2026 年 01 月 06 日

华鑫证券 CHINA FORTUNE SECURITIES

小红书 Video-Thinker 打破工具依赖，DeepSeek 推出 mHC

—计算机行业周报

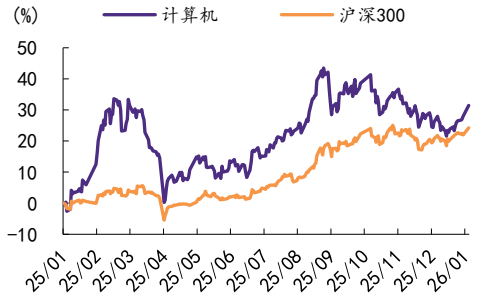
推荐(维持)

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)	4.3	-5.1	35.0
沪深 300	2.9	1.7	25.2

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：MiniMax 发布 MiniMaxM2.1，清华大学发布 TurboDiffusion》2025-12-31

2、《Monday.com(MNDY.O)：AI 多产品矩阵超预期贡献，获客渠道转型筑基长期增长》2025-12-31

3、《ServiceNow(NOW.N)：智能体 AI 驱动平台价值重塑，业绩指引上修》2025-12-31

投资要点

算力：算力租赁价格平稳，小红书 Video-Thinker 打破工具依赖

小红书推出 Video-Thinker 视频推理模型，打破行业外部工具依赖。其通过创新数据集构建与“SPO+GRPO”两阶段训练，实现核心能力内化。7B 版本在双重评测中确立 SOTA，关键基准准确率领先显著，核心指标大幅提升，性能优于外部工具方案，为视频推理开辟内生智能新范式。

AI 应用：Character.AI 周访问量环比+8.32%，DeepSeek 推出流形约束超连接

2026 年 1 月 1 日，由 DEEPSEEK 公司发布、其创始人梁文锋署名的一篇学术论文，正式公开了一种名为 mHC（流形约束超连接）的全新架构设计。这一方法能够在参数量为 27B 的模型上，仅引入约 6.7%的额外训练时间开销，即可实现显著的性能提升，标志着在提升模型能力与效率的探索道路上迈出了关键一步。

AI 融资动向：Kimi 完成 35 亿 C 轮融资，Meta 收购 Manus

2025 年 12 月 31 日，中国大模型企业月之暗面（Kimi）宣布完成 5 亿美元（约 35 亿元人民币）C 轮融资，阿里、腾讯、王慧文等投资方超额认购，投后估值达 43 亿美元。本轮融资后，Kimi 现金持有量超 100 亿元，资金将重点用于 K3 模型研发与显卡扩增，同时加大 2026 年人才激励力度。2025 年，Kimi 凭借 K2 系列开源模型实现技术突破，相关产品在核心基准测试中超越多款顶尖闭源模型，获全球行业领袖认可；商业化方面，依托 Agent 功能产品矩阵，公司 C 端付费用户月环比增长超 170%，海外 API 收入增长 4 倍，海外业务成为核心增长引擎。未来，Kimi 将聚焦 AGI 研发，目标 2026 年推出的 K3 模型追平全球前沿水平，同时聚焦 Agent 赛道实现营收量级增长。

美国当地时间 2025 年 12 月 29 日，Meta 宣布以 40 亿—50 亿美元收购中国创业公司研发的通用型 AIAgent 产品 Manus。这款 2025 年 3 月发布的产品，凭借“直接交付任务结果”的核心优势，上线 8 个月年度经常性收入突破 1 亿美元，年化运行率超 1.25 亿美元。收购谈判仅耗时十余

天，核心源于双方战略高度契合，Meta 正加速 AI 应用层布局以推进“超级智能”愿景。“90 后”创始人肖弘将出任 Meta 副总裁，掌舵相关技术落地，交易后 Manus 将在公司、团队及产品层面保持独立运营，技术将应用于 Meta 消费级与企业服务领域。

投资建议

2025 年 12 月，Meta 正式宣布完成对通用自主 AI 智能体公司 Manus 的数十亿美元收购交易，该笔交易位列 Meta 自成立以来的第三大收购案。交易完成后，Manus 将保持相对独立的运营模式，总部仍设于新加坡，继续沿用现有产品路线及服务体系。Manus 创始人、首席执行官肖弘将出任 Meta 副总裁，全面负责通用 AI 智能体全球业务板块。Meta 明确规划，将逐步推动 Manus 具备“端到端执行能力”的智能体系统，与面向消费者及企业的核心平台实现深度整合。公开资料显示，Manus 成立于 2022 年，自 2025 年 3 月核心产品正式上线后，业务规模实现快速扩张，年化营收突破 1.25 亿美元；技术层面，其累计处理 Token 规模超 147 万亿个，可支持逾 8000 万台虚拟计算机稳定运转。该公司在复杂任务执行领域积淀的核心技术实力，以及已验证的成熟商业化成果，构成此次收购交易达成的关键基础。此次收购亦是 Meta 自 2025 年启动 AI 业务重组、设立超级智能实验室并布局新一代模型体系以来的核心战略举措。其核心目标在于加速人工智能技术在商业自动化与生产系统场景的落地应用，进一步凸显下一阶段 AI 的价值核心——即更多取决于能否在现实场景中实现可靠、规模化的任务闭环能力。Meta 此次收购，是其深耕 AI 应用层、扩张生态的关键举措，核心目标为加码通用 AI 智能体赛道。此前，Meta 已通过成立超级智能实验室（MSL）、战略投资 ScaleAI、架构性重组 AI 业务等动作完善生态布局，此次收购进一步补齐其应用端执行能力的短板。通过本次战略动作，Meta 将 Manus 的端到端通用智能体技术与规模化商业落地能力纳入体系，强化了在复杂任务自动化赛道的卡位优势。Manus 智能体在多场景执行效率、商业化变现速度上具备显著优势，与 Meta 现有消费级及企业级产品生态形成互补，进一步夯实其从模型到应用的全栈护城河。此次收购印证了 AI 应用板块的高景气度，凸显通用智能体作为下一代应用形态的战略价值。龙头企业通过生态整合持续抬高行业技术落地与规模化运营门槛，我们维持对 AI 应用板块的长期看好态度。

