2026 年 06 月 02 日

华鑫证券 CHINA FORTUNE SECURITIES

Claude Opus 4.8 发布，小米 MiMo 大模型 API 永久降价

—计算机行业周报

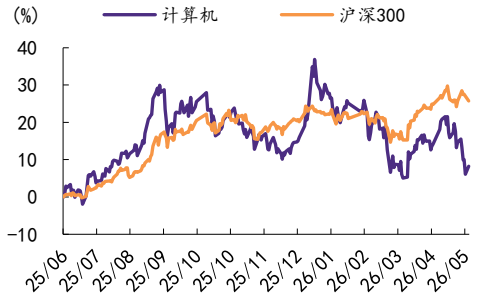
推荐(维持)

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)	-5.2	-11.5	8.5
沪深 300	0.8	2.4	26.1

市场表现



资料来源：Wind，华鑫证券研究

相关研究

1、《计算机行业点评报告：Symbotic（SYM）：Q2 营收增长势头强劲，调整后 EBITDA 同比翻倍》2026-05-31

2、《计算机行业点评报告：小马智行（PONY）：Robotaxi 商业化加速兑现，全球版图扩张与成本下探》2026-05-29

3、《计算机行业点评报告：Shopify（SHOP）：营收盈利结构优化，增值业务驱动利润高增》2026-05-28

投资要点

算力：算力租赁价格平稳，小米 MiMo 大模型 API 永久降价，最高降幅 99%

2026 年 5 月 27 日，小米宣布 MiMo-V2.5 系列 API 价格永久下调，取消上下文长度区分，执行统一定价。其中，MiMo-V2.5-Pro 输入缓存命中价格降至 0.025 元/百万 tokens，最高降幅达 99%。新版 API 价格直接对标 DeepSeek，国内大模型价格战升温。与此同时，小米 Token Plan 计费体系同步升级，在月费不变的前提下，各档位 Credits 额度普遍提升至原来的 5 至 8 倍。

AI 应用：Discord 周访问量环比+2.82%，Claude Opus 4.8 发布

2026 年 5 月 29 日，Anthropic 公司发布了 Claude Opus 4.8 模型。新模型在编程能力上取得了显著提升，几乎在所有主流基准测试中都占据了领先位置。与此同时，Anthropic 还展示了其 Claude Code 功能中的 dynamic workflows，该特性能够调动上百个 agent 并行处理任务，例如在 11 天内完成 75 万行代码的重写工作，且测试通过率高达 99.8%。

AI 融资动向：Cognition AI 完成超 10 亿美元 D 轮融资，投后估值达 260 亿美元

2026 年 5 月 28 日，AI 编程智能体公司 Cognition AI 宣布完成超 10 亿美元 D 轮融资，投后估值达 260 亿美元，本轮融资由 Lux Capital、General Catalyst、8VC 共同领投，Founders Fund、Elad Gil、Alpha Wave 等多家知名机构跟投。不到两年半时间内，该公司已实现超 70 倍估值增长。自今年年初以来，其核心产品 AI 智能体 Devin 的企业端使用量增长超过 10 倍，年化营收规模达 4.92 亿美元。

投资建议

2026 年 5 月 28 日，迈威尔科技发布 2027 财年第一季度业绩报告。本季度公司实现营收 24.2 亿美元，同比增长 28%，营收增量主要由数据中心业务拉动。盈利端，公司本季度基础毛利率为 52.15%，同比提升 1.9 个百分点。公司核心的数据中心业务本季度营收达 18.3 亿美元，同比增长 27%，增长主要由光互连产品驱动，该业务营收占比已达 76%。业绩预期方面，公司预计 2027 财年全年营收约 115 亿美元，同比增长约

40%；同时上调 2028 财年营收预期至 165 亿美元，较 2027 财年预计增长约 44%。中长期维度，公司确立 2029 财年定制芯片业务营收突破百亿美元的发展目标，目前已与美国全部超大规模云服务商开展定制芯片合作。整体来看，公司业绩增长依托 AI 定制 ASIC 芯片与光互连两大核心业务引擎，在定制 AI 芯片领域占据核心市场地位，与亚马逊、微软深度合作研发主流 AI 芯片。同时，公司凭借光互连领域的技术优势，适配 AI 数据中心升级迭代需求，叠加全球科技巨头持续加码 AI 基础设施投资，为企业长期稳健增长提供了坚实的行业支撑。

此次最新财报中，迈威尔科技光互连业务稳居核心增长支柱。作为全球高速光互连龙头，公司的业务高增是 AI 数据中心长期资本开支落地的直接佐证，充分验证高速光互连产品已是算力集群建设的刚需配套。公司上调中长期经营预期，叠加与全球头部云厂商的深度绑定合作，进一步印证下游需求具备高延续性，夯实了核心成长逻辑。当前全球 AI 基础设施建设持续推进，光通信作为算力网络的核心枢纽，在技术迭代升级与市场需求扩容的双向驱动下，行业向上趋势确立，长期景气度与发展空间持续向好。

中长期，建议关注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	2026-06-02 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E		
300757.SZ	罗博特科	595.98	-0.30	0.30	0.60	-1986.60	1986.60	993.30	买入
301196.SZ	唯科科技	118.50	2.53	3.34	3.98	46.84	35.48	29.77	买入
603859.SH	能科科技	43.90	0.92	1.21	1.50	47.72	36.28	29.27	买入
688615.SH	合合信息	136.62	3.24	4.22	5.25	42.17	32.37	26.02	买入

资料来源：Wind，华鑫证券研究

正文目录

图表目录

1、算力动态：算力租赁价格平稳，小米 MiMo 大模型 API 永久降价

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 5 月 25 日至 5 月 31 日，Token 周消耗量有所上升，调用量为 31.8T，环比上周增加 10.03%。在 Tokens 规模 Leaderboard 榜单中，DeepSeek 的 DeepSeek V4 Flash 以 3.11T tokens 位居榜首，Tencent 的 Hy3 preview 以 3.03T tokens 位居第二，Anthropic 的 Claude Opus 4.7 以 2.32T tokens 位居第三；Anthropic 的 Claude Sonnet 4.6 以 1.92T tokens 位列第四；OpenRouter 旗下的 Owl Alpha 以 1.66T tokens 位居第五；

