

2026 年 07 月 28 日



华鑫证券
CHINA FORTUNE SECURITIES

Anthropic 发布 Claude Opus 5 谷歌发布三款 Flash 模型

——计算机行业周报

推荐(维持)

投资要点

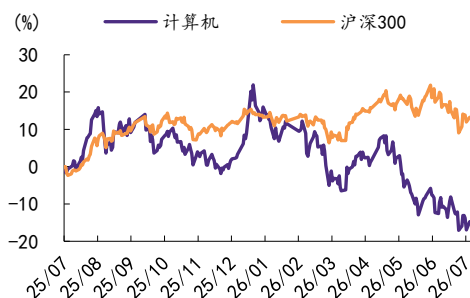
分析师：任春阳 S1050521110006

rency@cfsc.com.cn

行业相对表现

表现	1M	3M	12M
计算机(申万)	-2.7	-14.9	-14.2
沪深 300	-3.4	-1.2	13.7

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：月之暗面正式发布 KimiK3，StepAOS 与 STEPXNeo 发布》2026-07-22
- 2、《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
- 3、《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

算力：算力租赁价格平稳，Anthropic 发布新一代旗舰大模型 Claude Opus 5

2026 年 7 月 25 日，Anthropic 正式推出新一代旗舰级大模型 Claude Opus 5。该模型在性能上与此前广受关注的 Claude Fable 5 极为接近，但价格仅为后者的一半——输入 5 美元/百万 Token、输出 25 美元/百万 Token，与上一代 Opus 4.8 持平。模型同步引入快速模式及 Effort 调节机制，并在编程、复杂推理及科学研究等领域实现了显著的性能跃升，在 Frontier-Bench v0.1、ARC-AGI 3 等多项评测中位居榜首，并在生命科学任务中较前代提升明显。同时，该模型是 Anthropic 迄今为止对齐程度最高的模型，非对齐行为得分仅 2.3，为近期主流模型最低值。

AI 应用：Kimi 周访问量环比+86.92%，谷歌发布三款 Flash 模型

2026 年 7 月 22 日，谷歌正式发布了 Gemini 3.6 Flash、Gemini 3.5 Flash-Lite 以及 Gemini 3.5 Flash Cyber 三款新型模型，将重心转向了提升 Agent workflows 中的 Token 效率、响应速度以及运行可靠性。

AI 融资动向：CuspAI 完成 4.5 亿美元 B 轮融资，投后估值达 26 亿美元

2026 年 7 月 21 日，英国 AI 原生材料发现公司 CuspAI 宣布完成 4.5 亿美元 B 轮融资，投后估值达 26 亿美元。本轮融资由 Kleiner Perkins、NEA 联合领投，Bezos Expeditions 跟投，以及多家新老投资方共同参与。公司核心产品 MIRA 平台旨在将传统耗时数年的材料研发周期压缩至六个月以内，平台底层技术由 Meta UMA 系列模型及自研 kUPS 工具包共同驱动。同期，公司发布 "AI Materials Foundry" 全球网络计划，当前已吸引超 45 家行业领军企业加盟。

投资建议

2026 年 7 月 22 日，谷歌发布 2026Q2 财报：当季实现总营收 1197.96 亿美元，同比增长 24%；净利润 1121.07 亿美元，摊薄后每股收益 9.11 美元。分业务看，谷歌云增长最为强劲，当季营收 247.68 亿美元，同比大幅增长 82%，营业利润率由 20.7% 提升至 35.6%，在手积压订单由一季度的 4600 亿美元增

至 5140 亿美元；公司在财报电话会上表示，旗下 GeminiEnterprise 已渗透近 90%的《财富》100 强企业。谷歌服务营收 945.40 亿美元（同比+15%），其中搜索业务营收 632.71 亿美元（同比+17%），YouTube 广告营收 110.55 亿美元（同比+13%）。资本开支方面，公司单季资本开支 449 亿美元，同比翻倍；管理层在电话会上表示，其中约 60%投向服务器采购、40%投向数据中心与网络设备，并将全年资本开支指引由 1800 - 1900 亿美元上调至 1950 - 2050 亿美元。模型层面，公司表示正全力推进新一代旗舰模型 Gemini4 的预训练并保持按月迭代节奏；据 CEO 皮查伊披露，API 每分钟 Token 处理量已由上季度的 160 亿提升至 220 亿，公司并于当季首次向客户数据中心交付 TPU 系统。

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，谷歌云 82%的同比增速、5140 亿美元积压订单及 Token 消耗的爆发式增长表明企业级 AI 需求仍在加速，且供给持续紧张——管理层表示 Q3 起将租用第三方算力作为过渡，进一步印证算力缺口为结构性而非周期性。从传导路径看，全年 2000 亿美元级资本开支将直接转化为上游采购需求：约 60%投向服务器、40%投向数据中心与网络设备，叠加 TPU 首次对外交付客户数据中心并开辟新收入模式。我们认为，本轮资本开支上行由真实需求驱动：订单积压与对外租算力并存，表明算力供给仍滞后于需求增长，资本开支具备扎实的基本面支撑与持续性；在此背景下，AI 服务器、光模块与高速互联、数据中心网络设备、IDC 及电力配套等上游硬件环节将按资本开支投向结构直接承接订单，需求传导的确定性较强。

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	2026-07-28 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	427.12	-0.30	0.30	0.60	-1423.73	1423.73	711.87	买入

301196.SZ	唯科科技	77.30	2.53	3.34	3.98	30.55	23.14	19.42	买入
301369.SZ	联动科技	140.20	0.48	1.08	2.62	292.08	129.81	53.51	买入
603859.SH	能科科技	36.58	0.92	1.05	1.32	39.76	34.84	27.71	买入
688615.SH	合合信息	93.03	3.24	4.22	5.25	28.71	22.05	17.72	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，ANTHROPIC 发布新一代旗舰大模型 CLAUDE OPUS 5..... 5

1.1、 Tokens 跟踪..... 5

1.2、 数据跟踪：阿里云发布语音合成大模型 Qwen-Audio-3.0-TTS，同步推出轻量应用服务器智能体专用型实例..... 7

1.3、 产业动态：Anthropic 发布新一代旗舰大模型 Claude Opus 5 8

2、 AI 应用动态：KIMI 周访问量环比+86.92%，谷歌发布三款 FLASH 模型..... 13

2.1、 周流量跟踪：Kimi 周访问量环比+86.92%..... 13

2.2、 产业动态：谷歌发布三款 Flash 模型，聚焦效率与成本优化 13

3、 AI 融资动向：CUSPAI 完成 4.5 亿美元 B 轮融资，投后估值达 26 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）..... 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：大模型定价对比图 9