中长期，建议关注加快扩张算力业务的精密零部件龙头迈信林（688685.SH）、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	2026-01-06	EPS			PE			投资评级
		股价	2024	2025E	2026E	2024	2025E	2026E	
301196.SZ	唯科科技	79.86	1.76	2.53	3.34	45.38	31.57	23.91	买入
603859.SH	能科科技	44.22	0.78	0.96	1.18	56.69	46.06	37.47	买入
688615.SH	合合信息	246.20	4.01	3.37	4.11	61.40	73.06	59.90	买入
688685.SH	迈信林	55.58	0.31	1.64	2.26	179.29	33.89	24.59	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，小红书 VIDEO-THINKER 打破工具依赖 6
1.1、 Tokens 跟踪 6
1.2、数据跟踪：算力租赁价格平稳 7
1.3、产业动态：小红书 Video-Thinker 打破工具依赖，刷新视频推理 SOTA 7
2、 AI 应用动态：CHARACTER.AI 周访问量环比+8.32%，DEEPSEEK 推出流形约束超连接 10
2.1、 周流量跟踪：Character.AI 周访问量环比+8.32% 10
2.2、 产业动态：DeepSeek 推出流形约束超连接，引领 AI 架构效率新突破 10
3、 KIMI 完成 35 亿 C 轮融资，META 收购 MANUS 14
4、 行情复盘 16
5、 投资建议 18
6、 风险提示 18

图表目录

图表 1：TOKENS 规模 LEADERBOARD 示意 6
图表 2：市场份额占据示意 6
图表 3：上周算力租赁情况 7
图表 4：HINDSIGHT-CURATION 驱动的思维链构建示意图 7
图表 5：模型在视频推理数据集上表现对比结果示意 8
图表 6：模型与使用外部工具的视频推理方法在性能上的比较图 9
图表 7：2025.12.25-2025.12.31AI 相关网站流量 10
图表 8：残差连接范式示意图 11
图表 9：MHC 的计算过程，其中 Σ(·)表示 SIGMOID 函数。 11
图表 10：MHC 的通信－计算重叠机制 12
图表 11：MHC 的通信－计算重叠机制流形约束超连接（MHC）的训练稳定性，展示了(A)MHC 与 HC 相对于基线的绝对训练损失差距，以及(B)三种方法的梯度范数。 12
图表 12：27B 模型的系统级基准测试结果 12
图表 13：MHC 相比基线的扩展特性 13
图表 12：上周 AI 初创公司融资动态 15
图表 15：上周（2025.12.29-2026.1.02 日）指数日涨跌幅 16
图表 16：上周（2025.12.29-2026.1.02 日）AI 算力指数内部涨跌幅度排名 16
图表 17：上周（2025.12.29-2026.1.02 日）AI 应用指数内部涨跌幅度排名 17

图表 18：重点关注公司及盈利预测 18

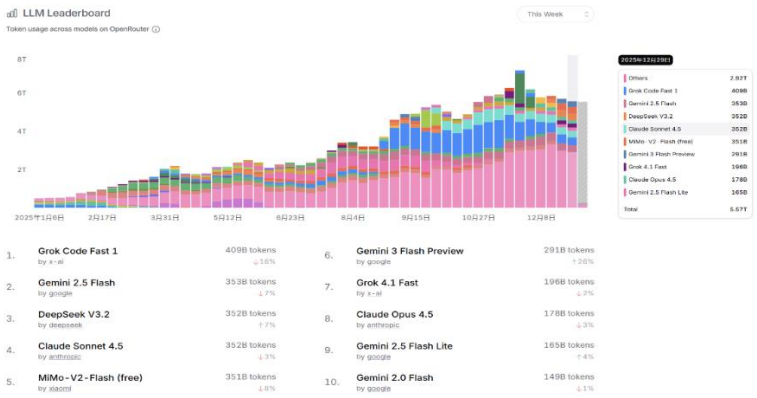
1、算力动态：算力租赁价格平稳，小红书 Video-Thinker 打破工具依赖

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2025 年 12 月 22 日至 12 月 28 日，周度 token 消耗量有所上升，调用量为 5.57T，环比上周-2.28%。在 tokens 规模 leaderboard 前五名中，xAI 旗下 GrokCodeFast1 以 409Btokens 位居榜首；Google 的 Gemini2.5Flash 以 353B 位列第二；DeepSeekV3.2 以 352B 排第三；Anthropic 的 ClaudeSonnet4.5 以 352Btokens 排第四；xiaomi 的 MiMo-V-Flash 以 351B 位列第五。

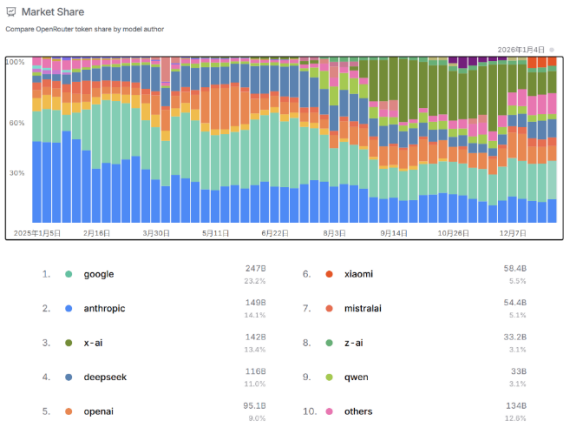
从市场份额维度来看，Google 以 247Btokens 占据 23.2%的份额，稳居首位；Anthropic 以 149Btokens 占比 14.1%，位列第二；xAI、Deepseek、OpenAI、则分别以 142B、116B、95.1Btokens，对应占据 13.4%、11.0%、9.0%的市场份额。

图表 1：tokens 规模 leaderboard 示意



资料来源：OpenRouter，华鑫证券研究

图表 2：市场份额占据示意



资料来源：OpenRouter，华鑫证券研究

1.2、数据跟踪：算力租赁价格平稳

上周算力租赁价格平稳。具体来看，显卡配置为 A100-40G 中，腾讯云 16 核+96G 价格为 28.64 元/时,阿里云 12 核+94GiB 价格为 31.58 元/时；显卡配置为 A800-80G 中，恒源云 16 核+256G 价格为 7.50 元/时。

