下载日志已记录，仅供内部参考

2026 年 09 月 01 日

华鑫证券 CHINA FORTUNE SECURITIES

智谱发布 GLM-5.3-Flash，Qwen3.8-Flash-Next 开源

—计算机行业周报

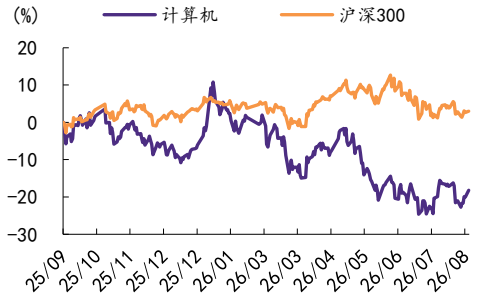
推荐(维持)

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)	2.6	-6.6	-21.5
沪深 300	0.8	-4.5	2.2

市场表现



资料来源：Wind，华鑫证券研究

相关研究

1、《计算机行业点评报告：Roblox：Q2 营收与自由现金流强劲增长，AI 赋能 UGC 创作生态再扩容》2026-08-30

2、《计算机行业点评报告：DoorDash（DASH）：会员飞轮驱动稳健增长，AI 与自动驾驶拓展长期空间》2026-08-30

3、《计算机行业周报：DeepSeek 上线全新多模态模型，CodexHarness 开源》2026-08-24

投资要点

算力：百度发布 Agent 沙箱夯实智能体底座，智谱开源 GLM-5.3-Flash 实现性能与成本双优化

2026 年 8 月 26 日智谱 AI 开源 GLM-5.3-Flash 原生多模态大模型，依托稀疏注意力等架构优化，在代码、Agent 任务性能达到国际前沿水平，推理成本大幅下降，同时完成国产算力适配，有望推动大模型在高频场景规模化落地

AI 应用：NotionAI 周访问量环比+7.76%，Qwen3.8-Flash-Next 开源

2026 年 8 月 26 日，阿里千问团队正式开源了基于下一代 Qwen4 架构构建的多模态 MoE 模型——Qwen3.8-Flash-Next。该模型主体参数量达到 125B，外配备了 51B 的 N-gram Embedding，每 token 激活 6B，在上下文处理能力方面，该模型原生支持 262144 token 的窗口长度，并可借助 YaRN 扩展技术将有效上下文进一步提升至 100 万 Token 的量级。

AI 融资动向：鹏行智能完成超 9 亿美元首轮融资，投后估值超 63 亿美元

鹏行智能完成超 9 亿美元首轮股权融资，投后估值 63 亿美元，由 IDG 资本领投。资金将投入人形机器人本体、AI 模型训练与产线建设，本轮融资体现资本对国内具身智能赛道的高度看好。

投资建议

近日，英伟达发布 2027 财年第二季度财报。当季收入 962.21 亿美元，同比增长 106%、环比增长 18%，增速连续第四个季度环比抬升；其中数据中心收入 890 亿美元，同比增长 117%、环比增长 18%，超大云厂商（hyperscale）收入 490 亿美元、环比增长 13%，边缘计算收入 72 亿美元，同比增长 27%；GAAP 与 Non-GAAP 毛利率均为 75.0%（上年同期 72.4%/72.5%）；GAAP 净利润 596.88 亿美元，同比增长 126%，摊薄 EPS2.46 美元，同比增长 128%；当季自由现金流 213.41 亿美元（上季 485.54 亿美元），环比回落主因应收账款增加及为 VeraRubin 量产备货；当季回购与股息合计回馈股东约 260 亿美元；当季新发债约 249 亿美元。指引方面，3QFY27 收入预期 1,080 亿美元±2%（不含中国数据中心计算收入），毛利率 74.0%±50bp；公司预计 FY2028 收入将增长

下载日志已记录，仅供内部参考

约 70%，并强调该预期为供给受限口径。公司 CFO 在财报电话会上表示，当前云行业积压订单已超 2 万亿美元，并预计前五大超大云厂商 2026/2027 年资本开支分别接近 8,000 亿/1.3 万亿美元。财报同步披露，公司与 Apollo、BlackRock、Blackstone、Brookfield、高盛、KKR 达成战略合作，拟设立独立算力融资平台、撬动超 5,000 亿美元第三方资本投向 AI 基础设施；VeraRubin 平台已进入全面量产，机架在 CoreWeave、GoogleCloud、MicrosoftAzure、OCI、Nebius 等伙伴处上线运行。

本次财报的核心信号在于需求端与投入端同时进一步加速：需求端，去年同期全球 AI 算力建设基本由单一大客户拉动，本季已变为多家前沿实验室集体大规模扩建，叠加 AI 创业公司、各国主权 AI 项目及机器人等新需求放量；1,080 亿美元的下季指引对应环比约 12%的增长，表明算力需求仍在加速兑现。投入端，公司首次通过"自己发债+联合六大资管机构设立算力融资平台"的方式，为下游客户筹措超 5,000 亿美元建设资金，商业模式从单纯出售芯片进一步延伸至为客户提供融资支持；库存与应收账款同步大增，也指向 VeraRubin 量产爬坡的确定性较高。我们认为，英伟达财报所呈现的需求信号与算力融资平台的设立相互印证，全球 AI 基础设施投入仍处加速期，公司管理层对前五大云厂商 2027 年资本开支接近 1.3 万亿美元的预期将进一步打开算力产业链空间。

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）、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	2026-09-01	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	569.75	-0.30	0.30	0.60	-1899.17	1899.17	949.58	买入
301196.SZ	唯科科技	95.40	2.05	1.92	2.39	39.28	49.65	39.86	买入
301369.SZ	联动科技	138.30	0.48	1.08	2.62	288.13	128.06	52.79	买入
603859.SH	能科科技	38.88	0.92	1.05	1.32	42.26	37.03	29.45	买入

