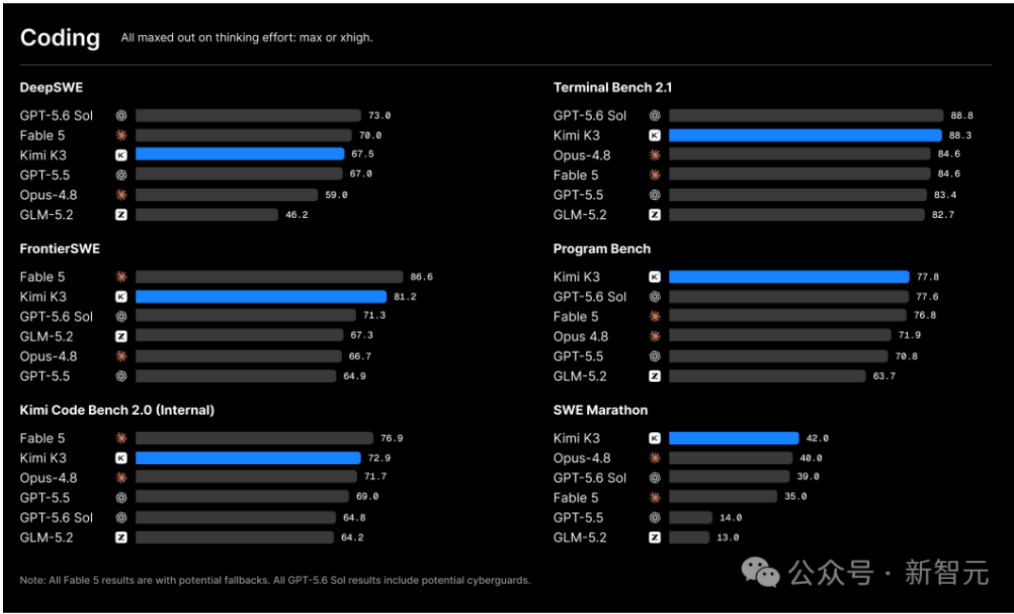


Figure 10: Completion curves on Camera Repair Management System, a black-box system replication task in which the agent reconstructs a hidden 3D-camera repair system as a web application through oracle queries. Completion denotes verifier-assessed task progress.

资料来源：新智元，华鑫证券研究

根据 Moonshot 官方及第三方评测，Kimi K3 的智能水平已进入全球第一梯队，在编程、代码 Agent、长上下文推理等多个关键场景中表现突出，全面超越多款主流闭源模型。例如，其能在真实极客任务中独立于 Linux 内核代码中发现未知的零日漏洞，可以开发出高效的 GPU 编译器，甚至在 48 小时内可以自主设计并验证了一款基于 45nm 工艺的 AI 推理芯片，展现了其在科研与工程领域的潜力。

图表 10: Kimi K3 在编程、代码 Agent、前端开发、长上下文推理等场景中的表现



资料来源：新智元，华鑫证券研究

3、AI 融资动向：Groundcover 完成 1 亿美元 C 轮融资，投后估值达 5 亿美元

2026 年 7 月 30 日，AI 云原生可观测性平台 Groundcover 宣布完成 1 亿美元 C 轮融资，投后估值达 5 亿美元。本轮融资由 One Peak 领投，Morgan Stanley Expansion Capital 及原有投资方 Zeev Ventures、Angular Ventures、Heavybit、Jibe 共同参投。当前，公司累计融资总额已达 1.6 亿美元，本轮融资将主要用于加速北美市场的商业化拓展，推进国际市场布局，并重点强化与云服务提供商的联合销售合作。

该公司定位为 AI 时代的全栈可观测性平台，专为自主运营、大规模高保真遥测、数据隐私与控制而设计，旨在让团队在不牺牲数据完整性、控制权与成本效益的前提下，全面监控云端运行的一切。随着云原生技术与 AI 的快速普及，生产环境所需的遥测数据量每隔数月便会翻倍，然而多数传统可观测性工具按数据量收费，导致成本持续攀升，迫使团队不得不牺牲数据的完整性与可见性，对遥测数据进行采样、限流或孤岛隔离。对此，Groundcover 从定价模式与技术架构两端提供了系统性解法。