图表 3：上周算力租赁情况

显卡配置	CPU	内存	磁盘大小（G）	平台名称	价格（每小时）	价格环比上周
A100-40G	16	96	可自定，额外收费	腾讯云	28.64/元	0.00%
	12 核	94G	可自定，额外收费	阿里云	31.58/元	0.00%
A100-80G	13	128	系统盘：20G 数据盘：50GB	恒源云	5.66/元	0.00%
A800-80G	16	256	系统盘：20G 数据盘：50GB	恒源云	7.50/元	0.00%

资料来源：腾讯云，阿里云，恒源云，华鑫证券研究

1.3、产业动态：小红书 Video-Thinker 打破工具依赖，刷新视频推理 SOTA

12 月底，小红书研究团队推出全新视频推理模型 Video-Thinker，以“ThinkingwithVideos”打破行业“工具依赖”困局，其 7B 参数量版本在 Video-Holmes 等多个高难度基准测试中刷新 SOTA 性能，为视频大模型的动态推理开辟了新路径。

针对现有视频推理方法往往受限于对外部工具的依赖或预设的提示词策略的痛点，Video-Thinker 提出将“时序定位（Grounding）”与“视觉描述（Captioning）”两大关键能力融入模型思维链（CoT），让模型在推理中自主定位关键帧、提取视觉线索。

Video-Thinker 的技术革新首先体现在高质量数据集的突破性构建上。针对现有开源视频数据集的二元割裂缺陷，团队整合 ActivityNet、STAR、LVBench 等六大主流数据集，创新推出 Hindsight-Curation 自动化流水线，构建了高质量样本的 Video-Thinker-10K 数据集。通过“双向信息补全”、“结构化思维链合成”、“后见之明验证”三步闭环，实现了时序定位、视觉描述与逻辑推理的深度融合，为模型内生能力培养奠定了坚实数据基础。

图表 4：Hindsight-Curation 驱动的思维链构建示意图



Figure 3: Data synthesis pipeline of Video-Thinker-10K where the data distribution is depicted in Figure 5 in Appendix B.

资料来源：机器之心，华鑫证券研究

其另一核心技术革新在于“监督微调（SFT）+组相对策略优化（GRPO）”的两阶段训练范式。SFT 阶段先通过<time>标签主动定位关键片段，再利用<caption>标签提取视觉细节，最后通过<think>标签进行逻辑整合，完成模型“冷启动”，强制模型习得“定位-感知-推理”的结构化思考范式，有效抑制无证据生成答案的幻觉倾向；GRPO 阶段则突破传统 PPO 算法的局限，通过对同一输入并行采样多组不同的推理轨迹，利用组内输出的相对优势来指导梯度更新。同时设计结果导向和过程约束的复合奖励函数，引入 KL 散度约束防止策略突变。这种训练模式不仅让模型将时序定位与视觉描述能力内化于思维链，涌现出顿悟时刻，使模型能主动修正推理偏差，真正摆脱外部工具依赖，实现视频内容的自主导航与深度推理。

为了全方位验证 Video-Thinker 的视频推理能力，研究团队构建了包含域内（In-Domain）与域外（Out-of-Domain）的双重评估体系。Video-Thinker-7B 在域内与域外双重评测体系中表现亮眼，成功确立 7B 参数量级模型的新 SOTA。评测涵盖 5 个域内测试集与 6 个高难度域外推理基准，对比了 5+12 个主流基础模型及视频推理模型，结果显示其域外泛化能力实现质变，Video-Holmes 榜单准确率达 43.22%，VRBench 综合基准准确率 80.69%，证明模型真正习得通用视频推理能力。帧数鲁棒性测试中，该模型在 16 帧、32 帧、64 帧不同输入设置下均持续压制基线模型，即便 16 帧低帧率条件下仍保持高推理精度，体现出高效的时序信息整合机制。深度归因分析表明，其时序定位平均交并比（mIoU）达 48.22%，较基础模型提升 75.5%，Recall@0.3 指标近乎基础模型两倍；视觉描述在三大核心指标上全面领先，较基础模型提升 31.2%，较 Video-R1 提升 61.0%，为推理提供了坚实信息基础。

图表 5：模型在视频推理数据集上表现对比结果示意

Table 1: Comparison of model performance on video reasoning datasets in both in-domain and out-of-domain settings. The best results are marked in **red bold** and the second best in blue.

Model	Out of Domain						In Domain				
	Video-Holmes	CG-Bench-Reasoning	VRBench	SciVideoBench	VdeoTT	VideoMME	ActivityNet	Star	ScaleLong	YouCook2	LVBench
Open-source Vanilla Models											
InternVL-2.5-8B	20.52%	19.39%	26.74%	15.50%	26.62%	29.89%	45.52%	49.85%	26.81%	40.84%	23.91%
InternVL-3-8B	18.67%	24.23%	41.14%	20.50%	28.42%	39.93%	48.56%	51.34%	29.34%	51.15%	25.93%
Qwen2.5-VL-7B-Instruct	34.02%	27.10%	63.42%	21.40%	34.57%	51.85%	70.96%	69.25%	40.06%	63.74%	33.33%
Qwen2.5-Omni-7B	29.99%	23.85%	49.04%	16.80%	36.27%	45.44%	63.92%	59.40%	36.91%	54.58%	31.65%
MiMo-VL-7B-RL-2508	28.96%	26.16%	57.94%	16.70%	34.15%	45.70%	49.84%	48.06%	35.33%	44.27%	28.62%
Open-source Reasoning Models											
Temporal-R1-7B	33.81%	25.27%	60.92%	20.80%	35.31%	51.63%	70.88%	70.15%	39.75%	63.74%	32.66%
Open-R1-Video-7B	21.83%	16.46%	50.15%	16.90%	18.66%	37.41%	55.76%	44.48%	31.86%	50.76%	26.94%
TW-GRPO-7B	33.32%	22.11%	53.46%	20.50%	33.93%	48.30%	70.00%	71.04%	39.12%	63.74%	29.97%
Video-R1-7B	38.54%	27.81%	69.25%	23.90%	41.04%	54.81%	76.00%	67.76%	47.32%	65.65%	34.68%
Time-R1-7B	34.73%	28.28%	66.48%	21.00%	34.78%	53.59%	72.00%	70.44%	44.47%	64.50%	32.65%
VideoChat-R1-7B	35.65%	29.26%	67.65%	22.80%	35.63%	54.41%	70.88%	**73.13%**	40.69%	69.08%	32.99%
VideoChat-R1-Thinking-7B	37.45%	29.44%	67.81%	20.30%	35.95%	54.15%	70.88%	71.64%	41.95%	66.79%	35.01%
GRPO-CARE-7B	34.34%	27.49%	66.39%	21.30%	35.74%	54.22%	70.96%	71.34%	40.69%	68.32%	33.33%
VersaVid-R1-7B	37.07%	28.58%	67.72%	21.80%	35.31%	54.78%	72.32%	71.94%	40.69%	66.79%	34.34%
VideoRFT-7B	24.39%	23.77%	61.54%	19.60%	37.65%	47.85%	50.72%	43.58%	36.59%	58.40%	26.94%
VR-Thinker-7B	25.37%	19.54%	53.43%	21.20%	31.07%	43.56%	55.36%	63.58%	32.18%	51.91%	30.98%
Video-RTS-7B	29.56%	18.09%	27.71%	20.30%	35.10%	42.63%	63.60%	65.07%	30.28%	65.27%	20.88%
SFT Model											
Video-Thinker-SFT-7B	31.52%	24.95%	62.40%	16.90%	33.93%	46.26%	70.80%	64.18%	43.22%	56.11%	35.69%
Pure RL Model											
Video-Thinker-Pure-RL-7B	30.70%	22.47%	63.06%	23.70%	36.37%	45.52%	64.24%	46.57%	41.96%	45.04%	30.64%
Our Model											
Video-Thinker-7B	**43.22%**	**33.25%**	**80.69%**	**26.30%**	**42.42%**	**54.96%**	**78.72%**	70.66%	**49.53%**	**73.66%**	**37.04%**