图表 4：CLAUDE OPUS 5 编程能力测试结果横向对比图..... 9

图表 5：CLAUDE OPUS 5 复杂推理能力测试结果横向对比图..... 10

图表 6：CLAUDE OPUS 5 安全能力测试结果横向对比图..... 12

图表 7：2026.7.17-2026.7.23 AI 相关网站流量..... 13

图表 8：GEMINI 3.6 FLASH 在能力、速度和成本上与其他模型对比..... 14

图表 9：3.1 PRO、3.5 FLASH 和 3.6 FLASH 的对比..... 14

图表 10：GEMINI 3 FLASH 和 GEMINI 3.5 FLASH-LITE 对比情况..... 15

图表 11：GEMINI 3.5 FLASH CYBER 在 CYBERGYM 测试中的表现..... 15

图表 12：上周 AI 初创公司融资动态 17

图表 13：上周（2026.7.20-2026.7.24 日）指数日涨跌幅..... 18

图表 14：上周（2026.7.20-2026.7.24 日）AI 算力指数内部涨跌幅度排名 18

图表 15：上周（2026.7.20-2026.7.24 日）AI 应用指数内部涨跌幅度排名 19

图表 16：FICONTEC2025 年年中至今公告订单..... 20

图表 17：重点关注公司及盈利预测 21

1、算力动态：算力租赁价格平稳，Anthropic 发布新一代旗舰大模型 Claude Opus 5

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 7 月 20 日至 7 月 26 日，周度 Token 消耗量有所下降，消耗量为 58T，环比上周减少 6.9%。在 Tokens 规模 Leaderboard 前五名中，Xiaomi 的 MiMo-V2.5 以 10.5T tokens 位居榜首；DeepSeek 的 DeepSeek V4 Flash 以 6.37T tokens 位列第二；Tencent 的 Hy3 以 3.94T tokens 位居第三；Z.AI 旗下的 GLM 5.2 以 3.29T tokens 位居第四；DeepSeek 的 DeepSeek V4 Pro 以 3.17T tokens 位居第五。

文本类模型的市场份额，Google 以 975M requests 占据 27.5% 的份额，位居第一；DeepSeek 以 816M requests 占据 23.0%，位列第二；OpenAI、Qwen、Xiaomi 则分别以 526M、181M、164M requests，对应占据 14.8%、5.1%、4.6% 的市场份额。

图像生成类模型调用量小幅上升，调用量为 55M requests，环比上周增加 4.12%。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)、Nano Banana 2 (Gemini 3.1 Flash Image Preview) 分别以 1.36M、851K、689K requests 位居第一、第二和第三；OpenAI 的 GPTImage 2 以 562K requests 位列第四；Google 的 Nano Banana Pro (Gemini 3 Pro Image) 以 508K requests 位列第五。从市场份额来看，Google 以 4.2M requests 占据 75.6% 的份额，位居第一；OpenAI 以 771K requests 占据 13.9%，位列第二；xAI、ByteDance Seed、Black Forest Labs 则分别以 266K、133K、121K requests，对应占据 4.8%、2.4%、2.2% 的市场份额。

嵌入模型调用量大幅上升，调用量为 274M requests，环比上周增加 18.1%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 94.5M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 66.8M requests 位列第二；BAAI 的 bge-m3 以 17M requests 位列第三；Perplexity 的 Embed V1 0.6B 以 15.2M requests 位列第四；OpenAI 的 Text Embedding 3 Large 以 14.3M requests 位列第五。从市场份额来看，OpenAI 以 112M requests 占据 40.8% 的份额，位居第一；Qwen 以 74.2M requests 占据 27.1%，位列第二；Google、Perplexity、BAAI 则分别以 34.9M、22M、19.8M requests，对应占据 12.8%、8.0%、7.2% 的市场份额。

重排序模型调用量有所上升，调用量为 3.26M requests，环比上周增加 2.52%。在 Rerank Leaderboard 中，NVIDIA 的 Llama Nemotron Rerank VL 1B V2 以 1.64M requests 位列榜首；Cohere 的 Rerank v3.5、Rerank 4 Fast、Rerank 4 Pro 以 704K、616K、300K requests 位列第二、第三和第四。从市场份额来看，NVIDIA 以 1.64M requests 占据 50.3% 的份额，位居第一；Cohere 以 1.62M requests 占据 49.7% 的份额，位列第二。

视频生成类模型调用量有所下降，调用量为 138K requests，环比上周减少 2.82%。在 Video Leaderboard 前五名中，Google 的 Veo 3.1 Fast 以 37K requests 位列第一；ByteDance 的 Seedance 2.0 以 22K requests 位居第二；Google 的 Veo 3.1 Lite、Veo 3.1 分别以 18K、

12K requests 位列第三、第四；ByteDance 的 Seedance 2.0 Fast 以 12K requests 位居第五。从市场份额来看，Google 以 67K requests 占据 49.0% 的份额，位居第一；ByteDance 以 40K requests 占据 29.0%，位列第二；xAI、KwaiVGI、Alibaba 则分别以 11K、11K、8K requests，对应占据 8.1%、7.8%、5.7% 的市场份额。

语音生成类模型调用量显著上涨，调用量为 1.53M requests，环比上周增加 56.76%。在 Speech Leaderboard 前五名中，Hexgrad 的 Kokoro 82M 以 648K requests 位列第一；Google 的 Gemini 3.1 Flash TTS Preview 以 457K requests 位居第二；xAI 的 Grok Voice TTS 1.0 以 335K requests 位列第三；Microsoft 的 MAI-Voice-2 以 42K requests 位列第四；Deepgram 的 Aura-2 以 8.25K requests 位列第五。从市场份额来看，Hexgrad 以 648K requests 占据 42.3%，位列第一；Google 以 457K requests 占据 29.9% 的份额，位居第二；xAI、Microsoft、MiniMax 则分别以 335K、46K、15K requests，对应占据 21.9%、3.0%、1.0% 的市场份额。

语音转写类模型调用量巨幅上涨，调用量为 8.32M requests，环比上周增加 109.05%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3 以 4.58M requests 位列第一；OpenAI 的 Whisper Large V3 Turbo 以 1.57M requests 位列第二；OpenAI 的 GPT-4o Mini Transcribe 以 745K requests 位列第三；Qwen 的 Qwen3 ASR Flash 以 412K requests 位列第四；Microsoft 的 MAI-Transcribe 1.5 以 302K requests 位列第五。从市场份额来看，OpenAI 以 7.22M requests 占据 86.7% 的份额，位居第一；Qwen 以 412K requests 占据 4.9%，位列第二；Microsoft、Mistral AI、NVIDIA 则分别以 302K、158K、105K requests，对应占据 3.6%、1.9%、1.3% 的市场份额。

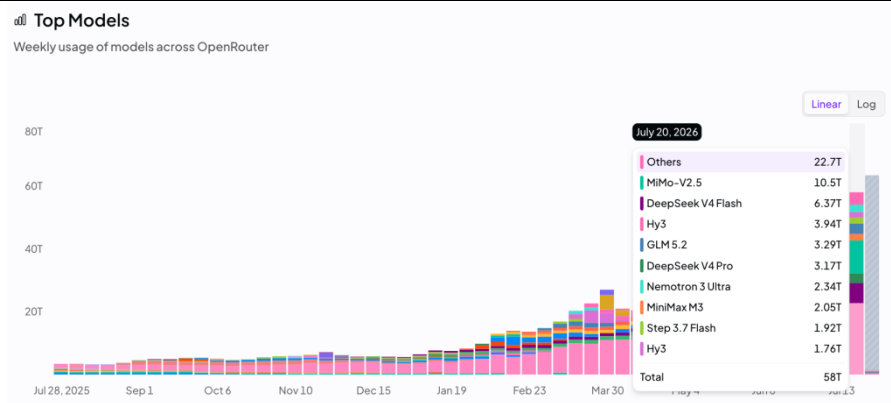
* 统计口径说明：以上数据均来源于 OpenRouter。基于官方统计口径调整，当前 Rankings 采用 Token 与 Request 两套统计维度：其中，仅 Tokens 规模 Leaderboard 仍以模型 Token 消耗量为统计口径，其余类目榜单及市场份额板块则统一采用 Request 调用量为统计口径，两者统计维度不同，不宜直接对应或进行横向比较。