下载日志已记录，仅供内部参考

688615.SH	合合信息	99.95	3.24	4.22	5.25	30.85	23.68	19.04	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：百度发布 AGENT 沙箱夯实智能体底座，智谱开源 GLM-5.3-FLASH 实现性能与成本双优化 .. 5
1.1、 Tokens 跟踪 5
1.2、 云厂动态：百度智能云发布智能体沙箱 AX，持续完善 Agent 云基础设施 7
1.3、 产业动态：智谱发布 GLM-5.3-Flash，性能持续提升，推理成本大幅下降 8
2、 AI 应用动态：NOTIONAI 周访问量环比+7.76%，QWEN3.8-FLASH-NEXT 开源 10
2.1、 周流量跟踪：NotionAI 周访问量环比+7.76% 10
2.2、 产业动态：Qwen3.8-Flash-Next 开源，GDN 与 QSA 双轨协同破解长文本记忆与检索难题 10
3、 AI 融资动向：鹏行智能完成超 9 亿美元首轮融资，投后估值超 63 亿美元 13
4、 行情复盘 15
5、 投资建议 17
6、 风险提示 18

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：GLM-5.3-FLASH 基准测试结果 8
图表 4：主流前沿模型性能与价格对比 9
图表 5：2026.8.21-2026.8.27 AI 相关网站流量 10
图表 6：QWEN3.8-FLASH-NEXT 与 DEEPSEEK-V4 系列模型价格对比(单位:元/百万 TOKENS) 11
图表 7：90% PREFIX CACHE 命中率下的相对 PREFILL 吞吐量 11
图表 8：QWEN3.8-FLASH-NEXT 模型架构 12
图表 9：上周 AI 初创公司融资动态 13
图表 10：上周（2026.8.24-2026.8.28 日）指数日涨跌幅 15
图表 11：上周（2026.8.24-2026.8.28 日）AI 算力指数内部涨跌幅度排名 15
图表 12：上周（2026.8.24-2026.8.28 日）AI 应用指数内部涨跌幅度排名 16
图表 13：FICONTEC2025 年年中至今公告订单 17
图表 14：重点关注公司及盈利预测 18

下载日志已记录，仅供内部参考

1、 算力动态：百度发布 Agent 沙箱夯实智能体底座，智谱开源 GLM-5.3-Flash 实现性能与成本双优化

1.1、 Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 8 月 24 日至 8 月 30 日，周度 Token 消耗量显著上涨，消耗量为 113T，环比上周增加 24.04%。在 Tokens 规模 Leaderboard 前五名中，stealth 的 Ox Alpha 以 15.7T tokens 位列榜首；Deepseak 的 Deepseek V4 Flash0731 以 12.3Ttokens 位居第二；Xiaomi 的 MiMo-V2.5 以 9.14T tokens 位居第三；Openai 的 GPT-5.6Luna 以 7.79T tokens 位居第四;Tencent 的 Hy3 以 6.66Ttokens 位居第五。

文本类模型的市场份额，Deepseek 以 1.11B requests 占据 23.5%的份额，位列第一。google,openai,z-ai 则分别以 1.01B、927M、295M requests，对应占据 21.5%、19.7%、6.3%的市场份额。

图像生成类模型调用量略上升，调用量为 5.76M requests，环比上周增加约 0.52%。在 Image Leaderboard 中，Google 的 Nano Banana 系列包揽榜单前四名席位，其中，Nano Banana 2 (Gemini 3.1 Flash Image)以 1.32M requests 位居第一；Banana 2 Lite (Gemini 3.1 Flash Lite Image)、Nano Banana 2 (Gemini 2.5 Flash Image)、Nano Banana 2 (Gemini 3.1 Flash Image Preview)分别以 1.13M、607K、479K requests 位居第二、第三和第四；OpenAI 的 GPT Image 2 以 456K requests 位列第五。从市场份额来看，Google 以 4.12M requests 占据 71.5%的份额，位居第一；OpenAI 以 137K requests 占据 12.5%，位列第二；xAI、Black Forest Labs、ByteDance Seed 则分别以 226K、219K、194K requests，对应占据 3.9%、3.8%、3.4%的市场份额。

嵌入模型调用量有所下降，调用量为 286M requests，环比上周下降 4.98%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 113M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 66.3M requests 位列第二；Google 的 Gemini Embedding 2 以 22.4M requests 位列第三；OpenAI 的 Text Embedding 3 Large 以 19.5M requests 位列第四；BAAI 的 bge-m3 以 18.6M requests 位列第五。从市场份额来看，OpenAI 以 137M requests 占据 47.8%的份额，位居第一；Qwen 以 70.3M requests 占据 24，6%，位列第二；Google、BAAI、Perplexity 则分别以 36.3M、19.4M、8.99M requests，对应占据 12.7%、6.8%、3.1%的市场份额。

重排序模型调用量小幅上升，调用量为 6.37M requests，环比上周增加 3.24%。在 Rerank Leaderboard 中，VoyageAI 的 rerank-2.5 以 2.72M requests 位列榜首；Cohere 的 Rerank 4 Fast 以 1.42M requests 位居第二；Cohere 的 Rerank v3.5、NVIDIA 的 Liama Nemotron Rerank V1LBV2、Rerank 4 Pro 分别以 963K、380K、328K requests 位列第三、第四和第五。从市场份额来看，V 以 2.99M requests 占据 47.0%的份额，位居第一；Cohere 以 2.71M requests 占据 42.5%的份额，位居第二；NVIDIA，Qwen 分别以 380M、291K requests 占据 6.0%、4.6%的份额。