资料来源：机器之心，华鑫证券研究

消融实验进一步验证了模型的核心优势，简单外挂工具会导致性能“负优化”，即便调用 72B 参数量专家工具，准确率仍低于 7B 基础模型；成熟工具调用方法 VideoMind-7B 虽超越基础模型，但仍落后 Video-Thinker 约 4.2 个百分点。而 Video-Thinker 的内生思维链实现“感知-定位-描述-推理”无缝融合，展现压倒性优势。同时实验揭示，SFT 阶段仅负责规范模型思考格式，GRPO 强化学习才是性能飞跃的关键，使其在 Video-Holmes 和 VRBench 上分别提升 11.70%和 18.29%，两阶段训练成为提升复杂推理能力的有效路径。这一系列实测结果证明，视频推理能力无需依赖“大参数+大数据”堆砌，通过精准培养内生能力，7B 量级模型即可实现超越现有方案的性能，为视频推理领域开辟了新范式。

图表 6：模型与使用外部工具的视频推理方法在性能上的比较图

Table 4: Performance comparison of our model with video reasoning methods with external tool use. The best results are marked in **red bold** and the second best in blue.

Model	Video-Holmes	CG-Bench-Reasoning	VRBench
Base Model			
Qwen2.5-VL-7B-Instruct	34.02%	27.10%	63.42%
Base Model + Plug-and-play Tools			
Grounding: Temporal-R1-7B Captioning: SkyCaptioner-V1-8B	30.58%	22.80%	55.09%
Grounding: Qwen2.5-VL-7B-Instruct Captioning: Qwen2.5-VL-7B-Instruct	31.23%	24.05%	59.14%
Grounding: Qwen2.5-VL-72B-Instruct Captioning: Qwen2.5-VL-72B-Instruct	33.96%	25.99%	60.54%
Existing Tool-use Method			
VideoMind-7B	38.98%	31.99%	75.39%
Our Model (Endogenous)			
Video-Thinker-7B	**43.22%**	**33.25%**	**80.69%**

资料来源：机器之心，华鑫证券研究

Video-Thinker 的核心价值，在于打破了“视频推理必须依赖外部工具”的固有认知，通过“高质量数据合成+精准强化训练”的全链路设计，让 MLLM 真正实现内生“时序定位”与“片段描述”能力，实现了端到端的自主视频思考。

2、 AI 应用动态：Character.AI 周访问量环比+8.32%，DeepSeek 推出流形约束超连接

2.1、 周流量跟踪：Character.AI 周访问量环比+8.32%

本期（2025.12.25-2025.12.31）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1090.0M）、Bing（749.0M）和 Gemini（384.0M），访问量环比增速第一为 Character.AI（8.32%）；平均停留时长前三位分别为 Character.AI（00:18:04）、Discord（00:10:42）和 Kimi（00:08:43）;平均停留时长环比增速第一为 Gemini（1.16%）。

图表 7：2025.12.25-2025.12.31AI 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	1090.0	-3.45%	6:32	0.26%
Bing	搜索	微软	749.0	1.33%	7:53	-0.63%
Gemini	聊天机器人	谷歌	384.0	0.21%	7:17	1.16%
Discord	游戏社区	微软	151.9	3.19%	10:42	0.16%
Canva	在线设计	Canva	150.2	-8.86%	6:09	0.82%
Github	代码托管	微软	110.3	0.46%	6:14	-1.84%
Character.AI	聊天机器人	Character.AI	46.62	8.32%	18:04	-0.18%
Perplexity	AI 搜索	Perplexity	34.51	-5.76%	4:29	-0.37%
NotionAI	文本/笔记	Notion	30.88	-3.89%	7:59	-1.03%
DeepI	翻译工具	DeepL	22.37	-7.83%	2:31	-1.95%
QuillBot	释义工具	QuillBot	7.10	-8.25%	3:08	-1.57%
Kimi	聊天机器人	MoonshotAI	7.04	-14.80%	8:43	-1.13%
文心一言	聊天机器人	百度	0.91	-9.03%	2:50	-1.73%