7月21日，阿里千问推出新一代图像生成模型 Qwen-Image-3.0。在内容承载上，该模型最大支持 4.5k token 输入，可精准生成包含海量文字的复杂版面，如报纸、分镜和试卷等；在细节表现上，该模型实现了 10px 小字的清晰渲染，能够逼真还原毛孔、发丝等微观质感；在语言能力上，该模型原生支持 12 种语言渲染，可结合丰富的世界知识构建网页、游戏及直播界面仿真等应用场景。

7月21日，北京市经信局披露，上半年全市数字经济增加值增速达 7.8%，其中核心产业增速为 9.8%。下半年，北京将围绕 Token 经济制定专项政策，在 Token 生产、分发、应用等关键环节布局建设 Token 工厂与分发平台，并推动其在工业、教育、文旅等领域创新应用。算力方面，全市上半年新增智能算力 2.2 万 P、总算力达 8.2 万 P，力争下半年再增 5 万 P，年内突破 13 万 P。

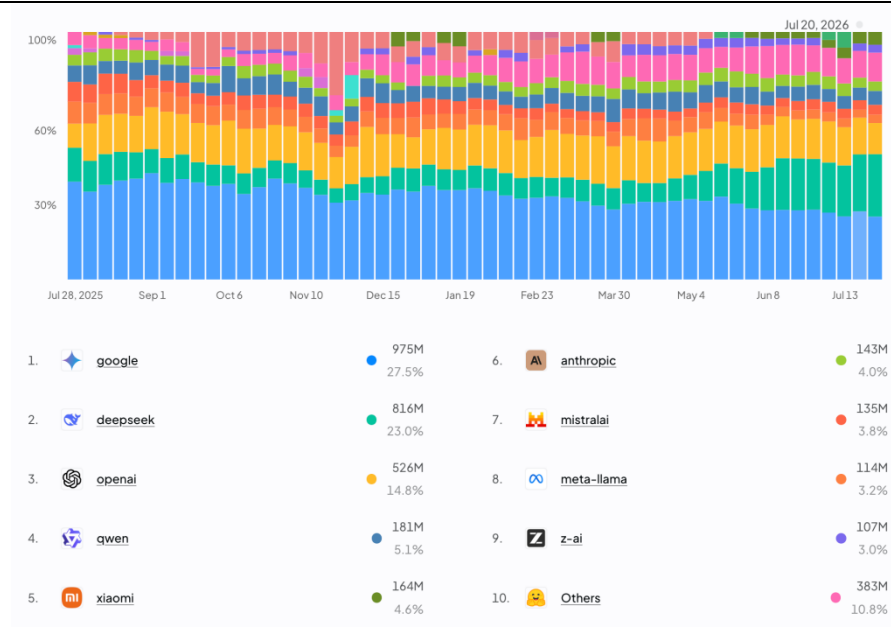
7月23日，腾讯云正式发布面向企业研发流程的云端智能体 CodeBuddy NPC。开发者只需派发任务，NPC 即可自主完成方案规划、代码开发、提交 PR、执行测试，并根据 CI 结果持续迭代，直至交付可验收成果。该智能体依托腾讯云研发平台 CNB，能够直接利用代码仓库、Issue、PR、CI/CD 等研发上下文，在现有流程中持续推进任务。经过持续优化，当前官方 NPC 的首轮 Token 消耗已从早期的 2 万多个降至约 2000，降幅超过 90%。

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：阿里云发布语音合成大模型 Qwen-Audio-3.0-TTS，同步推出轻量应用服务器智能体专用型实例

2026 年 7 月 20 日，阿里云正式发布语音合成大模型 Qwen-Audio-3.0-TTS。该模型在精细控制、指令遵循、多语种覆盖及声学鲁棒性等方面均实现显著升级，同时提供 Flash、Plus 两个版本。其中，Flash 版面向实时交互场景，首包延时低至 300 毫秒级别，Plus 版则面向高质量生成场景。当前，两款模型已在阿里云百炼开放调用。

细粒度标签控制方面，该模型支持 freestyle 自由风格模式，用户可通过角色设定自由调控情绪基调、场景氛围和语速节奏。同时，新模型进一步支持结构化标签控制，用户可

在合成文本中嵌入[gasp]（抽气）、[giggle]（轻笑）、[angry]（愤怒）等情感与动作标记，将控制精度细化至单字级别，以此适配叙事、游戏、配音等多种场景。

多语种与方言方面，新模型语言能力大幅扩展，在覆盖中、英、日、韩、德等 16 种语言的基础上，新增阿拉伯语、越南语、马来语及菲律宾语。此外，模型还新增对广东、重庆、上海、四川等 20 种中文方言的支持，通过针对性强化训练，使模型在方言韵律和声调还原上接近母语者水平。

声学鲁棒性方面，该模型通过真实声学仿真训练，在克隆流程中嵌入语音增强能力，具备在高噪声、高混响条件下过滤干扰的能力。音频质量方面，新模型提供开箱即用的精品音色库，输出音频从 24kHz 提升至 48kHz，单次语音合成可长达 3 分钟。

当前，Qwen-Audio-3.0-TTS-Plus 已在全球第三方权威榜单 Artificial Analysis 中登顶榜首，其预览版 Fun-Realtime-TTS 也曾一个月前拿到榜首席位。

同日，阿里云推出轻量应用服务器智能体专用型实例，旨在解决用户以往需分别采购云服务器、申请 API Key、配置公网 IP 及带宽的割裂链路问题。该产品将 vCPU、内存、云盘、200Mbps 峰值带宽（免流量费）与大模型 Tokens（1 亿至 32 亿不等）打包为预付费套餐，实现计算、存储、网络与 Token 额度的统一调配，控制台支持一键部署 OpenClaw、Hermes 等个人助理以及 Dify、MaxKB 等企业级平台。

该实例搭载了专为 Agent 场景设计的操作系统 ANOLISA（Alibaba Cloud Linux 4 Agentic 版），可在主流场景中节省 30% 的 Token，显著降低 Token 浪费，同时将主流 Bench 分数提升 10%、Agent 执行时长降低 30%、冷启动时长降低 20%，有效优化 Agent 运行表现，并具备三层纵深防御架构的全链路安全守护能力以及工作区快照恢复等功能，支持 Agent 长期稳定运行。此外，全系实例标配 200Mbps 峰值公网带宽且流量免费，可有效缓解对外服务时的网络瓶颈与成本焦虑。

定价方面，以入门级 2 核 2G 加 2 亿 Token 规格为例，包月购买享受 5 折优惠，相比分开购买，在 AI Coding 场景每月可节省 18%，在内容生成等高消耗场景每月可节省 68%。产品提供从 2 核 0.5G+1 亿 Tokens 到 16 核 64G+32 亿 Tokens 共 9 款阶梯规格，其中三档套餐享专属优惠，覆盖个人开发者到中小企业。当前，该产品已在北京、上海、12 州等地域上线，未来将支持更多国内及海外地域。