下载日志已记录，仅供内部参考

视频生成类模型调用量有所上升，调用量为 171K requests，环比上周增加 9.62%。在 Video Leaderboard 前五名中，Google 的 Veo3.1 Lite 以 30K requests 位列第一；ByteDance 的 Seedance 2.0 Mini 以 27K requests 位列第二；ByteDance 的 Seedance 2.0 Fast、Seedance 2.0、Veo3.1Fast 分别以 13K、13K、13K requests 并列第三。从市场份额来看，ByteDance 以 72K requests 占据 42.1%的份额，位列第一；Google 以 48K requests 占据 27.8%的份额，位居第二；xAI、Alibaba、KwaiVGI 则分别以 15K、13K、10K requests，对应占据 8.7%、7.8%、5.8%的市场份额。

语音生成类模型调用量大幅下降，调用量为 2.13M requests，环比上周下降 54.39%。在 Speech Leaderboard 前五名中，Google 的 Gemini3.1 Flash TTSP review 以 775K requests 位居第一；Hexgrad 的 Kokoro 82M 以 481K requests 位列第二;Fish-audio 的 S2.1Pro Free requests 以 325K 位列第三；xAI 的 Grok Voice TTS 1.0 以 215K requests 位列第四；Deepgram 的 FluxTTS 以 60K requests 位列第五。从市场份额来看，Google 以 775K requests 占据 36.4%的份额，位居第一；Hexgrad 以 481K requests 占据 22.6%，位列第二； fish-audio, x-ai, deepgram 则分别以 386K、215K、116K requests，对应占据 18.1%、10.1%、5.5%的市场份额。

语音转写类模型调用量有所上涨，调用量为 13.7M requests，环比上周增加 12.30%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3、Whisper Large V3 Turbo、GPT-4o Mini Transcribe 分别以 7.25M、3.57M、584K requests 位列第一、第二和第三；Qwen 的 Qwen3 ASR Flash 以 335K requests 位列第四；Mistralai 的 Voxtral Mini Transcribe 的 Whisper 1 以 333K requests 位列第五。从市场份额来看，OpenAI 以 12.1M requests 占据 87.9%的份额，位居第一；Qwen 以 605K requests 占据 4.4%，位列第二；Mistralai, nvidia, x-ai 则分别以 376K、313K、155K requests 占据 2.7%、2.3%、1.1%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。

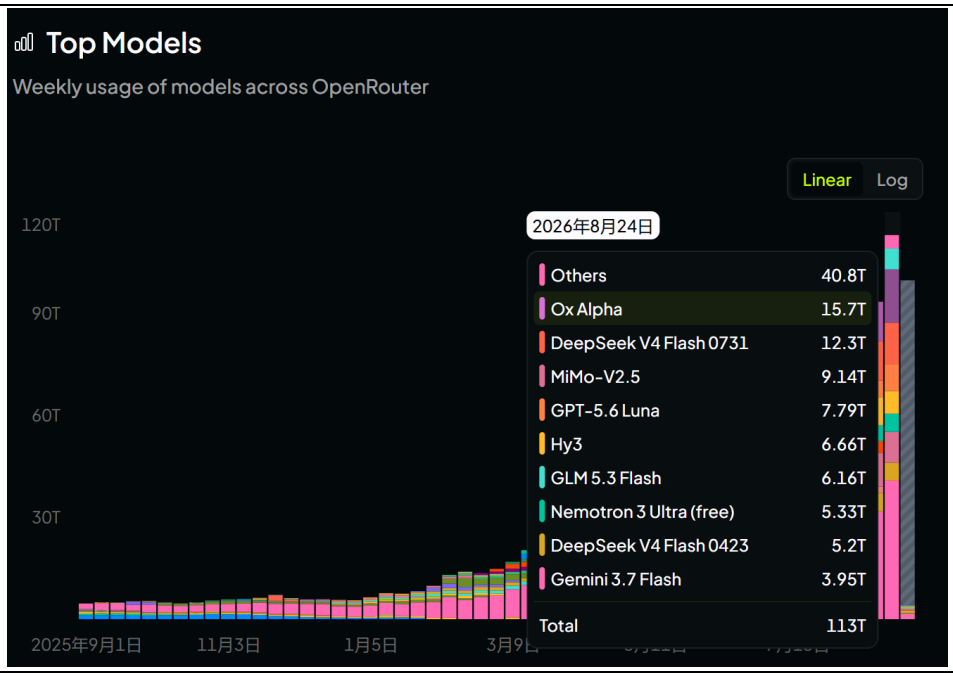
8 月 24 日浙江嘉兴市城投集团主导运营的长三角（嘉兴）Token 运营中心正式启动运行，该中心不直接生产算力，而是整合分散算力与各类 AI 大模型，以标准化服务形式向企业提供区域性 AI 算力和模型服务。

8 月 26 日 Token 经济发展大会在深圳国际会展中心召开，大会以“Token 经济重塑 AI 商业模式”为主线，汇聚腾讯云、阿里云等行业代表，集中探讨 Token 经济从概念走向规模化落地的核心路径。

8 月 28 日 2026 中国国际大数据产业博览会在贵阳开幕，大会将“词元——数据要素价值释放新路径”设为核心主题，行业代表现场提出 Token 价值释放的核心判断，推动数据基础设施向 AI-Ready 体系升级。