资料来源：similarweb,华鑫证券研究

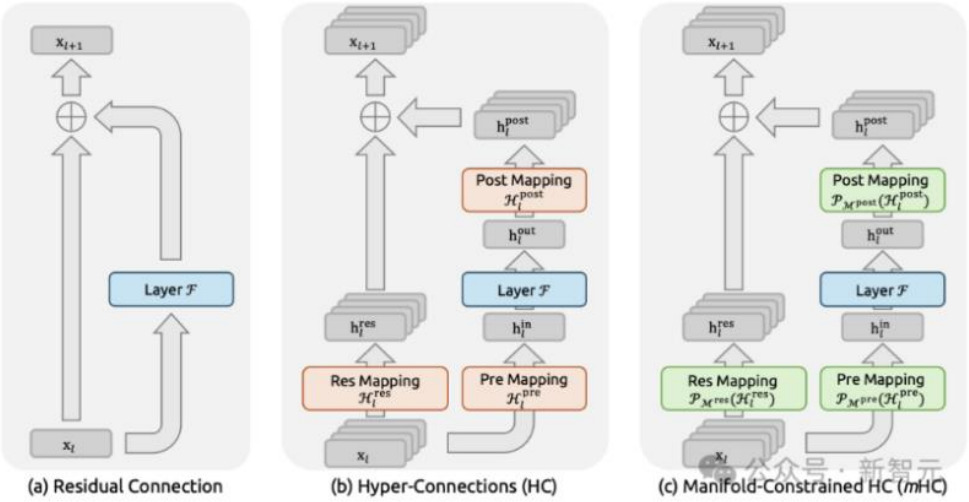
2.2、 产业动态：DeepSeek 推出流形约束超连接，引领 AI 架构效率新突破

2026 年 1 月 1 日，由 DEEPSEEK 公司发布、其创始人梁文锋署名的一篇学术论文，正式公开了一种名为 mHC（流形约束超连接）的全新架构设计。这一方法能够在参数量为 27B 的模型上，仅引入约 6.7%的额外训练时间开销，即可实现显著的性能提升，标志着在提升模型能力与效率的探索道路上迈出了关键一步。

DEEPSEEK 提出的 mHC 架构，正是针对传统残差流通道宽度瓶颈痛点的一次突破性尝试。其核心创新在于，将 Hyper-Connections 的参数空间约束到特定的流形（manifold）上，以

恢复恒等映射结构。具体而言，该方法要求矩阵元素非负，且每一行与每一列的元素之和均为 1。这种双随机性带来了多重理论优势：首先，该类矩阵的谱范数不大于 1，这意味着映射过程是非扩张的，有助于抑制梯度爆炸；其次，双随机矩阵在乘法运算下具有封闭性，确保了跨越多层的复合残差映射仍然是双随机的，从而在整个模型深度范围内保持稳定性；最后，该矩阵可视为一系列置换操作的凸组合，这提供了一种鲁棒的特征融合机制。

图表 8：残差连接范式示意图



资料来源：新智元，华鑫证券研究

在技术实现层面，mHC 架构对每一层参数的处理进行了精心设计。给定某一层的输入，模型会首先将其展平以保留完整的上下文信息，随后分别计算一个动态映射和一个静态映射，接着通过 Sigmoid 函数得到最终满足约束的映射。Sinkhorn－Knopp 算子首先通过指数运算保证所有元素为正，然后执行交替的迭代归一化过程，使矩阵的行和列分别归一到 1。尽管严格的归一化迭代过程需要无限迭代，但实验中进行约 20 次迭代即可获得一个实用的近似解，在效率与精度之间取得了良好平衡。

图表 9：mHC 的计算过程，其中 σ（•）表示 Sigmoid 函数。

$$\begin{cases} \hat{x}'_l = \mathrm{RMSNorm}(\hat{x}_l) \\ \tilde{H}^{\mathrm{pre}}_l = \alpha^{\mathrm{pre}}_l \cdot (\hat{x}'_l \phi^{\mathrm{pre}}_l) + b^{\mathrm{pre}}_l \\ \tilde{H}^{\mathrm{post}}_l = \alpha^{\mathrm{post}}_l \cdot (\hat{x}'_l \phi^{\mathrm{post}}_l) + b^{\mathrm{post}}_l \\ \tilde{H}^{\mathrm{res}}_l = \alpha^{\mathrm{res}}_l \cdot \mathrm{mat}(\hat{x}'_l \phi^{\mathrm{res}}_l) + b^{\mathrm{res}}_l, \end{cases}$$

$$\begin{cases} H^{\mathrm{pre}}_l = \sigma(\tilde{H}^{\mathrm{pre}}_l) \\ H^{\mathrm{post}}_l = 2\sigma(\tilde{H}^{\mathrm{post}}_l) \\ H^{\mathrm{res}}_l = \text{Sinkhorn-Knopp}(\tilde{H}^{\mathrm{res}}_l), \end{cases}$$

资料来源：新智元，华鑫证券研究

在基础设施设计层面，研究团队也实施了多层次的优化策略：首先，通过“内核融合”技术和混合精度策略，对高维隐藏状态进行除以范数的操作重新排序，在不牺牲计算速度的前提下最大化数值精度，并将多个具有共享内存访问模式的算子融合为统一的计算内核，显著减少了内存带宽瓶颈。其次，为了应对 n 路残差结构带来的内存开销，他们采用了巧妙

的重计算策略——在前向传播后丢弃 mHC 产生的中间激活，在反向传播需要时再即时重新计算。通过最小化与 L_r 对应的总内存占用确定了最优的重计算块大小。最后，在大规模训练中，团队扩展了现有的 DualPipe 调度策略。通过将部分计算内核置于独立的高优先级计算流，并避免使用长时间运行的持久化内核，以防止产生长时间的停顿。该设计允许对已重叠的注意力计算进行抢占，从而在保持计算设备处理单元高利用率的同时，实现更加灵活的调度。

图表 10：mHC 的通信 - 计算重叠机制

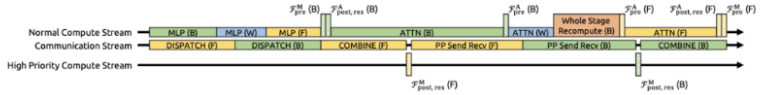


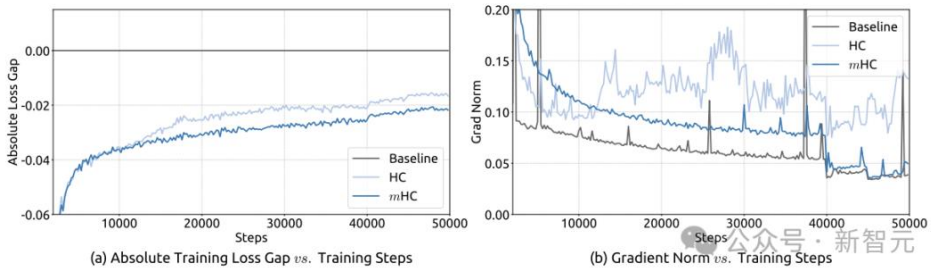
Figure 4 | **Communication-Computation Overlapping for *m*HC.** We extend the DualPipe schedule to handle the overhead introduced by *m*HC. Lengths of each block are illustrative only and do not represent actual duration. (F), (B), (W) refers to forward pass, backward pass, weight gradient computation, respectively. $\mathcal{F}^{A}$ and $\mathcal{F}^{M}$ represents kernels corresponded to Attention and MLP, respectively.