1.3、产业动态：Anthropic 发布新一代旗舰大模型 Claude Opus 5

2026 年 7 月 25 日，Anthropic 正式推出其新一代旗舰级大模型 Claude Opus 5。该模型在性能方面与此前广受关注的 Claude Fable 5 极为接近，但整体价格仅为 Fable 5 的一半。定价方面，该模型以每百万 Token 输入 5 美元、每百万 Token 输出 25 美元的价格，与上一代 Opus 4.8 持平。此外，该模型还提供了快速模式，运行速度约为默认模式的 2.5 倍，相应价格为基础价格的 2 倍，以满足高时效性场景需求。

官方表示，该模型高度适配日常高频使用场景，相较以往模型具备更高的工作效率。当前，该模型已经成为 Claude Max 的默认模型，并以其性能优势在 Claude Pro 中超越其他模型。与此同时，该模型同步引入了 Effort 调节机制，允许用户根据任务复杂度和成本预算灵活调节模型的推理深度，从而实现模型性能与经济效益之间的动态平衡。

图表 3：大模型定价对比图

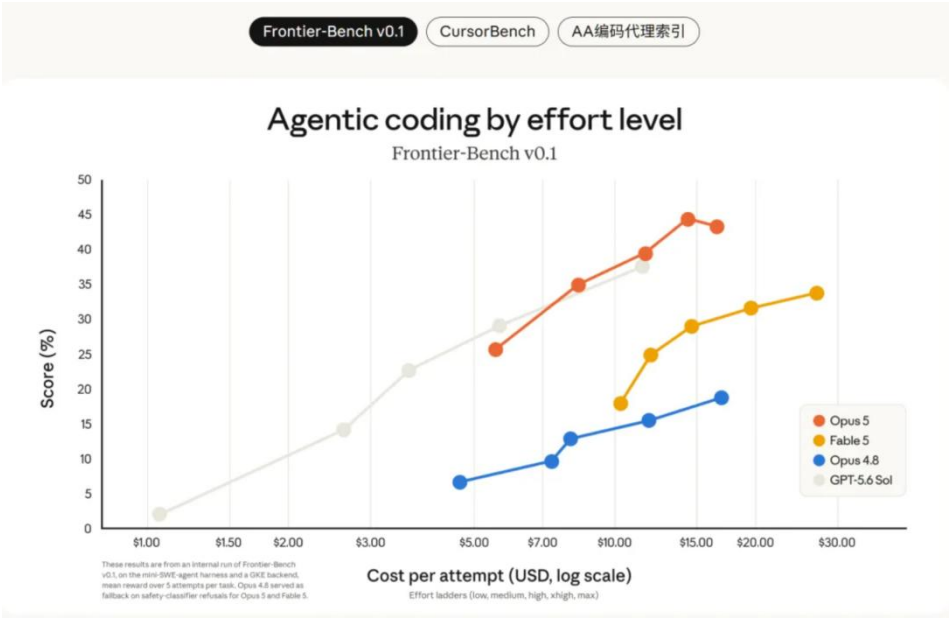
模型	输入单价（百万tokens）	输出单价（百万tokens）
GLM-5.2	8元	28元
Grok 4.5	2美元（约合13.59元）	6美元（约合40.78元）
Claude Opus 4.8	5美元（约合33.98元）	25美元（约合169.92元）
GPT-5.5	5美元（约合33.98元）	30美元（约合203.9元）
Claude Fable 5	10美元（约合67.97元）	50美元（约合339.83元）
Claude Opus 5	5美元（约合33.98元）	25美元（约合169.92元）

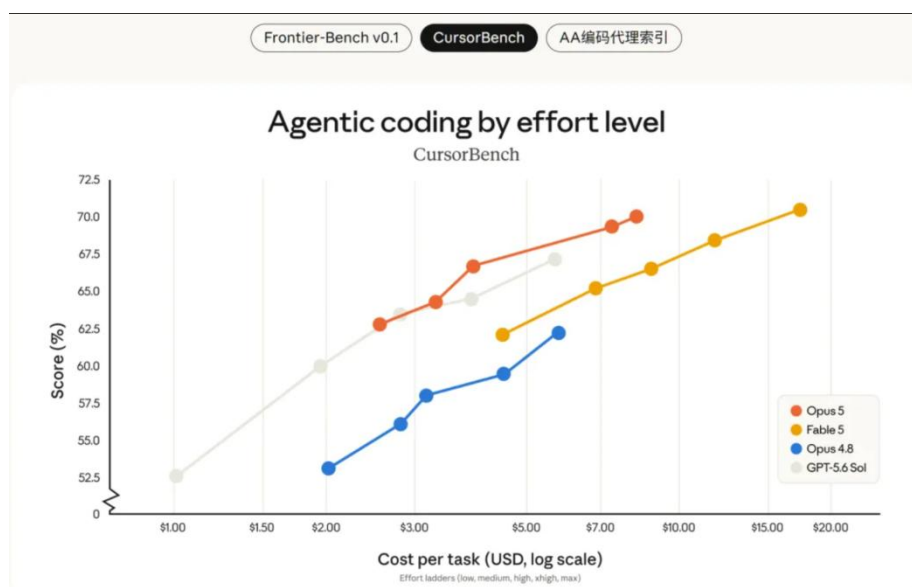
资料来源：智东西，华鑫证券研究

新一代模型 Opus 5 在高价值软件工程、复杂推理及科学研究等关键领域均实现了显著性能跃升。

编程能力方面，该模型在 Frontier-Bench v0.1 评估中超越了包含 Fable 5、GPT-5.6 Sol 在内的其他模型，在降低单项任务成本的同时，性能达到 Opus 4.8 的两倍以上。在 CursorBench 3.2 测试中，该模型 Max Effort 模式下的性能得分峰值仅与 Fable 5 相差 0.5%，但成本仅为后者的一半，不仅如此，该模型在相同成本下的性能表现均优于其他模型，展现出了极高的性价比优势。