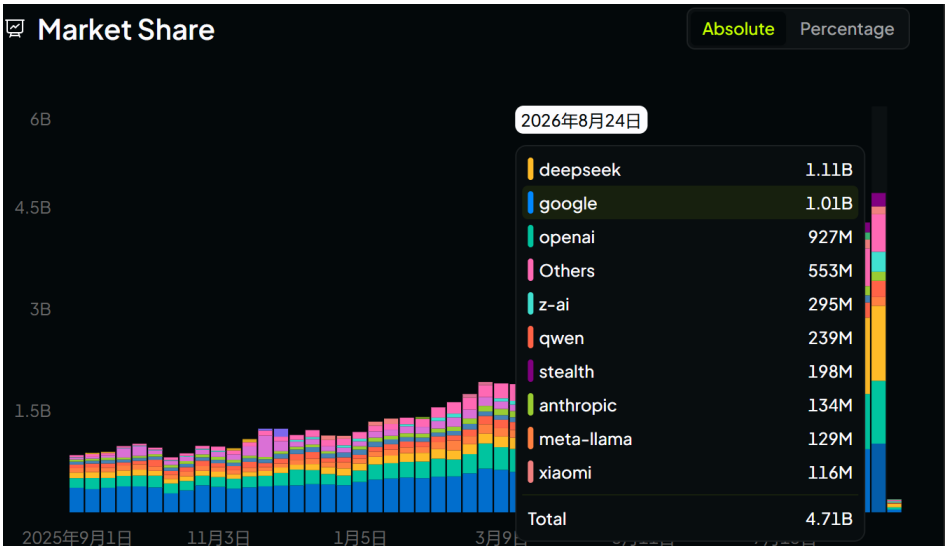
下载日志已记录，仅供内部参考

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、云厂动态：百度智能云发布智能体沙箱 AX，持续完善 Agent 云基础设施

2026 年 8 月 26 日，百度智能云正式发布智能体沙箱 AX（Agent Sandbox），围绕 AI 应用托管与 Sandbox 安全隔离执行两类核心能力，为 AI 应用提供从部署、运行到观测的一体化云服务。AX 支持代码执行、浏览器自动化、桌面控制及强化学习训练等多类场景，并支

下载日志已记录，仅供内部参考

持通过 SDK、REST API、MCP 及 Function Call 等方式调用。

当前，随着 AI Agent 由简单问答向复杂任务执行演进，Agent 在运行过程中对代码执行、浏览器操作、外部工具调用以及独立运行环境的需求持续提升。百度智能云 AX 通过虚拟机级安全隔离，为 AI 应用及外部 Agent 提供独立执行环境，并进一步统一模型、工具、网络、存储、权限及运行资源管理，降低企业部署 Agent 的基础设施门槛。

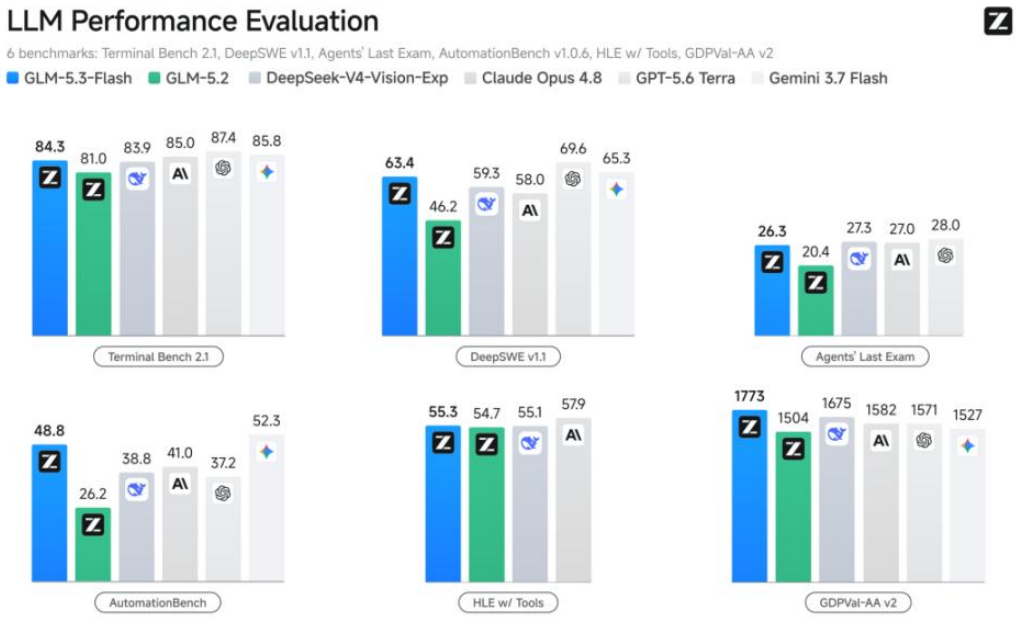
在应用场景方面，AX 不仅支持 AI 应用托管，还覆盖代码执行、Browser Use、Computer Use 及强化学习等场景，并支持万级并发沙箱及弹性扩展。百度智能云此次推出智能体沙箱，体现云厂商的竞争重点正由传统算力、存储等基础云资源，进一步向 Agent 运行环境及 AI 原生基础设施延伸，有望推动 AI 应用从模型调用向复杂任务执行加速落地。

1.3、 产业动态：智谱发布 GLM-5.3-Flash，性能持续提升，推理成本大幅下降

2026 年 8 月 26 日，智谱上线并开源 GLM-5.3-Flash，为 GLM-5 系列首个原生多模态模型。模型采用 320B 总参数、18B 激活参数架构，在代码、Agent、多模态理解等任务上进一步强化。根据智谱披露，GLM-5.3-Flash 在 Artificial Analysis Intelligence Index 中取得 57 分，进入全球前沿模型能力区间，编程表现与 Claude Opus 4.8 相当。

从基准测试结果来看，GLM-5.3-Flash 在多项文本 Agent 及复杂任务测试中表现较为突出。其中，Terminal Bench 2.1、NL2Repo 等测试得分分别达到 84.3 分和 57.7 分，体现出模型在代码生成、软件工程及复杂任务处理方面的能力。整体来看，模型性能已逐步向国际前沿模型水平靠拢。