公众号 · 新智元

资料来源：新智元，华鑫证券研究

实验部分全面验证了 mHC 架构的训练稳定性和收敛性。在 27B 参数模型的训练稳定性测试中，mHC 成功解决了原始 HC 方法中观察到的训练不稳定性问题，其最终训练损失较基线模型降低了 0.021，并且在整个训练过程中展现出显著优于 HC 的稳定性，与基线相当。在下游任务性能评估中，mHC 模型在包括 BBH、DROP 等在内的多个基准测试上，其零样本和少样本性能均持续超越基线模型，并且在绝大多数任务上也优于 HC 模型。特别是在 BBH 和 DROP 任务上，相较于 HC 取得了超过 2 个百分点的额外性能提升，证明了其在大规模预训练中的有效性。

图表 11：mHC 的通信 - 计算重叠机制流形约束超连接（mHC）的训练稳定性，展示了(a)mHC 与 HC 相对于基线的绝对训练损失差距，以及(b)三种方法的梯度范数。



(a) Absolute Training Loss Gap *vs.* Training Steps　　(b) Gradient Norm *vs.* Training Steps

公众号 · 新智元

资料来源：新智元，华鑫证券研究

图表 12：27B 模型的系统级基准测试结果

Table 4 | **System-level Benchmark Results for 27B Models.** This table compares the zero-shot and few-shot performance of the Baseline, HC, and *m*HC across 8 diverse downstream benchmarks. *m*HC consistently outperforms the Baseline and surpasses HC on the majority of benchmarks, demonstrating its effectiveness in large-scale pre-training.

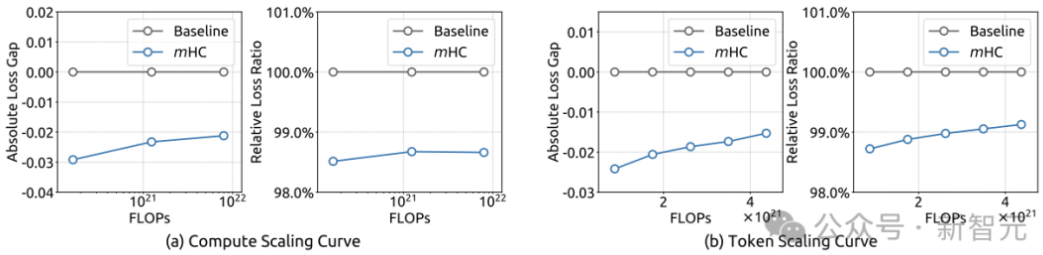
Benchmark (Metric)	BBH (EM)	DROP (F1)	GSM8K (EM)	HellaSwag (Acc.)	MATH (EM)	MMLU (Acc.)	PIQA (Acc.)	TriviaQA (EM)
# Shots	3-shot	3-shot	8-shot	10-shot	4-shot	5-shot	0-shot	5-shot
27B Baseline	43.8	47.0	46.7	73.7	22.0	59.0	78.5	54.3
27B w/ HC	48.9	51.6	53.2	74.3	**26.4**	63.0	79.9	56.3
27B w/ *m*HC	**51.0**	**53.9**	**53.8**	**74.7**	26.0	**63.4**	**80.5**	**57.6**

公众号 · 新智元

资料来源：新智元，华鑫证券研究

研究团队还深入分析了 mHC 的扩展特性。通过在不同模型规模——从 3B 到 27B 参数，和不同计算预算下的对比实验，结果表明 mHC 带来的性能增益在不同规模下均能稳健保持，在更高的计算预算下仅有轻微衰减，验证了 mHC 架构在大规模训练场景下的普遍有效性。综上所述，DEEPSEEK 提出的 mHC 架构，成功地将“扩展残差流宽度”这一理论设想推向了实用化，同时实现了增强模型处理能力和保证训练稳定性的双重目标。

图表 13：mHC 相比基线的扩展特性



资料来源：新智元，华鑫证券研究

3、AI 融资动向：Kimi 完成 35 亿 C 轮融资，Meta 收购 Manus

2025 年 12 月 31 日，中国大模型初创公司月之暗面（Kimi）宣布完成 5 亿美元 C 轮融资（约合 35 亿元人民币），阿里、腾讯、王慧文等投资方均超额认购，投后估值达 43 亿美元。同日，Kimi 创始人、CEO 杨植麟发布内部信，披露公司在技术突破、商业化落地等方面的显著进展。

Kimi 在 2025 年实现关键技术跨越，11 月推出的开源长思考模型 KimiK2Thinking 凭借交错推理与工具调用核心创新，支持 200-300 步连续工具调用，在 HumanitysLastExam（HLE）中取得 44.9%的成绩，超越 GPT-5、ClaudeSonnet 等顶尖闭源模型；K2 系列模型作为中国首个万亿参数基座模型和首个开源 AgenticModel，不仅在核心基准测试中达成行业顶尖水平（sota），独具特色的创意写作风格获得全球用户青睐，更首次成功采用二阶优化器完成大规模训练，已在硅谷及全球技术圈产生广泛影响，登上 Twitter 全球热搜，获得 Nvidia 创始人、Anthropic 联合创始人等行业领袖及诺奖得主的高度评价，硅谷创业加速器 FoundersSpace 创始人史蒂文·霍夫曼直言其新的 Agentic 模型“非常强大”。

技术突破直接驱动商业化进入爆发期，2025 年 5 月以来，Kimi 高频推出 Researcher、OKComputer、PPT、KimiCode 等多款 Agent 功能产品，功能矩阵持续完善，依托 K2 模型的顶尖表现，公司 C 端业务实现指数级增长，9-11 月海外及国内付费用户数平均月度环比增长（MoM）超 170%；目前 Kimi 的商业模式由 C 端订阅付费与 B 端 API 调用构成，其 API 工具调用能力在金融研报等复杂任务中具备显著优势，已成为多家国内金融研报 AI 工具的默认内置模型，本轮融资后公司现金持有量已超 100 亿元，为业务扩张提供充足保障。