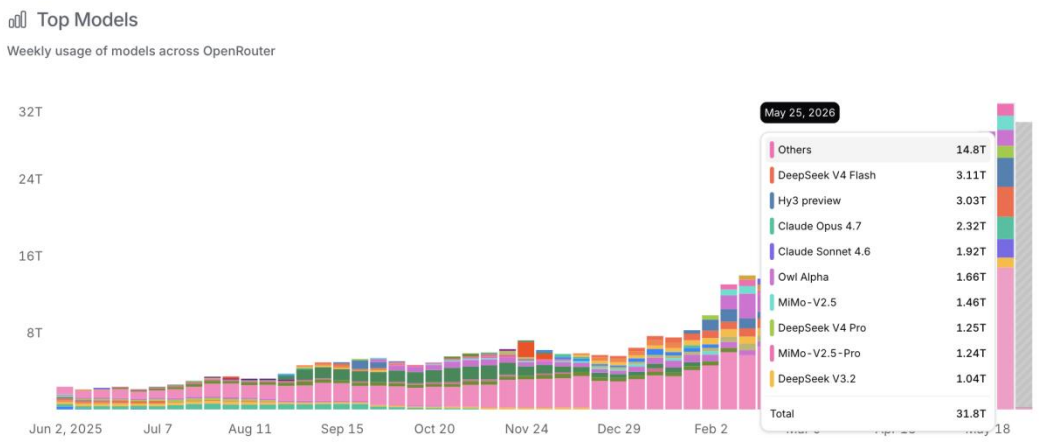
从市场份额维度来看，DeepSeek 以 5.6T tokens 占据 17.6%的份额，稳居首位；Anthropic 以 5.57T tokens 占据 17.5%，位列第二；Google、Tencent、OpenAI 则分别以 4.2T、3.03T、2.94T tokens，对应占据 13.2%、9.5%、9.2%的市场份额。

5 月 25 日，迅策科技发布全球首款 TokenOS 操作系统 TokenONE。该系统旨在将每一次的模型调用转化为可量化、可追溯的业务价值产出，让企业数据成为可以直接被模型调用的数据 Token，以推动各行业 Token 工厂大规模商业化落地。

5 月 26 日，蚂蚁集团旗下支付宝推出全球首个 Token Pay 服务，作为目前市场唯一的模型付款解决方案，该服务将连同 AI 付、AI 收与 AI 钱包，共同构成一套覆盖授权、支付、结算、管理、安全的全栈 AI 原生支付体系。

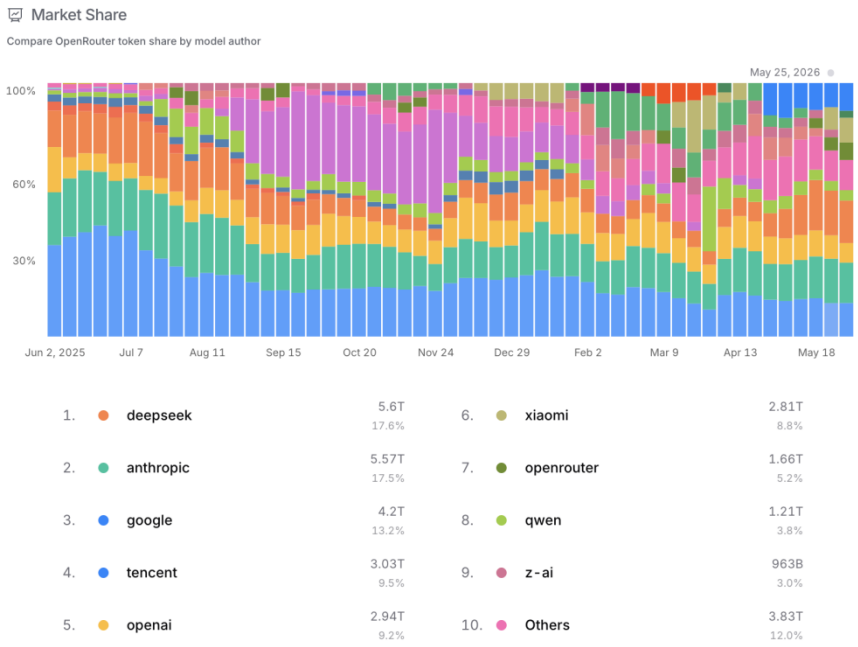
5 月 28 日，华为于数据通信创新峰会正式发布星河 AI 数据中心网络方案。该方案依托于独家的网算存协同技术，可使传输带宽提升 8 倍，Token 生产效率提升 2 至 5 倍，并且结合分钟级光链路脏污检测手段，确保 Token 稳定生产零中断。同等成本下，Token 产能显著提升。

图表 1：Tokens 规模 Leaderboard



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：阿里云发布全新海外 AI 产品官网 Qwen Cloud，AI 出海开启加速模式

2026 年 5 月 26 日，阿里云面向海外市场发布全新 AI 产品官网 Qwen Cloud。与传统云控制台中的 AI 子模块不同，该平台面向 Agent 时代，官方将其定位为全新的云产品入口。基于云服务产品主要消费者由人向 Agent 的转变趋势，该平台采用三入口设计：网站入口供开发者浏览、试用、比较模型并接入 OpenAI 兼容 API；Skills 入口将平台能力封装为 Agent 可读的标准化指令；CLI 入口为开发者和智能体提供稳定可重复的命令行操作层

围绕人类设计的界面和交互逻辑被重构，阿里云将模型服务、推理调用等核心能力封装为标准化的 Skills 和 CLI 工具，使智能体能够直接解析指令、自主掌握平台全部能力并按需调用。在模型方面，该平台汇聚了阿里千问系列模型，覆盖文本、视觉、音频、图像、视频等任务类型，并将持续引入更多第三方模型。