图表 3：GLM-5.3-Flash 基准测试结果



资料来源：智谱，华鑫证券研究

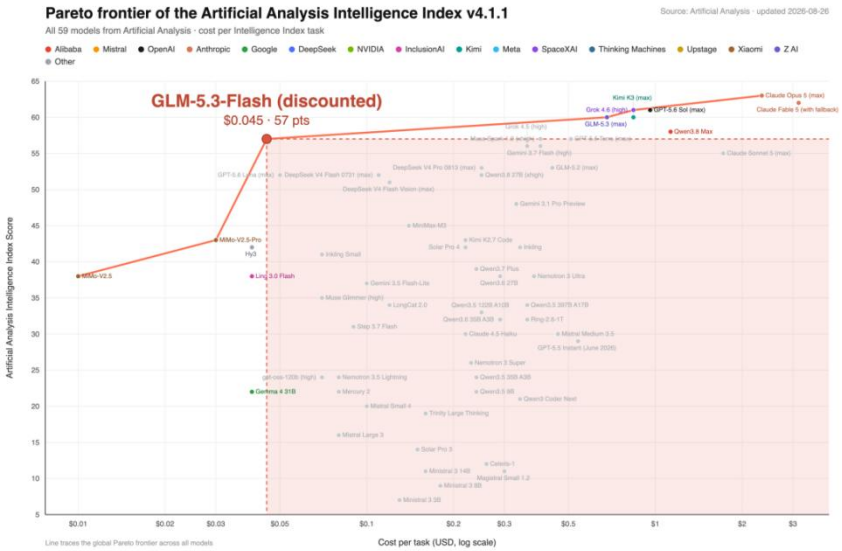
价格方面，模型推理成本显著下降与性能提升同步，GLM-5.3-Flash 进一步降低模型使

下载日志已记录，仅供内部参考

用成本。根据智谱披露，其定价约为 GLM-5.2 的 1/10，约为 Claude Opus 4.8 的 1/40。在模型能力持续提升的同时，推理价格大幅下降，有望降低开发者调用大模型及部署 AI Agent 的成本。

从模型性能与价格的综合比较来看，GLM-5.3-Flash 位于较高性价比区间，在保持前沿模型能力的同时，其 API 调用价格明显低于海外主流模型。模型性能提升与推理成本下降并行，有望进一步推动大模型在代码、Agent 等高频应用场景中的规模化使用。

图表 4：主流前沿模型性能与价格对比



资料来源：智谱，华鑫证券研究

模型架构持续优化，降低推理计算需求从技术架构来看，GLM-5.3-Flash 采用稀疏注意力与线性注意力混合架构，并引入 Manifold-Constrained Hyper-Connections（mHC）等技术，在保持长上下文能力的同时降低上下文服务成本。模型总参数量达到 320B，但激活参数仅 18B，相较 GLM-4.5 的 92B 激活参数明显下降，体现出模型厂商由单纯扩大参数规模向提升参数利用效率和推理效率转变。

此外，GLM-5.3-Flash 在发布前进行了国产算力环境下的大规模测试，相关请求流量实现国产芯片算力支持，反映国产大模型与国产 AI 算力的适配能力进一步加强。

下载日志已记录，仅供内部参考

2、AI 应用动态：NotionAI 周访问量环比+7.76%，Qwen3.8-Flash-Next 开源

2.1、周流量跟踪：NotionAI 周访问量环比+7.76%

本期（2026.8.21-2026.8.27）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1306.0M）、Bing（852.2M）和 Gemini（607.7M），访问量环比增速第一为 NotionAI（7.76%）；平均停留时长前三位分别为 Character.AI（00:15:58）、Discord（00:11:00）和 NotionAI（00:08:32）;平均停留时长环比增速第一为 ChatGPT（0.50%）。

图表 5：2026.8.21-2026.8.27 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	1306.0	1.24%	6:39	0.50%
Bing	搜索	微软	852.2	-0.94%	7:17	-1.13%
Gemini	聊天机器人	谷歌	607.7	1.67%	6:55	-0.48%
Canva	在线设计	Canva	203.7	3.66%	5:55	-0.28%
Github	代码托管	微软	148.2	0.34%	6:26	0.00%
Discord	游戏社区	微软	147.3	-1.41%	11:00	-0.60%
Character.AI	聊天机器人	Character.AI	34.91	-1.75%	15:58	-1.44%
DeepI	翻译工具	DeepL	22.57	0.18%	2:28	0.00%
Perplexity	AI 搜索	Perplexity	21.50	0.14%	4:25	0.00%
Kimi	聊天机器人	Moonshot AI	8.78	-19.15%	6:53	-2.59%
QuillBot	释义工具	QuillBot	7.60	4.50%	2:50	-0.58%
NotionAI	文本/笔记	Notion	2.79	7.76%	8:32	0.39%
文心一言	聊天机器人	百度	0.000672	-18.55%	1:12	-62.50%

资料来源：similarweb,华鑫证券研究

2.2、产业动态：Qwen3.8-Flash-Next 开源，GDN 与 QSA 双轨协同破解长文本记忆与检索难题

2026 年 8 月 26 日，阿里千问团队正式开源了基于下一代 Qwen4 架构构建的多模态 MoE 模型——Qwen3.8-Flash-Next。该模型主体参数量达到 125B，并额外配备了 51B 的 N-gram Embedding，每 token 激活 6B 参数。在上下文处理能力方面，该模型原生支持 262144 token 的窗口长度，并可借助 YaRN 扩展技术将有效上下文进一步提升至 100 万 Token 的量级。

从成本效益来看，Qwen3.8-Flash-Next 的训练开销仅为此前 Qwen3.7-Plus 模型的大约九分之一，但在编程开发与办公自动化等任务场景中展现出更为强劲的性能表现。在千问团队公布的基准测试结果中，该模型在绝大多数评测项目上均超越了 DeepSeek-V4-Flash 正

下载日志已记录，仅供内部参考

式版。定价策略上，该模型即将通过千问 AI 平台对外提供 API 服务，其输入 Token 每百万收费一元，输出 Token 每百万收费三元，相较于 DeepSeek-V4-Flash 在非高峰时段的定价，输入和输出价格均降低了约三分之一。