本轮融资资金将重点用于“更加激进地扩增显卡，加速 K3 模型的训练和研发”，同时预留部分资金用于 2026 年激励计划和期权回购计划，2025 年 Kimi 已通过调薪、期权奖励等方式激励 324 人次，2026 年春节前将发放 K2Thinking 及后续产品专项奖励，全年平均激励力度预计为 2025 年的 200%，期权回购额度也将大幅上调；未来 Kimi 将以“超越 Anthropic 等前沿公司，成为世界领先的 AGI 企业”为核心目标，2026 年推出的 K3 模型将通过技术改进和规模扩容，使等效 FLOPs 提升至少一个数量级，预训练水平追平全球前沿模型，同时垂直整合模型训练与 Agent 产品体验，商业化方面聚焦 Agent 赛道，不追求绝对用户数量，着力突破智能上限、创造生产力价值，目标实现营收规模数量级增长；关于上市计划，杨植麟透露 KimiB/C 轮融资金额已超绝大部分 IPO 募资及上市公司定向增发规模，短期不急于上市，未来将把上市作为加速 AGI 发展的手段，择机而动。

同时在大洋彼岸的美国，当地时间 12 月 29 日，美国科技巨头 Meta 正式宣布收购中国创业公司研发的通用型 AIAgent 产品 Manus，交易金额高达数十亿美元，业内预估区间为 40 亿—50 亿美元。这款 2025 年 3 月才正式发布的产品，仅用 9 个月就完成了从亮相到被全球科技巨头收购的跨越式发展，其“90 后”创始人肖弘将出任 Meta 副总裁，掌舵相关技术在 Meta 旗下产品的落地应用，引发全球科技圈关注。

作为聚焦 AI 应用层的明星产品，Manus 的核心优势在于“直接交付任务结果”：用户只需一句口令，其智能体即可自动调用多种工具，完成写代码、制定旅行方案、数据分析等复杂任务，区别于传统聊天机器人仅提供答案的模式，精准契合全球 AI 行业“让 AI 真正执行任务”的核心需求。其市场表现同样亮眼，上线 8 个月年度经常性收入（ARR）即突破 1 亿美元，成为全球同类初创公司中增长最快的产品，目前总收入年化运行率超 1.25 亿美元，月度复合增长率保持 20%以上，Meta 创始人马克·扎克伯格亦为其长期用户。

创始人肖弘拥有十余年连续创业经历，始终聚焦“打造好用的工具”，避开基础大模型研发红海，深耕 AI 应用层创新。此次收购谈判仅耗时十余天，背后是 Manus 产品与 Meta 战略的高度契合——在基础大模型竞争中调整重心后，Meta 正加速布局 AI 生态与应用层，通过收购吸纳顶级应用人才与技术，推动“超级智能”愿景落地，Manus 的加入成为其该领域布局的关键落子。交易完成后，Manus 将在公司、团队及产品层面保持独立运营，其技术将同时应用于 Meta 消费级产品与企业服务领域。

图表 12：上周 AI 初创公司融资动态

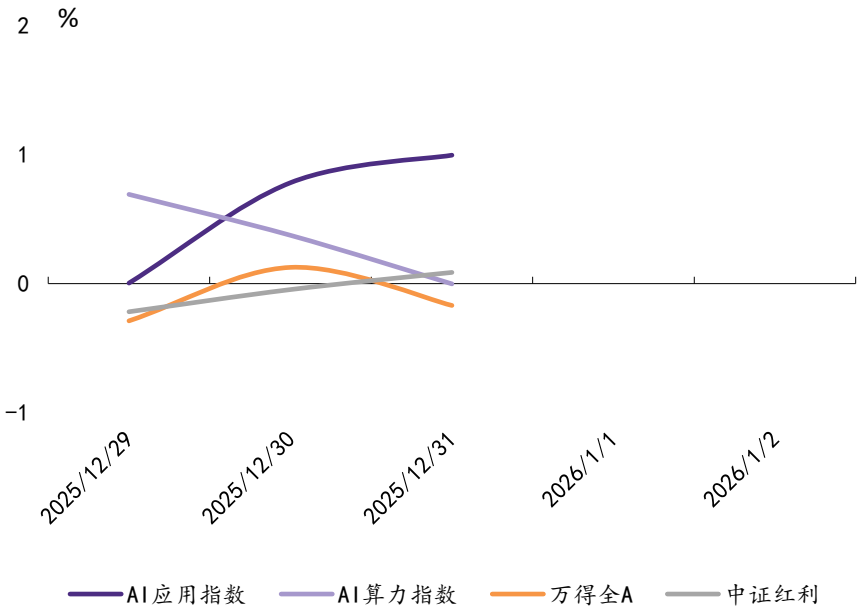
应用	应用类型	领投方	融资轮	融资额	目前累计融资额	目前估值
Kimi	LLM 大语言模型	IDG 资本	C 轮	5 亿美元	19800 万美元	43 亿美元
NeuralConcept	企业级 AI 工程设计平台	高盛另类投资部门（GrowthEquity）	C 轮	1 亿美元	13800 万美元	-
LemonSlice	AI 实时交互	MatrixPartners	种子轮	1050 万美元	1050 万美元	-