同时，大会面向海外发布了 Agent 产品 MuleRun。该产品支持多 Agent 并行工作、任务分解与协作，并内置代码生成、数据分析、文档处理、图像视频生成等场景的标准化 Skills 模块。同期的产品发布与更新还包括：智能体编程平台 Qoder 1.0 版本、桌面智能体 QoderWork 自定义工作台模式，以及数字员工产品 QoderWake 公测。

在生态层面，阿里千问大模型的海外合作持续提速。一方面，Fireworks AI 宣布与千问达成战略合作，并将于其平台提供千问模型的优化部署服务，为全球开发者和企业带来低延迟、低成本、高安全性和高可扩展性的推理和微调方案。另一方面，Hermes Agent 也将与千问围绕 Qwen3.7-Max 展开新一轮合作，在北美开发者社区，进一步扩大千问模型的覆盖度。

当前，海外市场对 AI 的需求持续旺盛，而 Agent 的爆发则让模型调用量和云资源消耗

呈指数级增长。为应对这一趋势，阿里云正在面向海外进行全栈升级，范围覆盖模型、入口、Agent 产品和云基础设施，其根本目的在于实现阿里云 AI 能力在全球范围内的无缝衔接。与此同时，近年来，阿里云持续加码全球化投入，加快基础设施、产品与技术的全球化拓展。目前，阿里云已在全球设有 31 个地域、101 个可用区，海外市场规模五年增长 20 倍，是中国最大、亚太第一的云服务商。

1.3、产业动态：MiMo 大模型 API 永久降价，最高降幅 99%

2026 年 5 月 27 日，小米正式宣布 MiMo-V2.5 系列 API 价格永久下调。此次调整，MiMo 取消了此前根据上下文长度制定的价格区分，无论 256K 还是 1M 上下文窗口，均执行统一价格。具体来看：MiMo-V2.5 输入缓存命中价格 0.02 元/百万 tokens，降幅最高 98%；未命中输入 1 元/百万 tokens，降幅最高 82%；输出 2 元/百万 tokens，降幅最高 93%。MiMo-V2.5-Pro 输入缓存命中 0.025 元/百万 tokens，降幅最高 99%；未命中输入 3 元/百万 tokens，降幅最高 79%；输出 6 元/百万 tokens，降幅最高 86%。

图表 3：MiMo-V2.5 系列 API 最新价格

资料来源：智东西，华鑫证券研究

此前不久，DeepSeek 刚刚宣布从 6 月 1 日起 DeepSeek-V4-Pro 模型 API 不再恢复原价，将价格永久降低至原价的 1/4，并把输入缓存命中价格进一步压到原价 1/10。相距不到一周，小米几乎直接对标 DeepSeek 做出价格调整，国内 API 价格战再度升温。

图表 4：DeepSeek-V4 系列与小米 MiMo-V2.5 系列 API 价格对比表

计费维度	DeepSeek		MiMo	
	v4-flash	v4-pro	V2.5	V2.5-Pro
输入（缓存命中）	0.02	0.025	0.02	0.025
输入（缓存未命中）	1.0	3.0	1.0	3.0
输出	2.0	6.0	2.0	6.0
上下文长度	1M	1M	256k-1M	256k-1M

资料来源：智东西，华鑫证券研究

此次调整，小米同步对 Token Plan 计费体系进行了升级。改版后的 Token Plan 在月费不变的前提下，各档位 Credits 额度普遍提升至原来的 5 至 8 倍。从当前国内主流大模型订阅制套餐来看，经过此次调整，小米 Lite 套餐在入门档方面与 Kimi、字节、阶跃星辰等厂商的最低档位接近，最低仍为腾讯混元 Hy 的 Lite 档，标价 28 元/月。高阶档方面，小米 Max 套餐标价远低于阿里尊享版的 1398 元/月、字节 Agent Plan Max 档的 950 元/月，及 MiniMax Ultra 极速版的 749.17 元/月。

图表 5：国内大模型订阅制套餐价格对比图

平台	套餐类型	套餐名称	月单价（元）	核心额度/权益
阿里	Coding Plan	Pro	200	每月至高90000次请求 (Lite套餐已停购/续费)
	Token Plan	标准版	198	25000 Credits/月
		高级版	698	100000 Credits/月
		尊享版	1398	250000 Credits/月
百度	Coding Plan	Lite	40	每月18000次请求 用量达Claude Pro 3倍
		Pro	200	每月90000次请求 用量达Claude Max 3倍
阶跃星辰	Flash	Mini (入门版)	38	5小时限额：100次Prompt (~1500次模型调用)
		Plus (进阶版)	78	5小时限额：约400次Prompt (~6000次模型调用)
		Pro (专业版)	155	5小时限额：约1500次Prompt (~22500次模型调用)
		Max (旗舰版)	555	5小时限额：约5000次Prompt (~75000次模型调用)

MiniMax	Token Plan（标准版）	Starter	24.17	600次模型调用/5小时 每周可用额度为「每5小时额度」的10倍
		Plus	40.83	1500次模型调用/5小时 每周可用额度为「每5小时额度」的10倍
		Max	99.17	4500次模型调用/5小时 每周可用额度为「每5小时额度」的10倍
	Token Plan（极速版）	Plus-极速版	81.67	1500次模型调用/5小时 每周可用额度为「每5小时额度」的10倍
		Max-极速版	165.83	4500次模型调用/5小时 每周可用额度为「每5小时额度」的10倍
		Ultra-极速版	749.17	30000次模型调用/5小时 每周可用额度为「每5小时额度」的10倍
小米	Token Plan	Lite	34.32	原：6000万 Credits/月 现：41亿 Credits/月 非高峰期0.8x消耗
		Standard	87.12	原：2亿 Credits/月 现：110亿 Credits/月 非高峰期0.8x消耗
		Pro	289.52	原：7亿 Credits/月 现：380亿 Credits/月 非高峰期0.8x消耗
		Max	579.92	原：16亿 Credits/月 现：820亿 Credits/月 非高峰期0.8x消耗
腾讯	Hy Token Plan	Lite	28	3500万 Tokens/月
		Standard	78	1亿 Tokens/月
		Pro	238	3.2亿 Tokens/月
		Max	468	6.5亿 Tokens/月
月之暗面	Kimi Code Plan	Andante	39	基础Agent额度 Kimi Code可调用
		Moderato	79	2倍Agent额度 Kimi Code 4倍额度
		Allegretto	159	4倍Agent额度 Kimi Code 20倍额度
		Allegro	559	10倍Agent额度 Kimi Code 60倍额度
智谱	Coding Plan	Lite	39.2	3x Claude Pro 用量额度
		Pro	119.2	5x Lite 用量额度 + Lite 全量权益
		Max	375.2	20x Lite 用量额度 + Pro 全量权益
字节	Agent Plan（方舟）	Small	38	20000 Agent燃料值/月
		Medium	190	100000 Agent燃料值/月
		Large	475	250000 Agent燃料值/月
		Max	950	500000 Agent燃料值/月
	Coding Plan	Lite plan	40	用量数倍于Claude Pro
		Pro plan	200	5倍Lite用量