在技术与架构方面，平台主要采用自带云（BYOC）架构，确保数据私密、安全地存储于客户自有云环境中，赋予企业对遥测数据的完全控制权。在数据采集层面，平台基于 eBPF 技术将 Linux 内核转化为高效的可观测性数据采集机制，在隔离的沙箱环境中运行内核扩展，规避了传统内核改造的可靠性与安全风险，其自适应 eBPF 传感器无需修改代码，数分钟内即可完成部署。与此同时，平台结合原生 OpenTelemetry 支持，能够实现跨基础设施、应用及 AI 工作负载的全保真、无采样遥测数据采集，且无任何采样或基数限制。在定价模式上，平台采用与数据采集量脱钩的计费方式，使企业可对低优先级工作负载进行遥测采集而无需担忧成本失控，解决了可观测性成本与数据完整性之间互相矛盾的问题。

在产品演进方面，平台自主构建 Agent Mode，为工程师提供在生产环境中发现、排查和修复问题的智能体体验，将可观测性从被动的监控工具升级为主动的运维智能体。此外，公司计划通过本轮融资重新定义工程师与 AI 智能体之间的协作方式，推动生产运营实现真正的自主化。当前，公司年度经常性收入实现三倍增长，平台已拥有超过 250 家企业用户，其中多家在过去一年内签署了七位数金额的合同，充分验证了其企业级市场的认可度与商业模式的可持续性。

图表 11：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	目前估值
					融资额	
Groundcover	AI 云原生	One Peak	C 轮	1 亿美元	1.6 亿美元	5 亿美元

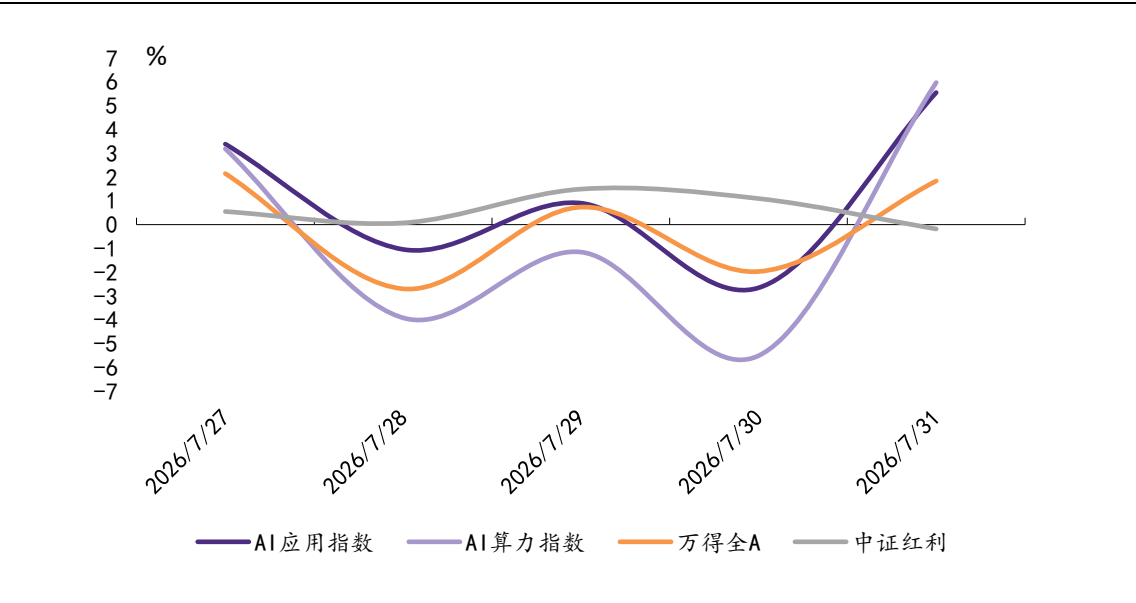
Enigma	具身智能	Index Ventures、 Ribbit Capital	种子轮	7100 万美 元	7100 万美 元	未披露
Fish Audio	AI 语音	Coreline Ventures、Capital Today	种子轮	5200 万美 元	5200 万美 元	未披露