资料来源：Saasverse，wind，第一财经，华鑫证券研究

4、 行情复盘

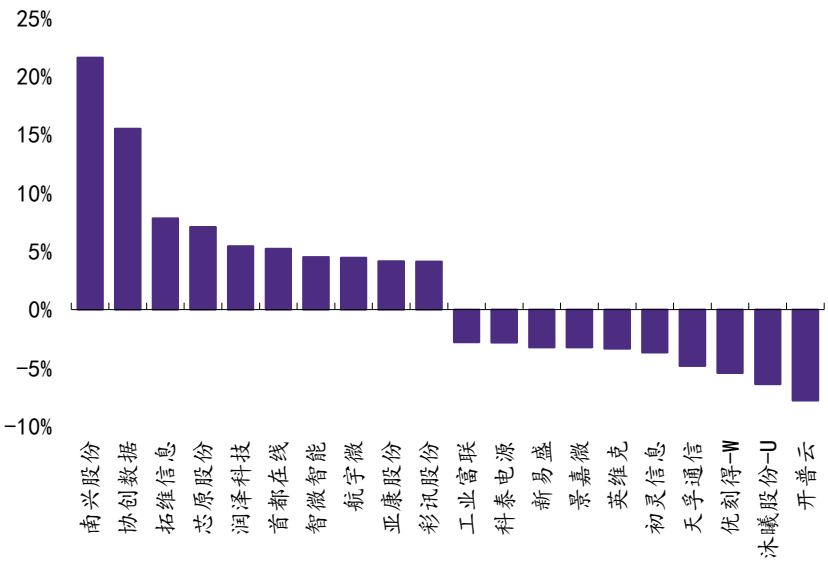
上周（2025.12.29-2026.1.02 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 0.99%/0.69%/0.13%/0.09%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为 0%/0%/-0.29%/-0.22%。AI 算力指数内部，南兴股份以+21.61%录得上周最大涨幅，开普云以-7.80%录得上周最大跌幅。AI 应用指数内部，蓝色光标以+30.02%录得上周最大涨幅，亿纬锂能以-8.41%录得上周最大跌幅。

图表 15：上周（2025.12.29-2026.1.02 日）指数日涨跌幅



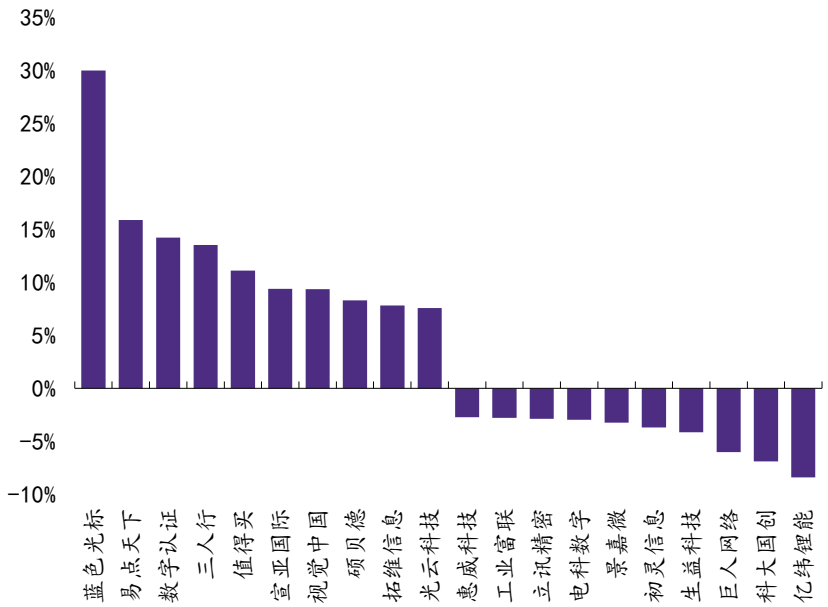
资料来源：wind，华鑫证券研究

图表 16：上周（2025.12.29-2026.1.02 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 17：上周（2025.12.29-2026.1.02 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

2025 年 12 月，Meta 正式宣布完成对通用自主 AI 智能体公司 Manus 的数十亿美元收购交易，该笔交易位列 Meta 自成立以来的第三大收购案。交易完成后，Manus 将保持相对独立的运营模式，总部仍设于新加坡，继续沿用现有产品路线及服务体系。Manus 创始人、首席执行官肖弘将出任 Meta 副总裁，全面负责通用 AI 智能体全球业务板块。Meta 明确规划，将逐步推动 Manus 具备“端到端执行能力”的智能体系统，与面向消费者及企业的核心平台实现深度整合。公开资料显示，Manus 成立于 2022 年，自 2025 年 3 月核心产品正式上线后，业务规模实现快速扩张，年化营收突破 1.25 亿美元；技术层面，其累计处理 Token 规模超 147 万亿个，可支持逾 8000 万台虚拟计算机稳定运转。该公司在复杂任务执行领域积淀的核心技术实力，以及已验证的成熟商业化成果，构成此次收购交易达成的关键基础。此次收购亦是 Meta 自 2025 年启动 AI 业务重组、设立超级智能实验室并布局新一代模型体系以来的核心战略举措。其核心目标在于加速人工智能技术在商业自动化与生产系统场景的落地应用，进一步凸显下一阶段 AI 的价值核心——即更多取决于能否在现实场景中实现可靠、规模化的任务闭环能力。Meta 此次收购，是其深耕 AI 应用层、扩张生态的关键举措，核心目标为加码通用 AI 智能体赛道。此前，Meta 已通过成立超级智能实验室（MSL）、战略投资 ScaleAI、架构性重组 AI 业务等动作完善生态布局，此次收购进一步补齐其应用端执行能力的短板。通过本次战略动作，Meta 将 Manus 的端到端通用智能体技术与规模化商业落地能力纳入体系，强化了在复杂任务自动化赛道的卡位优势。Manus 智能体在多场景执行效率、商业化变现速度上具备显著优势，与 Meta 现有消费级及企业级产品生态形成互补，进一步夯实其从模型到应用的全栈护城河。此次收购印证了 AI 应用板块的高景气度，凸显通用智能体作为下一代应用形态的战略价值。龙头企业通过生态整合持续抬高行业技术落地与规模化运营门槛，我们维持对 AI 应用板块的长期看好态度。

中长期，建议关注加快扩张算力业务的精密零部件龙头迈信林（688685.SH）、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 18：重点关注公司及盈利预测

公司代码	名称	2026-01-06	EPS			PE			投资评级
		股价	2024	2025E	2026E	2024	2025E	2026E	
301196.SZ	唯科科技	79.86	1.76	2.53	3.34	45.38	31.57	23.91	买入
603859.SH	能科科技	44.22	0.78	0.96	1.18	56.69	46.06	37.47	买入
688615.SH	合合信息	246.20	4.01	3.37	4.11	61.40	73.06	59.90	买入
688685.SH	迈信林	55.58	0.31	1.64	2.26	179.29	33.89	24.59	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&主题&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

陶欣怡：毕业于上海交通大学，于 2023 年 10 月加入团队。

倪汇康：金融学士，2025 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。