图表 4：Claude Opus 5 编程能力测试结果横向对比图



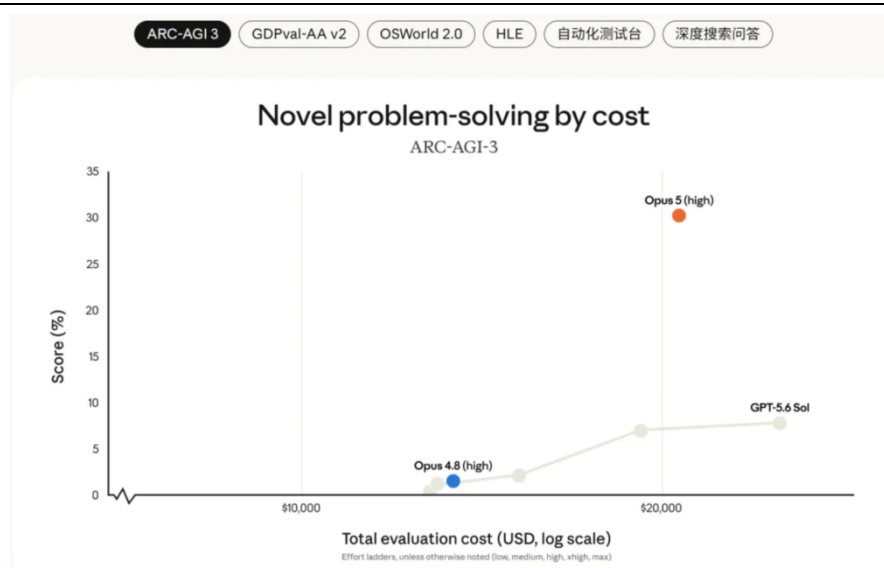


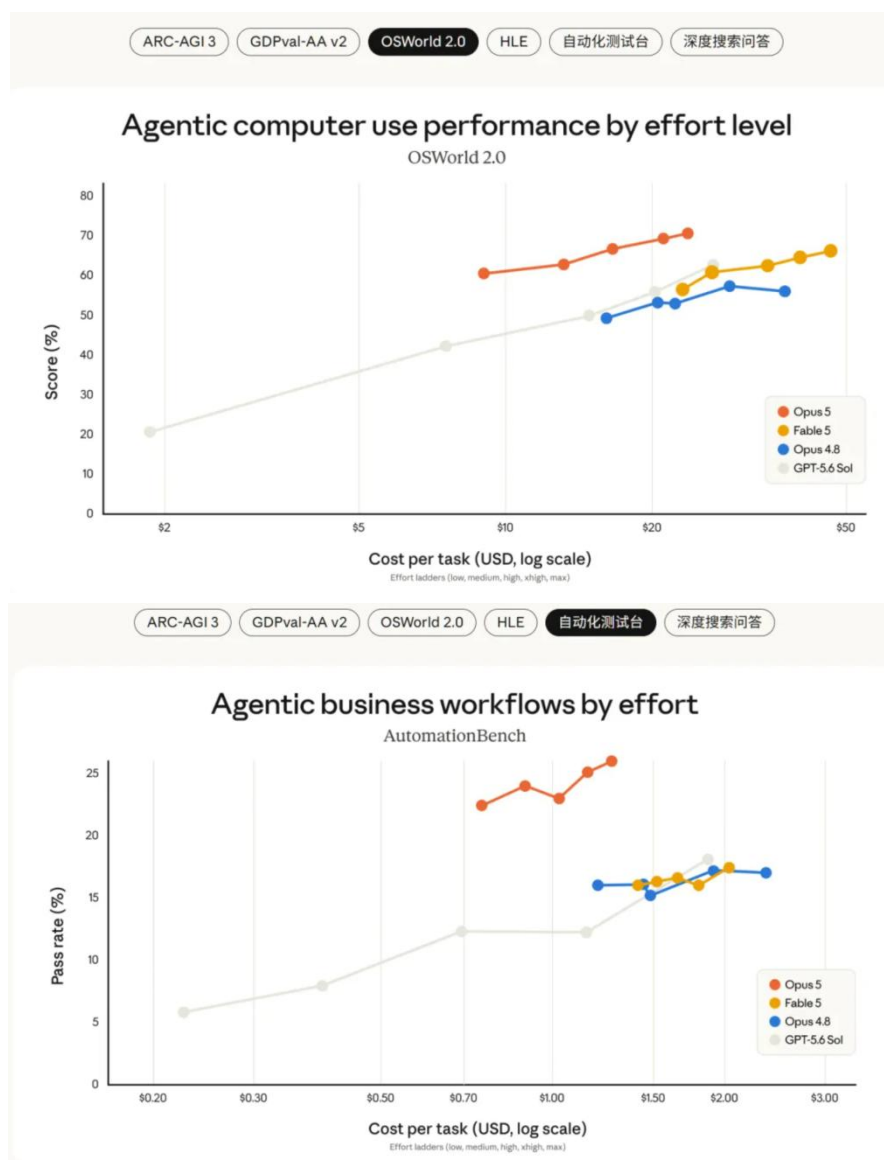
资料来源：智东西，华鑫证券研究

在复杂推理领域，该模型在考验模型面向全新未知问题解决能力的 ARC-AGI 3 测试中，以第二名模型 3 倍的得分拔得头筹；在衡量模型商业自动化任务完成能力的 Zapier AutomationBench 测试中，该模型以第二名模型 1.5 倍的通过率，远超其他模型；在计算机操作使用基准测试 OSWorld 2.0 中，任意同等成本下，该模型表现均优于其他所有模型，且达成 Fable 5 的最佳成绩仅需消耗后者略高于三分之一的成本。此外，该模型在包含 HLE、AutomationBench、DeepSearchQA 在内的多项测试中，均显示出了性能优、性价比高的特性。

在科学研究方面，该模型在有机化学任务中得分较 Opus 4.8 提升 10.2 个百分点，在蛋白质相关任务中得分较 Opus 4.8 提升 7.7 个百分点，在涵盖结构生物学、有机化学、生物信息学等领域的全部生命科学评估中，该模型表现均超越前代。此外，新模型具备自动生成空气动力学风洞风流图解、细胞结构互动图示等复杂可视化内容的能力，在视觉输出生成方面能力大幅增强。