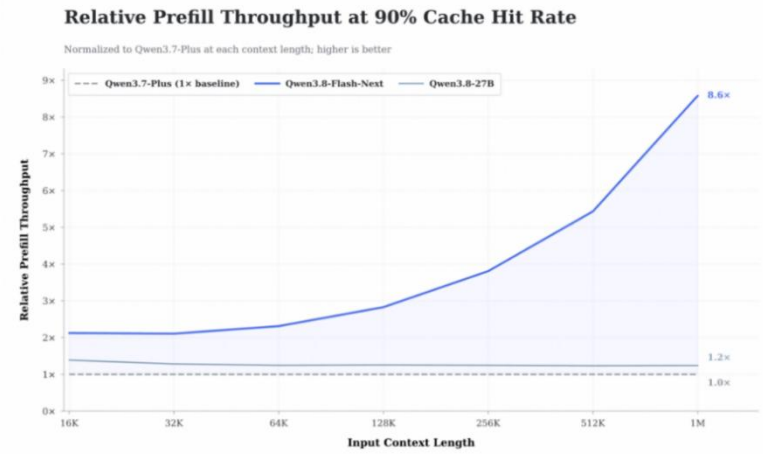
图表 6：Qwen3.8-Flash-Next 与 DeepSeek-V4 系列模型价格对比(单位:元/百万 Tokens)

模型	计费项	价格	
		空闲时段价格/半价	高峰时段价格/正常价格
DeepSeek-V4-Flash	输入(缓存命中)	0.05	0.1
	输入(缓存未命中)	1.5	3.0
	输出	4.5	9.0
DeepSeek-V4-Pro	输入(缓存命中)	0.15	0.3
	输入(缓存未命中)	4.5	9.0
	输出	13.5	27.0
Qwen3.8-Flash-Next	输入(缓存命中)	-	
	输入(缓存未命中)	1	
	输出	3	
GLM-5.3-Flash（上线前两周实行半价）	输入(缓存命中)	0.115 元	0.23 元
	输入(缓存未命中)	0.4 元	0.8 元
	输出	1.4 元	2.8 元

资料来源：智东西，华鑫证券研究

在模型架构设计上，Qwen3.8-Flash-Next 围绕 Attention、Residual、Embedding 和 Optimization 四个维度进行了系统性升级。具体而言，其创新性地将 GDN 与 QSA（通义稀疏注意力）相结合，延续 Qwen3.5 的架构设计采用 GDN+Attention 的 Hybrid 架构：每四层网络中有三层采用 Gated DeltaNet（GDN），将历史信息持续压缩至固定大小的状态中；剩余一层保留全局 Attention，专注于精确检索全局信息，实现高效记忆与精准查找的分工协作。同时对全局 Attention 进一步引入 Qwen Sparse Attention（QSA），通过只关注重要上下文来减少长序列下的计算。通过这种设计，在 1M token 的上下文长度下，QSA 的 Attention Kernel 在 Prefill 和 Decode 阶段分别实现了最高 7.6 倍和 4.9 倍的加速效果，而在高缓存复用场景中，Prefill 吞吐量可达到 Qwen3.7-Plus 的 8.6 倍。

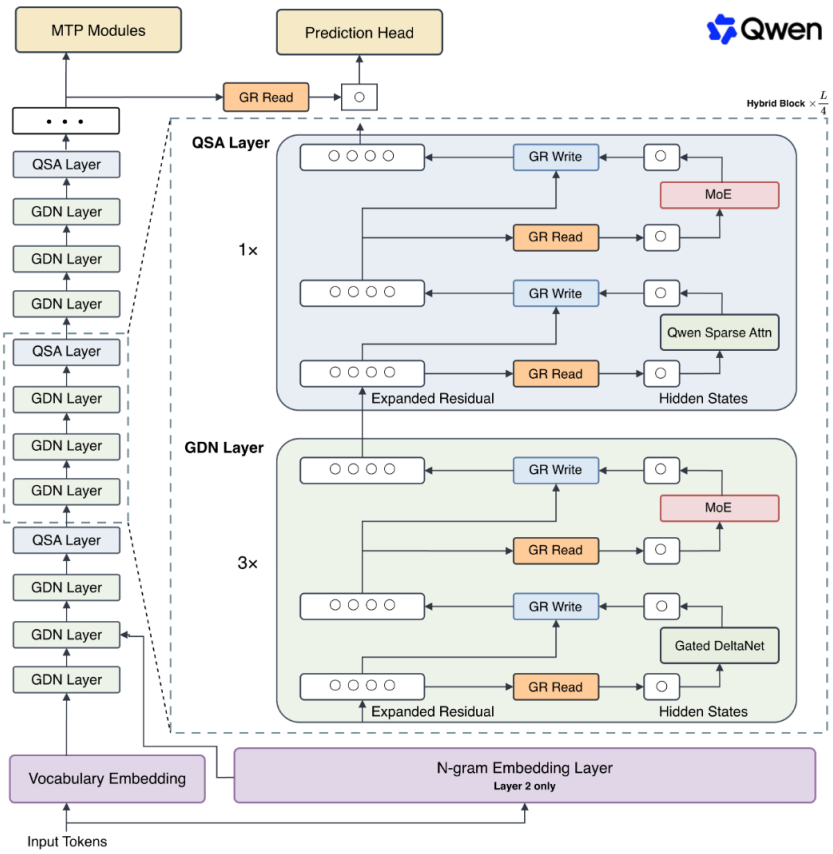
图表 7：90% Prefix Cache 命中率下的相对 Prefill 吞吐量