资料来源：wind，Saasverse，华鑫证券研究

4、行情复盘

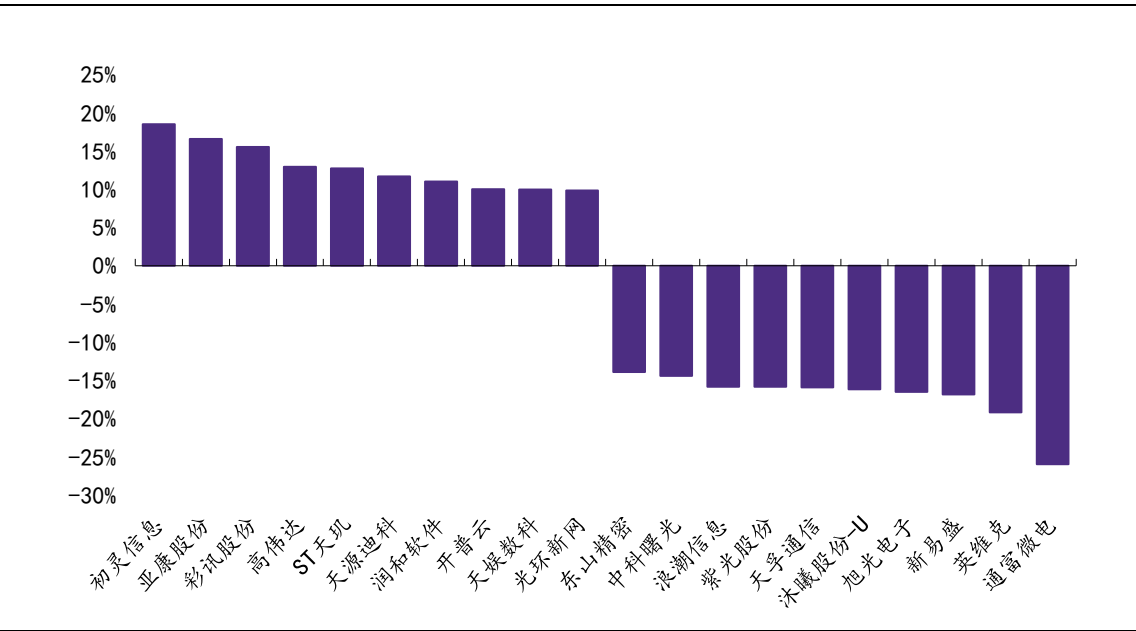
上周（2026.7.27-2026.7.31），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 5.55%/5.97%/2.15%/1.49%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-2.66%/-5.51%/-2.7%/-0.18%。AI 算力指数内部，初灵信息以 18.51% 录得上周最大涨幅，通富微电以-25.98%录得上周最大跌幅。AI 应用指数内部，中文在线以 34.14%录得上周最大涨幅，通富微电以-25.98%录得上周最大跌幅。

图表 12：上周（2026.7.27-2026.7.31 日）指数日涨跌幅



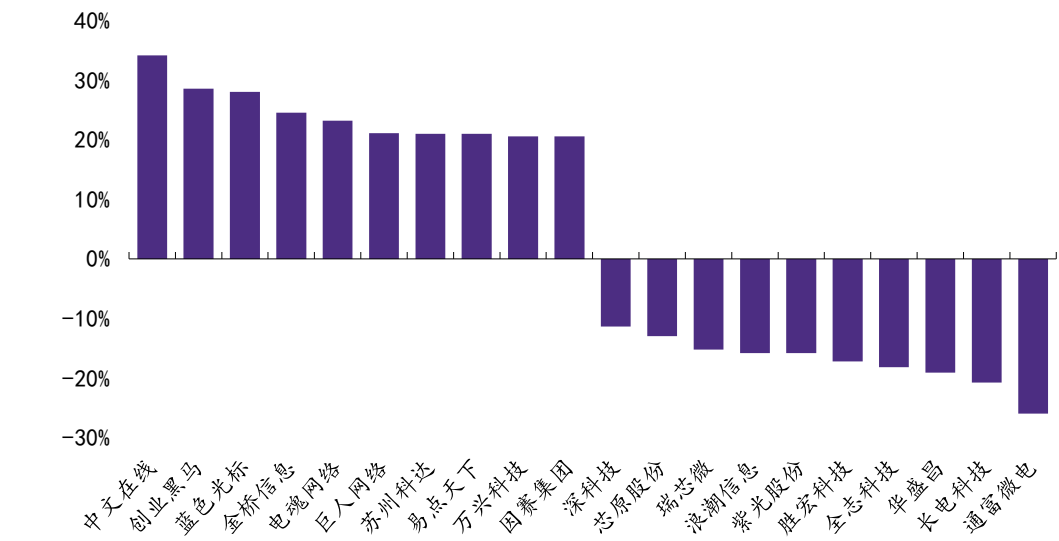
资料来源：wind, 华鑫证券研究

图表 13：上周（2026.7.27-2026.7.31 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