资料来源：智东西，华鑫证券研究

与此同时，小米同步给出了高缓存命中场景下，不同套餐大致对应的实际 Token 规模。以高缓存命中场景（95%以上）为例：Lite 档（39 元/月）下，MiMo-V2.5 可用超 5 亿 Token，MiMo-V2.5-Pro 可用超 1.9 亿 Token；Standard 档（99 元/月）下，模型的 Token 可用量分别为 V2.5 的 13 亿以上和 V2.5-Pro 的 5 亿以上；Pro 档（329 元/月）下，模型的 Token 可用量分别为 V2.5 的 47 亿以上和 V2.5-Pro 的 18 亿以上；Max 档（659 元/月）下，模型的 Token 可用量分别为 V2.5 的 100 亿以上和 V2.5-Pro 的 40 亿以上。同时，小米特别指出，在 Agent 与代码类场景中，缓存命中率更高，实际可用 Token 数量会进一步增加。

图表 6：高缓存命中场景下不同套餐对应实际 Token 规模概览

资料来源：智东西，华鑫证券研究

降价背后是推理系统的持续优化。小米基于 SGLang HiCache 完整支持 SWA，将 KV Cache 在 GPU 显存、CPU 内存、SSD 等多级存储间的数据搬运量降至优化前的约 1/7，可缓存 Token 数量提升至约 5 倍。不仅如此，小米还同时优化了专家并行方案、输入长度分桶策略等机制，进一步提升集群输入吞吐能力。总体而言，其核心逻辑在于通过更激进的缓存命中策略和更高的推理吞吐效率，降低单位 Token 成本，与 DeepSeek 逻辑类似。

据不完全统计，近半年以来，已有至少五家国产大模型厂商对自家套餐体系进行过明显调整，但策略却有所不同。部分厂商如小米、DeepSeek，选择进一步降低定价，而另一部分厂商则开始缩减低价套餐、减少额度，整体价格有所上涨，其中，阿里、字节暂停了 Coding Plan 中低价套餐，智谱 2026 年一季度 API 调用定价提升了 83%。

2、AI 应用动态：Discord 周访问量环比+2.82%，Claude Opus 4.8 发布

2.1、周流量跟踪：Discord 周访问量环比+2.82%

本期（2026.5.22-2026.5.28）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1223.0M）、Bing（824.7M）和 Gemini（659.9M），访问量环比增速第一为 Discord（2.82%）；平均停留时长前三位分别为 Character.AI（00:14:15）、Discord（00:11:05）和 Kimi（00:08:29）;平均停留时长环比增速第一为文心一言（4.61%）。

图表 7：2026.5.22-2026.5.28AI 相关网站流量

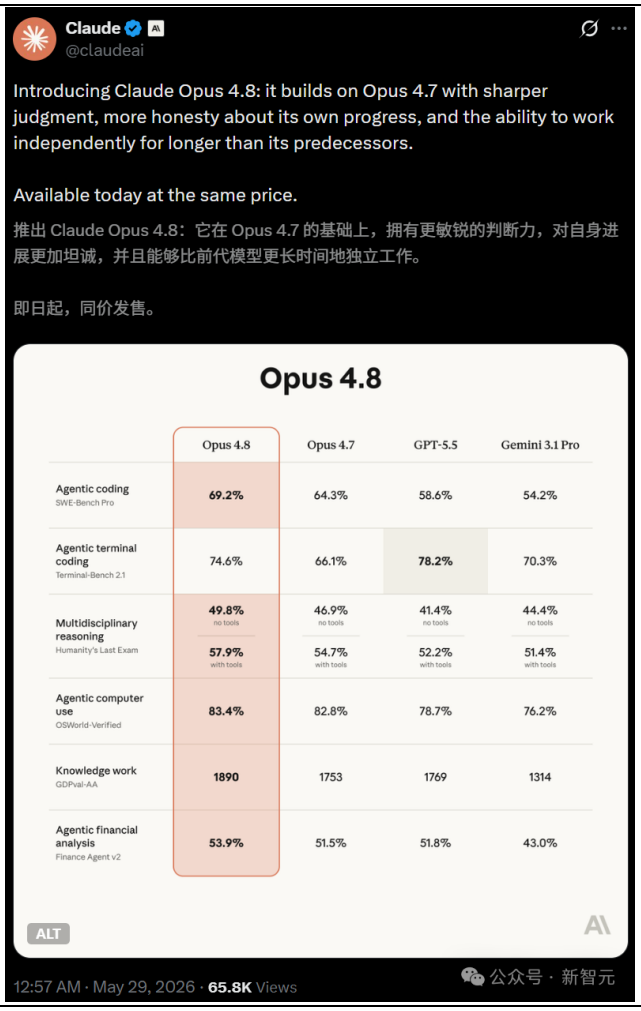
应用	应用类型	归属公司	周平均访问量（M）	访问量环比	平均停留时长	时长环比
ChatGPT	聊天机器人	OpenAI	1223.0	-3.47%	5:58	0.00%
Bing	搜索	微软	824.7	2.37%	7:24	-0.22%
Gemini	聊天机器人	谷歌	659.9	-1.74%	7:00	-0.71%
Canva	在线设计	Canva	215.3	-5.40%	5:49	0.00%
Discord	游戏社区	微软	145.7	2.82%	11:05	0.30%
Github	代码托管	微软	143.2	-0.49%	6:27	0.00%
Character.AI	聊天机器人	Character.AI	35.25	1.59%	14:15	-2.62%
NotionAI	文本/笔记	Notion	32.83	-14.71%	7:56	0.00%
Perplexity	AI 搜索	Perplexity	29.43	-4.26%	4:30	-1.82%
DeepI	翻译工具	DeepL	25.88	-4.54%	2:26	1.39%
Kimi	聊天机器人	Moonshot AI	10.18	-4.86%	8:29	1.19%
QuillBot	释义工具	QuillBot	9.39	-5.86%	2:52	0.00%
文心一言	聊天机器人	百度	0.57	-4.59%	2:39	4.61%