图表 5：Claude Opus 5 复杂推理能力测试结果横向对比图



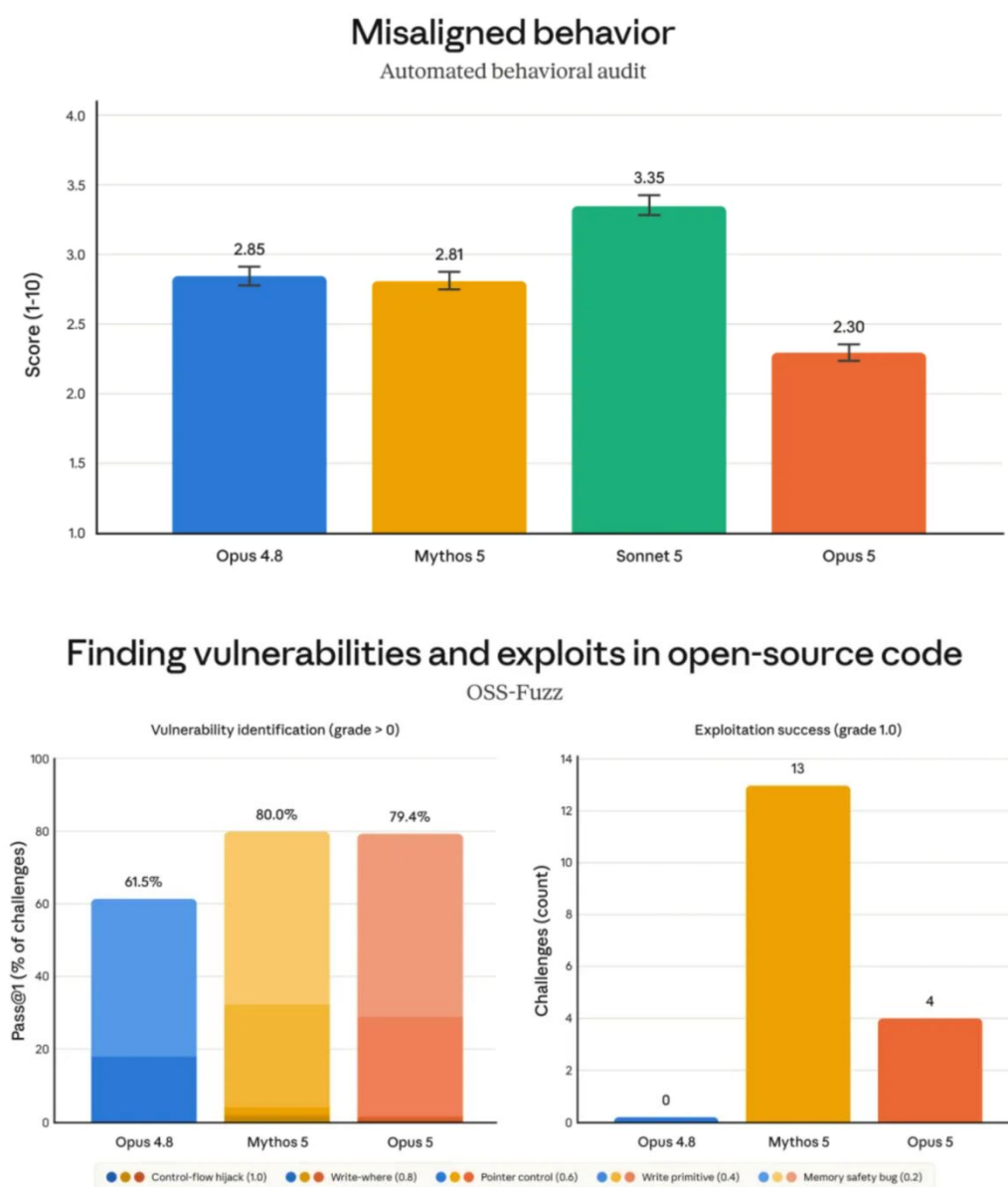


资料来源：智东西，华鑫证券研究

与此同时，该模型在自我校验和审慎迭代方面的能力提升尤为突出，其主动性和严谨性在多场景测试中得到充分验证。例如，在未提供直接查看图纸接口的情况下，模型可自行编写计算机视觉管线，并从工程图纸中提取特征并完成 3D 模型重建；在面对热门开源包管理器中社区补丁遗漏的边缘情况时，模型能精准定位并修复根本原因，而不仅仅是修补表层问题；在为新建交易所构建数据接入源时，模型能自主构建测试套件，以模拟方式验证代码的准确性和鲁棒性。这些案例表明，Opus 5 不仅具备更强的任务执行能力，更在任务规划、工具调用和质量自检等方面展现出了主动性和系统性思维。

此外，官方披露，Opus 5 是公司迄今为止对齐程度最高的模型，其欺骗行为发生率最低，最不容易被误导或诱导滥用。在自动化行为审计中，该模型的整体非对齐行为得分仅为 2.3，为近期主流模型中的最低值。在网络安全领域，该模型在漏洞发现能力上已逼近 Mythos 5，但在漏洞利用方面仍大幅落后于后者。为防范潜在风险，Opus 5 的网络安全分类器会对基于二进制的漏洞扫描、渗透测试及 Exploits 生成等高风险操作进行拦截，并自动降级回退至 Opus 4.8 的安全策略。在生物安全方面，该模型则延续了与 Opus 4.8 类似的安全防护框架，官方称其目前可用于科学研究的最强通用模型。

图表 6: Claude Opus 5 安全能力测试结果横向对比图



资料来源：智东西，华鑫证券研究

2、AI 应用动态：Kimi 周访问量环比+86.92%，谷歌发布三款 Flash 模型

2.1、周流量跟踪：Kimi 周访问量环比+86.92%

本期（2026.7.17-2026.7.23）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1236.0M）、Bing（859.4M）和 Gemini（599.9M），访问量环比增速第一为 Kimi（86.92%）；平均停留时长前三位分别为 Character.AI（00:16:22）、Discord（00:11:15）和 NotionAI（00:08:33）；平均停留时长环比增速第一为 NotionAI（12.75%）。

图表 7：2026.7.17-2026.7.23 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	1236.0	1.06%	6:32	1.03%
Bing	搜索	微软	859.4	2.01%	7:16	0.23%
Gemini	聊天机器人	谷歌	599.9	-1.85%	7:02	-0.47%
Canva	在线设计	Canva	184.2	-0.49%	5:59	0.00%
Discord	游戏社区	微软	150.7	0.74%	11:15	0.15%
Github	代码托管	微软	145.8	3.11%	6:24	0.52%
Character.AI	聊天机器人	Character.AI	36.07	-4.65%	16:22	0.31%
DeepL	翻译工具	DeepL	23.74	-0.34%	2:28	-0.67%
Perplexity	AI 搜索	Perplexity	23.14	-4.02%	4:18	-0.39%
Kimi	聊天机器人	Moonshot AI	20.43	86.92%	7:26	-5.31%
QuillBot	释义工具	QuillBot	7.00	-2.94%	2:52	-0.58%
NotionAI	文本/笔记	Notion	2.615	-1.43%	8:33	12.75%
文心一言	聊天机器人	百度	0.267504	-13.11%	2:09	-7.19%

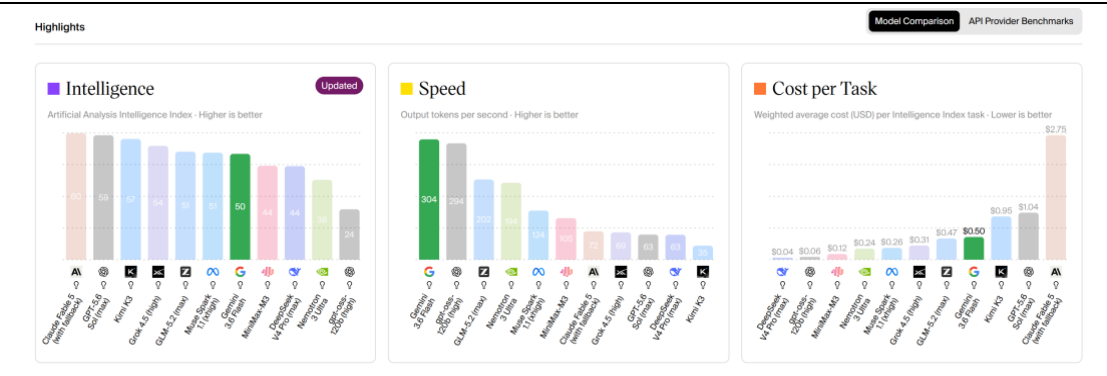
资料来源：similarweb, 华鑫证券研究

2.2、产业动态：谷歌发布三款 Flash 模型，聚焦效率与成本优化

2026 年 7 月 22 日，谷歌正式发布了 Gemini 3.6 Flash、Gemini 3.5 Flash-Lite 以及 Gemini 3.5 Flash Cyber 三款新型模型，将重心转向了提升 Agent workflows 中的 Token 效率、响应速度以及运行可靠性。