图表 14：上周（2026. 7. 27-2026. 7. 31 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

5、投资建议

2026 年 7 月 29 日，微软发布 2026 财年第四财季（截至 6 月 30 日）财报：当季实现总营收 900 亿美元，同比增长 18%；净利润 358 亿美元，同比增长 31%；摊薄后每股收益 4.81 美元，同比增长 32%。分业务看，智能云部门当季营收 393 亿美元，同比增长 32%，其中 Azure 及其他云服务营收增长 43%，较上季度提速 3 个百分点；公司首次披露 Azure2026 财年全年营收突破 1000 亿美元；商业剩余履约义务（RPO）达 6780 亿美元，同比增长 84%。生产力与业务流程部门营收 378 亿美元，同比增长 14%；Microsoft365Copilot 付费席位突破 3000 万，较三个月前披露的约 2000 万增长约 50%；Microsoft365 商用云营收经调整后增长 16%。更多个人计算部门营收 129 亿美元，同比下降 4%，其中 WindowsOEM 和设备营收下降 7%，Xbox 内容和服务营收下降 10%，搜索广告收入（不含流量获取成本）增长 10%。资本开支方面，公司单季资本开支 410 亿美元，同比增长约 70%；2026 财年全年资本开支 1453 亿美元；公司此前预计 2026 自然年资本开支约 1900 亿美元，较 2025 年增长 61%，公司解释称其中约 250 亿美元来自芯片短缺导致的内存和组件成本上升；大规模投入下公司现金流仍保持正值。

本次财报印证了 AI 算力需求的结构性高景气与 CSP 资本开支向硬件产业链的强传导。需求端看，Azure 云的大幅增速、6780 亿美元商业剩余履约义务及 Copilot 付费席位单季约 50%的环比跃升表明企业级 AI 需求仍在加速；管理层在电话会上明确表示，当季商业 RPO 的全部环比增量均由前沿模型公司以外的客户承诺贡献，印证 AI 算力需求正从少数基础模型公司向更广泛的企业级市场扩散，需求结构更趋多元。从传导路径看，我们推测 2026 自然年约 1900 亿美元的资本开支将直接转化为上游采购需求；电话会披露，当季 Azure 收入超预期主要得益于 CPU 和 GPU 资源效率提升及流程改进带来的新增容量更快交付，且本季度新增容量已迅速实现收入（quickly monetized），资本投入与收入确认之间的时间差显著缩短；管理层也重申客户需求持续超过可用容量，当前资本开支本质上是为已签署合同扩充交付能力，具备明确支撑。展望层面，公司指引下一季度 Azure 按固定汇率计算收入增长约 45%，并预计财年上半年（对应 2026 自然年下半年）增长将继续加速，表明算力供给缺口仍在扩大而非收敛。我们认为，本轮资本开支上行由真实需求驱动：积压订单高增、新增容量当季变现、后续指引加速三者共同表明算力供给仍滞后于需求增长，叠加芯片短缺推高的内存与组件成本为全行业共性压力，资本开支具备扎实的基本面支撑与持续性；对国内市场而言，微软资本开支的持续性为上游硬件供应链提供了较高的订单能见度，AI 服务器、数据中心与网络设备、IDC 及电力配套等环节的配置价值凸显。

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 15: 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
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 16：重点关注公司及盈利预测

公司代码	名称	2026-08-04	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	410.80	-0.30	0.30	0.60	-1369.33	1369.33	684.67	买入
301196.SZ	唯科科技	62.05	2.53	3.34	3.98	24.53	18.58	15.59	买入
301369.SZ	联动科技	109.41	0.48	1.08	2.62	227.94	101.31	41.76	买入
603859.SH	能科科技	40.10	0.92	1.05	1.32	43.59	38.19	30.38	买入
688615.SH	合合信息	114.10	3.24	4.22	5.25	35.22	27.04	21.73	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。

■ 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。

■ 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。