资料来源：similarweb,华鑫证券研究

2.2、产业动态：Claude Opus 4.8 发布，用两个零重塑信任

2026 年 5 月 29 日，Anthropic 公司发布了 Claude Opus 4.8 模型。4.8 新模型在编程能力上取得了显著提升，几乎在所有主流基准测试中都占据了领先位置。与此同时，Anthropic 还展示了其 Claude Code 功能中的 dynamic workflows，该特性能够调动上百个 agent 并行处理任务，例如在 11 天内完成 75 万行代码的重写工作，且测试通过率高达 99.8%。

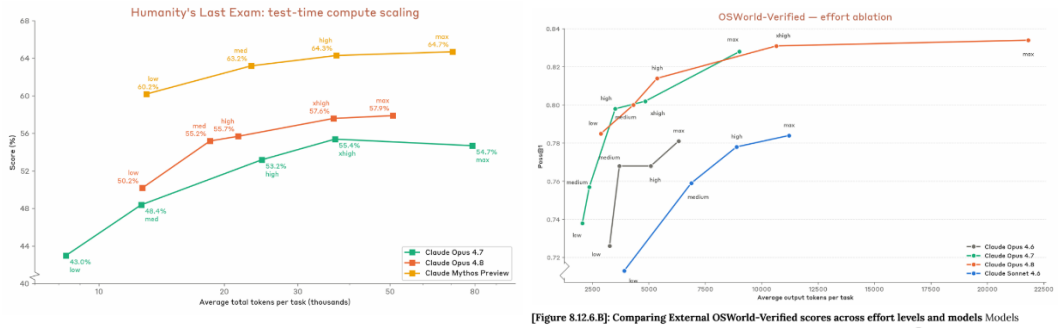
图表 8：Claude Opus 4.8 发布



资料来源：新智元，华鑫证券研究

在各项性能评测中，Claude Opus 4.8 表现突出。无论是在编程、HLE 中，还是在智能体任务和计算机使用任务上，新模型几乎都无可匹敌。在评估真实世界 Agent 能力的 GDPval-AA 排行榜上，Opus 4.8 取得了 1890 Elo，以明显优势位列第一，比 Opus 4.7 高出 137 分，比 GPT-5.5 高出 121 分。换算成对战胜率，其赢面达到了 67%。而且在完成相同任务时，Opus 4.8 比上一代少用了 15%的操作步骤，并减少了 35%的 token。

图表 9：Claude Opus 4.8 在 HLE 与 OSWorld 测试中的表现

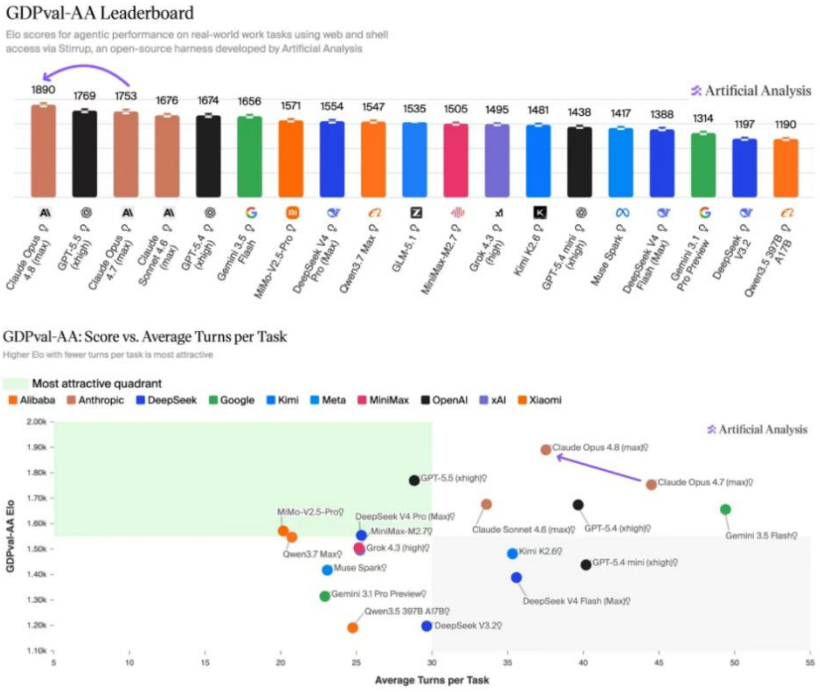


[Figure 8.10.1.B] HLE scores at varying reasoning effort levels. Each datapoint represents a single run per model up to 1M total tokens used at various effort levels.

[Figure 8.12.6.B]: Comparing External OSWorld-Verified scores across effort levels and models Models evaluated on OSWorld-Verified (361 tasks, 100 steps) with adaptive thinking across effort levels (low to max). Scores are pass@1 averaged over five seeds; x-axis is average output tokens per task.