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

图表 8：Qwen3.8-Flash-Next 模型架构



资料来源：智东西，华鑫证券研究

此外，该模型引入了 Gated Residual（门控残差），将原本单一的 residual stream 扩展为四条并行分支，使得模型能够根据当前内容动态决定从各分支读取或写入信息的比例。GR 还进一步简化了 Hyper-Connection，归一化后的 Gate 可以有效抑制激活异常值，提升训练稳定性，并支持使用 FP8 格式存储残差状态以降低访存开销。在嵌入层面，该模型引入 N-gram Embedding，并额外引入了 51B N-gram Embedding 参数，通过结合当前 Token 与其前序若干 Token 的局部上下文进行查表，为常见短语和局部模式提供了补充表征，且这些参数可存放于 Host Memory 中，通过异步 Prefetch 与模型计算重叠，无需长期占用 GPU 显存，同时几乎不增加每个 Token 的计算量。在 Optimization 方面，团队针对 Muon 优化器在大规模训练中的使用进行了精细化调整，对于二维线性映射的参数，例如 Attention、GDN 和 MoE Expert 中的主要权重使用 Muon；对于 Embedding、MoE Router、GR 低秩参数等则继续使用 AdamW，并重新拟合了 Scaling Law，使得模型能够稳定采用更大的学习率和批次大小，从而提升收敛效率与并行训练吞吐。

在其余架构组件上，该模型沿用了 Qwen3-Next 系列经过反复验证的设计，包括极致稀疏 MoE，即在维持全局负载均衡的前提下固定激活专家数量而持续扩充专家总池，并辅以共享专家机制；Multi-Token Prediction 模块则以多步方式训练，保持训练与推理的一致性，提升了推测解码的接受率；训练稳定性方面则通过保留 Zero-Centered RMSNorm 并对 norm weight 施加 weight decay、注意力输出门控机制，以及 MoE 路由参数初始化的归一化策略加以保障。

下载日志已记录，仅供内部参考

3、AI 融资动向：鹏行智能完成超 9 亿美元首轮融资，投后估值超 63 亿美元

8 月 24 日，小鹏汽车旗下人形机器人业务宣布完成首轮融资，融资金额超过 9 亿美元，投后估值超过 63 亿美元。本轮融资由 IDG 资本领投，腾讯、阿里巴巴、高榕资本等多家机构参与。此次融资是小鹏机器人业务成立以来完成的首轮外部融资，也成为近期国内人形机器人领域规模较大的融资事件之一。

小鹏机器人主要聚焦人形机器人及具身智能技术的研发与产业化，业务覆盖机器人本体、运动控制、人工智能模型以及相关软硬件系统。此前，小鹏已推出人形机器人 IRON，并持续推进机器人在运动能力、智能交互和复杂环境作业等方面的技术研发。通过持续迭代产品和技术，小鹏正在逐步完善人形机器人从研发到生产制造的产业体系。

根据公司披露，本轮融资资金将主要投入产品研发、AI 模型训练、数据采集以及生产设施建设等环节。其中，AI 模型训练和数据采集是人形机器人实现自主感知、决策和执行的重要基础，而生产设施建设则将进一步支持后续产品的规模化制造。

在产品研发方面，小鹏机器人持续推进人形机器人核心技术能力建设，包括机器人运动控制、环境感知、智能决策和人机交互等。通过将人工智能技术与机器人硬件结合，小鹏希望进一步提升机器人在真实环境中的自主作业能力，并推动相关产品从技术验证逐步走向实际应用。

从融资参与机构来看，本轮融资汇集了 IDG 资本、腾讯、阿里巴巴、高榕资本等机构。多家头部投资机构参与本轮融资，为小鹏机器人后续研发投入、团队建设和生产设施布局提供资金支持。

在产业化方面，小鹏机器人正在推进人形机器人产品的量产准备工作，并围绕研发、供应链、生产制造和数据体系进行持续建设。随着本轮融资完成，小鹏机器人将进一步加大在人工智能模型、机器人硬件及生产能力等方面的投入，推动人形机器人产品研发及商业化进程。

图表 9：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
小鹏机器人	人形机器人	IDG 资本	首轮	9 亿美元	9 亿美元	63 亿美元

下载日志已记录，仅供内部参考

Instinct	AI 安全	A16z	B 轮	2.5 亿美元	4.1 亿美元	25 亿美元
Gatik	自动驾驶货运 AI	QIA, ARK Investment management	D 轮	2 亿美元	5 亿美元	18 亿美元