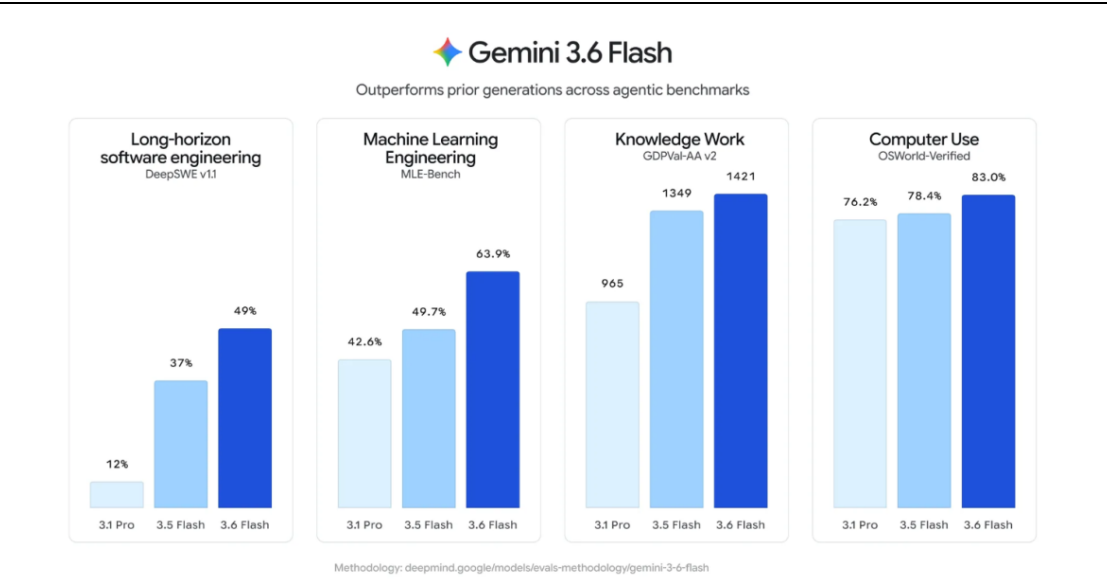
其中，Gemini 3.6 Flash 被定位为面向编程、知识密集型工作以及多模态处理任务的主力模型。根据 Artificial Analysis 的 Intelligence Index 显示，该模型在完成任务时所消耗的 Token 较前代 3.5 Flash 减少了 17%，而在 DeepSWE 测试中，其输出 Token 消耗量最多减少约 65%，这种显著的效率提升转化为每项 Agent 任务的运行成本下降。在定价层面，3.6 Flash 的输入价格为每 100 万个 Token 1.5 美元，输出价格为每 100 万个 Token 7.5 美元，输出价格低于 3.5 Flash。与此同时，该模型在软件工程 Agent 评测 DeepSWE v1.1 和机器学习工程测试 MLE-Bench 中的得分均高于 3.5 Flash，并且在测试 Computer Use 的 OSWorld-Verified 中也取得了 83.0% 的成绩。谷歌还展示了模型的应用案例，涵盖精准高效的金融数据与会议记录分析、基于多 Agent 协同的代码迁移自动化、面向 3D 流程的照片纹理提取，以及与 tldraw 离线编辑器创建交互式主题设计工具。安全层面，该模型还强化了针对化学、生物、放射性与核风险以及网络攻击滥用的前沿防护，在提升抗攻击能力的同时，尽可能减少对正常使用场景的错误干涉。

图表 8：Gemini 3.6 Flash 在能力、速度和成本上与其他模型对比



资料来源：智东西，华鑫证券研究

图表 9：3.1 Pro、3.5 Flash 和 3.6 Flash 的对比

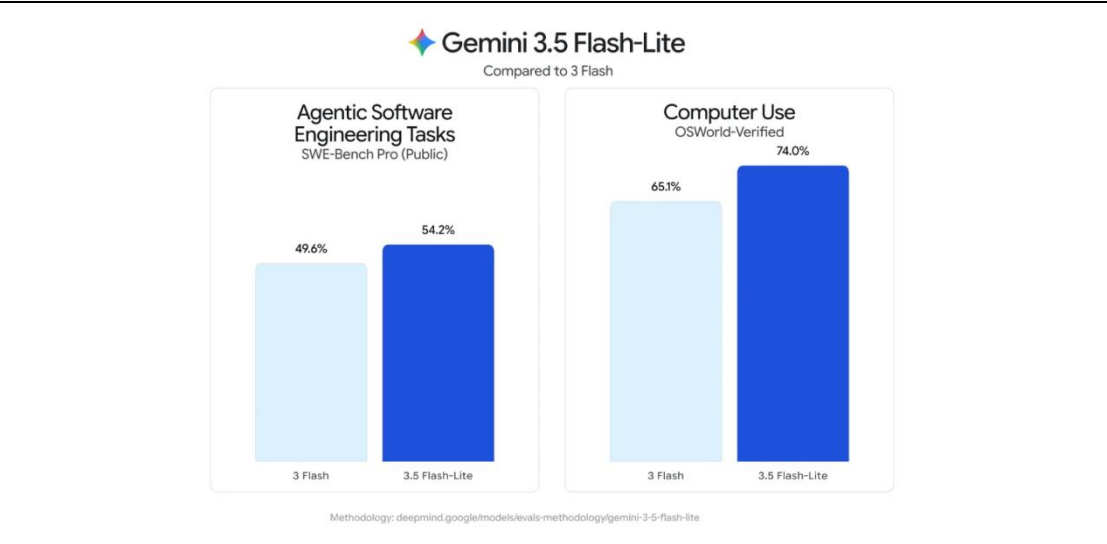


资料来源：智东西，华鑫证券研究

Gemini 3.5 Flash-Lite 作为 3.5 系列中速度最快且定价最低的模型，主要面向对低延迟和高吞吐量的应用场景，例如 Agent 搜索和文档处理。据 Artificial Analysis 的数据显示，该模型的生成速率达到了每秒 350 个输出 Token。其输入和输出价格分别为每 100 万个

Token 0.3 美元和 2.5 美元。与 3.1 Flash-Lite 相比，它在 Agent 终端编程测试 Terminal-Bench 2.1、长上下文测试 GDM-MRCR v2 以及知识工作测试 GDPval-AA v2 中的得分均有提升，成绩分别为 54%、72.2%、1140，甚至在部分测试中的表现超越了定位更高的 Gemini 3 Flash。在 SWE-Bench Pro 中，Gemini 3.5 Flash-Lite 和 Gemini 3 Flash 得分分别为 54.2% 和 49.6%；在 OSWorld-Verified 中，两者得分分别为 74.0% 和 65.1%。在实际业务场景中，如金融票据提取等任务里，该模型也成功地平衡了识别准确率、响应延迟与成本。

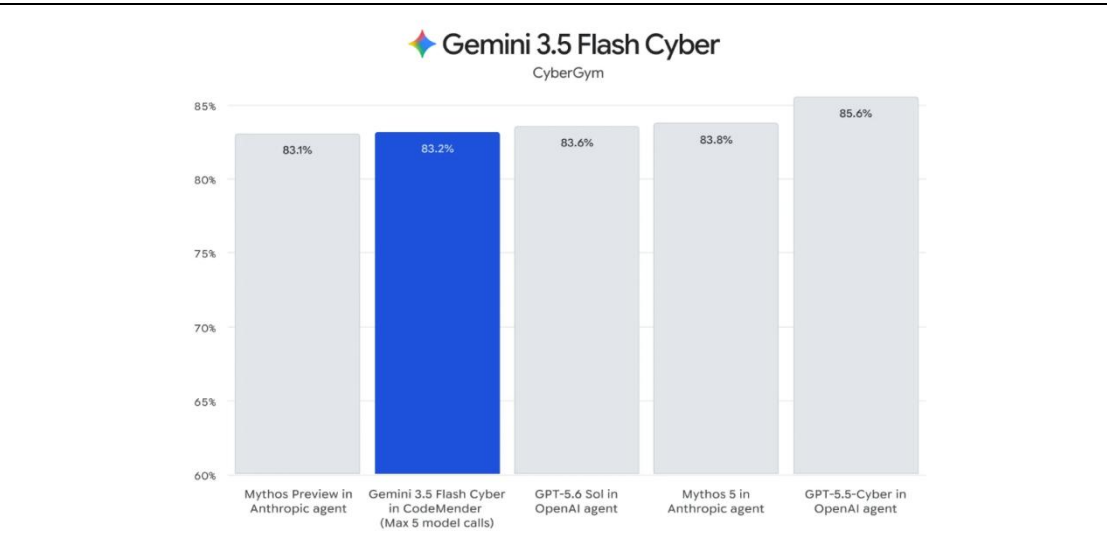
图表 10: Gemini 3 Flash 和 Gemini 3.5 Flash-Lite 对比情况