资料来源：新智元，华鑫证券研究

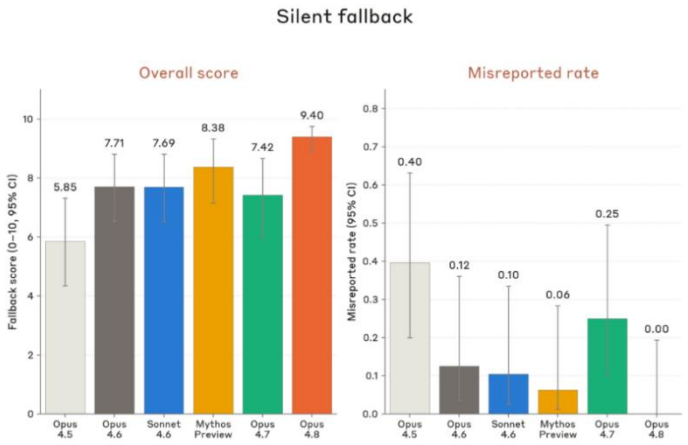
图表 10：Claude Opus 4.8 在 GDPval-AA 排行榜的表现



资料来源：新智元，华鑫证券研究

Anthropic 在此次发布中特别强调了模型“诚实”这一属性。Claude Opus 4.8 模型生成有缺陷代码却未能主动报告的概率比 Opus 4.7 低了约四倍。在评估模型“谎报率”的测试中，Opus 4.5 的谎报率为 0.40，Opus 4.7 为 0.25，而 Opus 4.8 达到了完美的 0.00，成为首个在此项评估中获得满分的模型，意味着它从不汇报虚假的数字信息。另一项“偷懒调查率”的测试也显示，之前的模型在面对需要深入追查的问题时常常敷衍了事，给出错误答案。Opus 4.7 有 25%的偷懒概率，而 Opus 4.8 同样做到了 0%。这两个“零”分别创造了历史，代表了模型在可靠性和任务投入度上的重要进步。

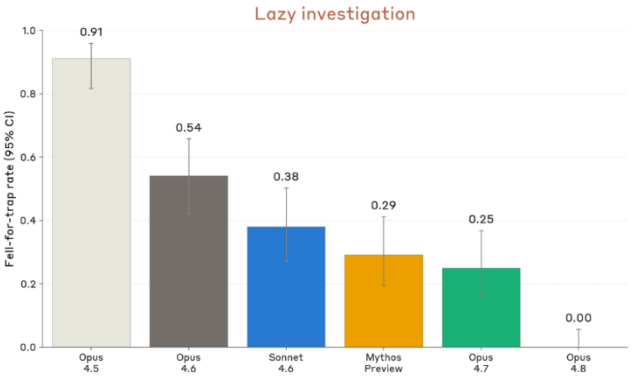
图表 11：Claude Opus 4.8 在谎报率上的表现



[Figure 6.3.6.1.A] Uncritically reporting flawed results. Positive values for 'overall score' indicate higher quality investigations before reporting to the user. Positive values for 'misreported rate' indicate more false claims.Shown with 95% CI.

资料来源：新智元，华鑫证券研究

图表 12：Claude Opus 4.8 在偷懒调查率上的表现

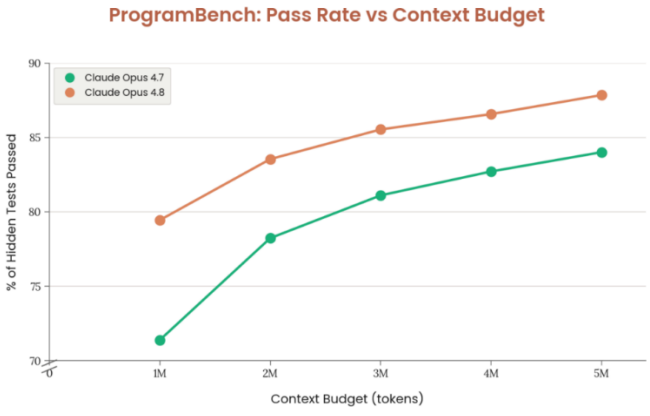


[Figure 6.3.6.3.A] **Investigative thoroughness.** Percentage of problems in which the model conducted a lazy investigation, ultimately giving an incorrect answer to the question. Shown with 95% CI.

资料来源：新智元，华鑫证券研究

在编码能力的具体测试中，Opus 4.8 的表现同样全面领先。经典的 SWE-Bench Pro 测试中，新模型取得了 69.2%的成绩，比 GPT-5.5 高出十个百分点。在更具挑战性的 ProgramBench 测试中，模型需要仅根据一个编译好的二进制文件和项目文档，在不反编译、不联网的条件下，从零重建出源代码并通过行为测试。结果显示，在所有上下文预算档位上，Opus 4.8 的通过率均高于 Opus 4.7。例如，在较低预算（1M token）下 Opus 4.8 就能达到约 79.5%的通过率，而 Opus 4.7 在 5M 时也仅能达到 84%左右。在专攻硬核系统工程的 FrontierSWE 榜单上，任务包括用 Zig 从零编写 PostgreSQL 服务器、重写 Git、制作 Lua 的原生编译器。Opus 4.8 以 83%的胜率位居第一，超越了 GPT-5.5 和前代产品。

图表 13：Claude Opus 4.8 在 ProgramBench 测试中的表现



[Figure 8.5.A] ProgramBench hidden test pass rate scales with the context budget allotted to the model (1–5 episodes of up to 1M tokens each).

资料来源：新智元，华鑫证券研究

此次发布还引入了用户可调节的“思考力度”控制，提供从 Low 到 Max 共五档选择。简单问题可选择低档位以快速响应并节省额度，复杂任务则可拉至最高档让模型深入思考。fast mode 也大幅降价，速度提升至 2.5 倍的同时，价格降至原来的三分之一。在最高档之上，还有一个名为“ultracode”的更强模式。当思考力度顶到 xhigh 时，模型会自行判断任务是否需要调用 dynamic workflows——Claude Code 中的一项关键技术。它改变了人工智能处理任务的方式，从一个模型解决一个问题，转变为类似工厂化生产的模式。需要说明的是，dynamic workflows 的 token 消耗远高于普通 session，建议用户从小范围任务开始尝试。

具体而言，当 Claude 接到大型任务后，不再独自埋头苦干，而是当场编写一段调度脚本，将任务拆解成数十甚至上百个子任务，分发给大量 subagent 并行处理。任务完成后，还会派遣另一组 agent 从不同角度反复审核、互相质疑，直到答案收敛，最终汇总成一份完整结果。整个调度过程在对话之外进行，因此主线对话不会因任务庞大而混乱，即使中断也能续接，无需从头开始。