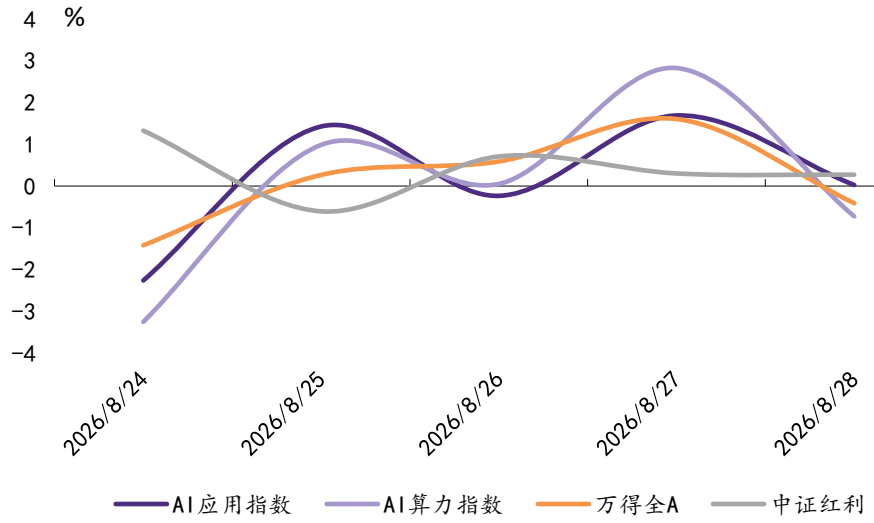
资料来源：wind，Saasverse，华鑫证券研究

下载日志已记录，仅供内部参考

4、 行情复盘

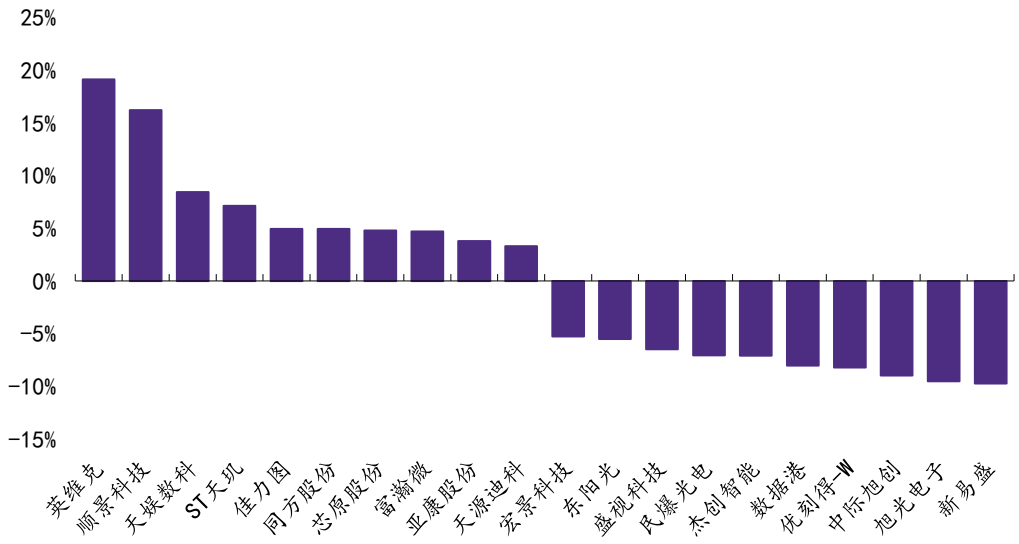
上周（2026.8.24-2026.8.28 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 1.69%/2.83%/1.61%/1.33%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-2.26%/-3.25%/-1.42%/-0.61%。AI 算力指数内部，英维克以 19.14%录得上周最大涨幅，新易盛以-9.73%录得上周最大跌幅。AI 应用指数内部，金桥信息以 12.45%录得上周最大涨幅，科沃斯以-13.3%录得上周最大跌幅。

图表 10：上周（2026.8.24-2026.8.28 日）指数日涨跌幅



资料来源：wind，华鑫证券研究

图表 11：上周（2026.8.24-2026.8.28 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 12：上周（2026.8.24-2026.8.28 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

近日，英伟达发布2027财年第二季度财报。当季收入962.21亿美元，同比增长106%、环比增长18%，增速连续第四个季度环比抬升；其中数据中心收入890亿美元，同比增长117%、环比增长18%，超大云厂商（hyperscale）收入490亿美元、环比增长13%，边缘计算收入72亿美元，同比增长27%；GAAP与Non-GAAP毛利率均为75.0%（上年同期72.4%/72.5%）；GAAP净利润596.88亿美元，同比增长126%，摊薄EPS2.46美元，同比增长128%；当季自由现金流213.41亿美元（上季485.54亿美元），环比回落主因应收账款增加及为VeraRubin量产备货；当季回购与股息合计回馈股东约260亿美元；当季新发债约249亿美元。指引方面，3QFY27收入预期1,080亿美元±2%（不含中国数据中心计算收入），毛利率74.0%±50bp；公司预计FY2028收入将增长约70%，并强调该预期为供给受限口径。公司CFO在财报电话会上表示，当前云行业积压订单已超2万亿美元，并预计前五大超大云厂商2026/2027年资本开支分别接近8,000亿/1.3万亿美元。财报同步披露，公司与Apollo、BlackRock、Blackstone、Brookfield、高盛、KKR达成战略合作，拟设立独立算力融资平台、撬动超5,000亿美元第三方资本投向AI基础设施；VeraRubin平台已进入全面量产，机架在CoreWeave、GoogleCloud、MicrosoftAzure、OCI、Nebius等伙伴处上线运行。

本次财报的核心信号在于需求端与投入端同时进一步加速：需求端，去年同期全球AI算力建设基本由单一大客户拉动，本季已变为多家前沿实验室集体大规模扩建，叠加AI创业公司、各国主权AI项目及机器人等新需求放量；1,080亿美元的下季指引对应环比约12%的增长，表明算力需求仍在加速兑现。投入端，公司首次通过"自己发债+联合六大资管机构设立算力融资平台"的方式，为下游客户筹措超5,000亿美元建设资金，商业模式从单纯出售芯片进一步延伸至为客户提供融资支持；库存与应收账款同步大增，也指向VeraRubin量产爬坡的确定性较高。我们认为，英伟达财报所呈现的需求信号与算力融资平台的设立相互印证，全球AI基础设施投入仍处加速期，公司管理层对前五大云厂商2027年资本开支接近1.3万亿美元的预期将进一步打开算力产业链空间。

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13：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

下载日志已记录，仅供内部参考

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 14：重点关注公司及盈利预测

公司代码	名称	2026-09-01	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	569.75	-0.30	0.30	0.60	-1899.17	1899.17	949.58	买入
301196.SZ	唯科科技	95.40	2.05	1.92	2.39	39.28	49.65	39.86	买入
301369.SZ	联动科技	138.30	0.48	1.08	2.62	288.13	128.06	52.79	买入
603859.SH	能科科技	38.88	0.92	1.05	1.32	42.26	37.03	29.45	买入
688615.SH	合合信息	99.95	3.24	4.22	5.25	30.85	23.68	19.04	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。