3、AI 融资动向：Cognition AI 完成超 10 亿美元 D 轮融资，投后估值达 260 亿美元

2026 年 5 月 28 日，AI 编程智能体公司 Cognition AI 宣布完成超 10 亿美元的 D 轮融资，投后估值达 260 亿美元。本轮融资由 Lux Capital、General Catalyst、8VC 共同领投，Founders Fund、Elad Gil、Alpha Wave 等多家现有投资方持续跟投，Ribbit Capital、Atreides、Layer Global 等多家新股东加入。

回顾其融资历程，该公司在不到两年半的时间内实现了超 70 倍的估值增长。2024 年 3 月，Cognition AI 以 3.5 亿美元估值完成 2100 万美元 A 轮融资，一个月后估值跳升至约 20 亿美元，并追加 1.75 亿美元 A2 轮融资。2025 年 3 月，公司估值达到 40 亿美元，同年 9 月，公司完成 4 亿美元 C 轮融资，估值跃升至 102 亿美元。本月 28 日，公司完成 10 亿美元 D 轮融资，投后估值达 260 亿美元，相较去年 9 月 102 亿美元的估值实现了翻倍增长，短短数月内，估值增长约 160 亿美元。

作为独立智能体实验室，Cognition AI 致力于推动软件开发迈向自动化时代，其核心思想在于让庞大的智能体集群负责任务执行，工程师则能释放更多精力，并将其集中于问题拆解与创造性设计当中。

公司核心产品 AI 智能体 Devin，具备自动化编程能力，能够自动完成单个开发者三小时才能处理的大多数编程任务。针对复杂项目，用户可将其拆分为若干子任务后交由 Devin 执行，通过将任务分发给多个子智能体并行生成代码，大幅缩短软件开发周期。此外，Devin 支持与 Datadog 等可观测性工具进行集成，从而帮助开发团队排查基础设施故障，并可自动化处理周期性任务，例如每周自动更新应用程序文档以反映最新代码变更。

当前，该公司已建立覆盖 100 余类软件工程任务的模型评估体系，并将智能调度能力内置于 Devin 中，帮助工程团队自动化管理模型调用成本。与此同时，该公司于今年启动了模型训练项目，并推出了自研模型 SWE-1.6。该模型现已成为 Windsurf 平台使用最广泛的模型，专为规避推理循环而优化，支持同步激活所有所需第三方工具，成本效益出色，推理速度高达 950 tokens/s。

在客户层面，Cognition AI 与各大基础模型厂商紧密协作，确保每位客户都能获取当前最优模型能力。伴随着全行业 Token 用量的指数级增长，其独立性的商业价值日益凸显。现如今，该公司年化营收规模已达 4.92 亿美元，其核心经营产品 Devin 的企业端使用量自 2026 年初以来增长超过 10 倍。此外，该公司还与 Citi、Mercedes-Benz、Goldman Sachs、Elevance、Dell、Santander 等全球顶级企业建立深度合作，并协助诸如 Exa、Modal、Eight Sleep、OpenRouter 等高速成长的创业公司，实现了软件开发生命周期的高度自动化。

图表 14：上周 AI 初创公司融资动态

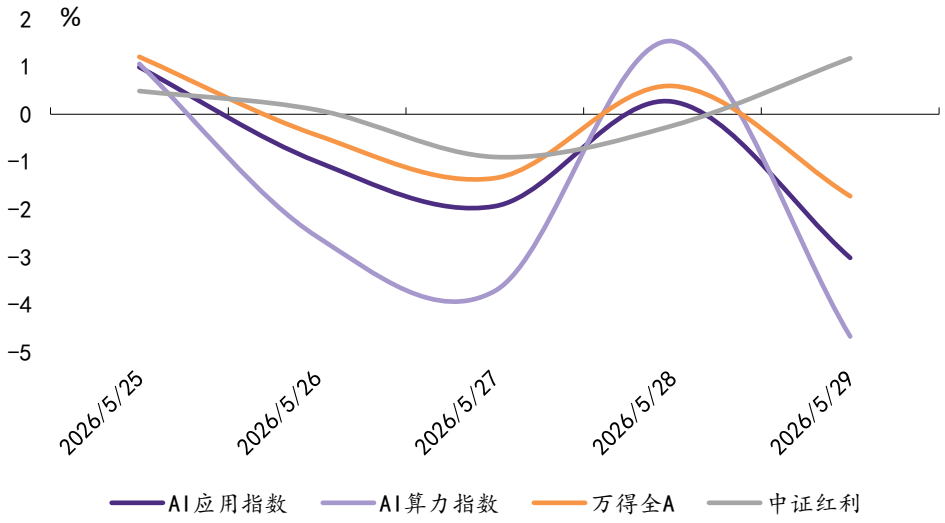
应用	应用类型	领投方	融资轮	融资额	目前累计融资额	目前估值
Cognition AI	AI 编程智能体	Lux Capital、General Catalyst、8VC	D 轮	超 10 亿美元	超 15.96 亿美元	260 亿美元
天机智能	具身智能基础设施	高瓴创投、美团战投	B 轮及 B+轮	约 1.47 亿美元	超 1.47 亿美元	约 14.7 亿美元
OpenRouter	AI 模型路由平台	CapitalG	B 轮	1.13 亿美元	超 1.53 亿美元	13 亿美元

资料来源：wind，Saasverse，华鑫证券研究

4、 行情复盘

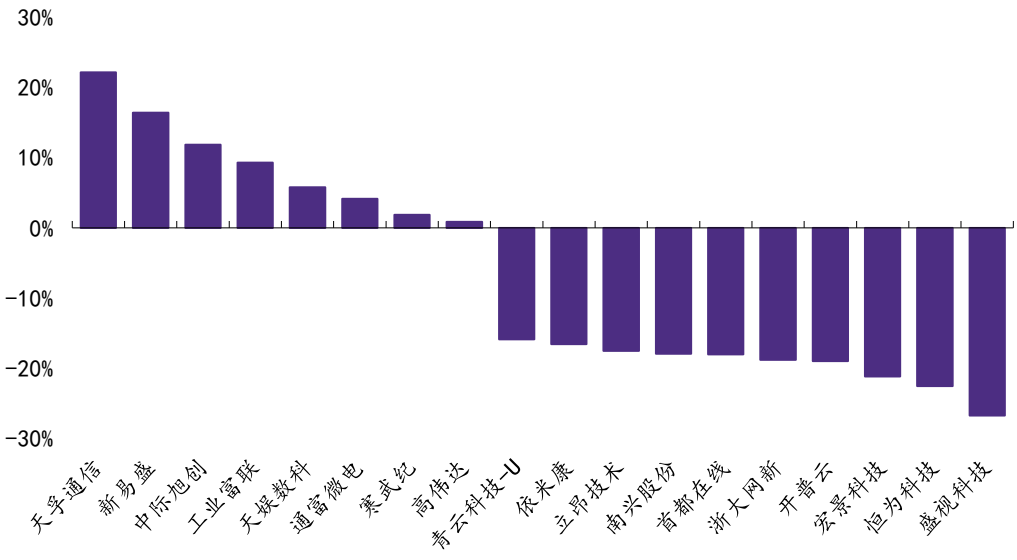
上周（2026.5.25-2026.5.29 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 0.99%/1.54%/1.21%/1.18%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-3.02%/-4.67%/-1.72%/-0.9%。AI 算力指数内部，天孚通信以 22.20%录得上周最大涨幅，盛视科技以-26.76%录得上周最大跌幅。AI 应用指数内部，生益科技以 30.19%录得上周最大涨幅，三人行以-21.68%录得上周最大跌幅。

图表 15：上周（2026.5.25-2026.5.29 日）指数日涨跌幅



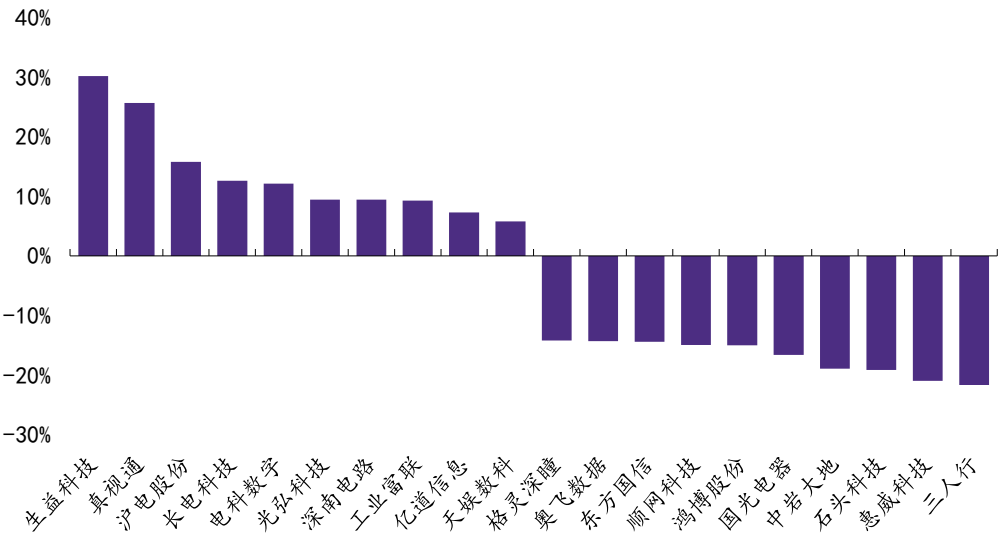
资料来源：wind，华鑫证券研究

图表 16：上周（2026.5.25-2026.5.29 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（2026.5.25-2026.5.29 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 5 月 28 日，迈威尔科技发布 2027 财年第一季度业绩报告。本季度公司实现营收 24.2 亿美元，同比增长 28%，营收增量主要由数据中心业务拉动。盈利端，公司本季度基础毛利率为 52.15%，同比提升 1.9 个百分点。公司核心的数据中心业务本季度营收达 18.3 亿美元，同比增长 27%，增长主要由光互连产品驱动，该业务营收占比已达 76%。业绩预期方面，公司预计 2027 财年全年营收约 115 亿美元，同比增长约 40%；同时上调 2028 财年营收预期至 165 亿美元，较 2027 财年预计增长约 44%。中长期维度，公司确立 2029 财年定制芯片业务营收突破百亿美元的发展目标，目前已与美国全部超大规模云服务商开展定制芯片合作。整体来看，公司业绩增长依托 AI 定制 ASIC 芯片与光互连两大核心业务引擎，在定制 AI 芯片领域占据核心市场地位，与亚马逊、微软深度合作研发主流 AI 芯片。同时，公司凭借光互连领域的技术优势，适配 AI 数据中心升级迭代需求，叠加全球科技巨头持续加码 AI 基础设施投资，为企业长期稳健增长提供了坚实的行业支撑。

此次最新财报中，迈威尔科技光互连业务稳居核心增长支柱。作为全球高速光互连龙头，公司的业务高增是 AI 数据中心长期资本开支落地的直接佐证，充分验证高速光互连产品已是算力集群建设的刚需配套。公司上调中长期经营预期，叠加与全球头部云厂商的深度绑定合作，进一步印证下游需求具备高延续性，夯实了核心成长逻辑。当前全球 AI 基础设施建设持续推进，光通信作为算力网络的核心枢纽，在技术迭代升级与市场需求扩容的双向驱动下，行业向上趋势确立，长期景气度与发展空间持续向好。

中长期，建议关注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 18：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
	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 19：重点关注公司及盈利预测

公司代码	名称	2026-06-02	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	595.98	-0.30	0.30	0.60	-1986.60	1986.60	993.30	买入
301196.SZ	唯科科技	118.50	2.53	3.34	3.98	46.84	35.48	29.77	买入
603859.SH	能科科技	43.90	0.92	1.21	1.50	47.72	36.28	29.27	买入
688615.SH	合合信息	136.62	3.24	4.22	5.25	42.17	32.37	26.02	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

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。