资料来源：智东西，华鑫证券研究

第三款模型 Gemini 3.5 Flash Cyber 则具有更强的专用属性，它基于 3.5 Flash 进行微调，专门用于网络安全漏洞的发现、验证与修复。该模型与谷歌代码安全 Agent CodeMender 协同工作，同时运行多个 3.5 Flash Cyber Agent 并合并结果。在 CyberGym 测试中，CodeMender 最多调用 5 次模型时，3.5 Flash Cyber 得分为 83.2%。不过，鉴于漏洞发现与修复技术具有明显的双重用途属性，谷歌目前并未向普通开发者开放这一模型，而是计划近期通过 CodeMender 平台启动小范围试点，仅供政府机构和可信合作伙伴使用。

图表 11: Gemini 3.5 Flash Cyber 在 CyberGym 测试中的表现



资料来源：智东西，华鑫证券研究

3、AI 融资动向：CuspAI 完成 4.5 亿美元 B 轮融资，投后估值达 26 亿美元

2026 年 7 月 21 日，英国 AI 原生材料发现公司 CuspAI 宣布完成 4.5 亿美元 B 轮融资，投后估值达到 26 亿美元。本轮融资由顶尖风投机构 Kleiner Perkins 与 NEA 联合领投，亚马逊创始人 Jeff Bezos 旗下 Bezos Expeditions 跟投，新投资方阵容涵盖 Glade Brook Capital Partners、Lux Capital、AMD Ventures 等多家机构，Temasek、Basis Set Ventures、Giant Ventures 等现有投资方持续跟投。

CuspAI 成立于 2024 年，是一家致力于通过人工智能加速新材料发现与开发的深度科技企业，总部设于英国剑桥，并在伦敦、阿姆斯特丹、柏林、东京、新加坡及美国设有业务团队。公司核心团队包括联合创始人兼首席执行官 Chad Edwards 博士、首席技术官及机器学习领域权威学者 Max Welling 教授，以及首席科学官 Aron Walsh 教授。其顾问委员会则汇聚了人工智能与半导体制造领域的泰斗级人物，包括诺贝尔奖得主 Geoffrey Hinton、图灵奖得主 Yann LeCun、前 ASML 总裁兼首席技术官 Martin Van Den Brink，以及前 BP 首席执行官 Lord John Browne 等，强大的学术与产业背景为公司提供了前沿的学术视野与产业战略支撑，夯实了技术研发的底层根基。

核心产品方面，公司推出了专为工业级材料发现设计的 MIRA AI 平台，旨在解决传统材料研发周期长、试错成本高的行业痛点。该平台将材料发现流程整合为五个核心环节，其中包括定义目标与规格、生成新型候选材料、模拟高保真属性、合成与实验室验证，以及规模化至生产就绪成果。经过这一系列闭环流程，传统耗时数年的材料研发周期预计可压缩至六个月以内。

技术方面，该平台的底层技术架构由两大核心模型驱动，主要包含 Meta Platforms 去年发布的 UMA 系列模型，以及公司自研的 kUPS 工具包。其中，UMA 系列 AI 模型用于在原子层面精确模拟化合物的电子特性，从而帮助研究者理解材料间的相互作用，kUPS 工具包则专注于分子层面的模拟分析，尤其是针对晶体等物质的内部结构，同时该工具具备自动化部分编码任务的能力，并可深度优化 GPU 处理效率，简化模拟过程中的错误排查。此外，公司已获得多个顶尖学术数据集在 AI 训练方面的独家授权，并结合主流科学期刊的公开数据，构建了高壁垒的训练数据池，确保模型的预测精度与泛化能力。

在企业发展布局上，伴随本轮融资，公司同步推出 "AI Materials Foundry" (AI 材料铸造厂) 全球网络计划，旨在整合全球分散的算力、数据、实验室与专业知识资源，打造材料领域的 AI 基础设施。当前，该企划已吸引超 45 家行业领军企业加入，创始合作伙伴涵盖多个层面：核心技术层面，NVIDIA 提供加速计算支持，Meta 则供应开源材料科学大模型；产业生态层面，Applied Materials、Lam Research、imec、KIOXIA、Tokyo Electron 等半导体行业巨头，以及 Hyundai Motor Group、Fujifilm、3M、Mitsui Chemicals、Kemira 等产业公司集体加盟，共同打通商业化闭环；同时，ICSD (无机晶体结构数据库)、CCDC (剑桥晶体数据中心) 及学术出版商 Wiley 等数据与学术伙伴的加入，则进一步确保了数据来源的广度与权威性。

图表 12：上周 AI 初创公司融资动态

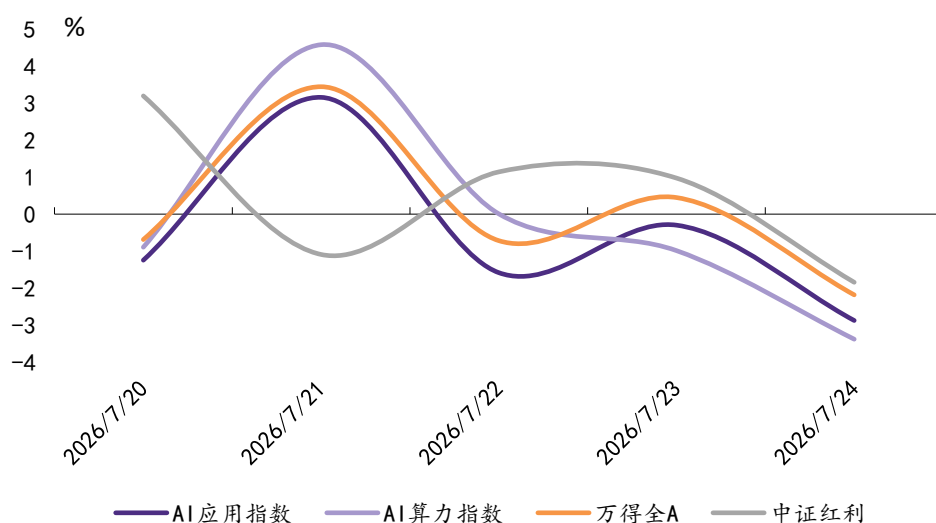
应用	应用类型	领投方	融资轮	融资额	目前累计 融资额	目前估值
CuspAI	AI 材料发现	Kleiner Perkins、 NEA	B 轮	4.5 亿美元	7 亿 美元	26 亿美元
Meshy	AI 3D	未披露	B 轮	近 4 亿美元	近 4.5 亿 美元	15 亿美元
Etched	AI 芯片	Sequoia Capital	C 轮	3 亿美元	9.2 亿美 元	103 亿美 元

资料来源：wind，Saasverse，华鑫证券研究

4、行情复盘

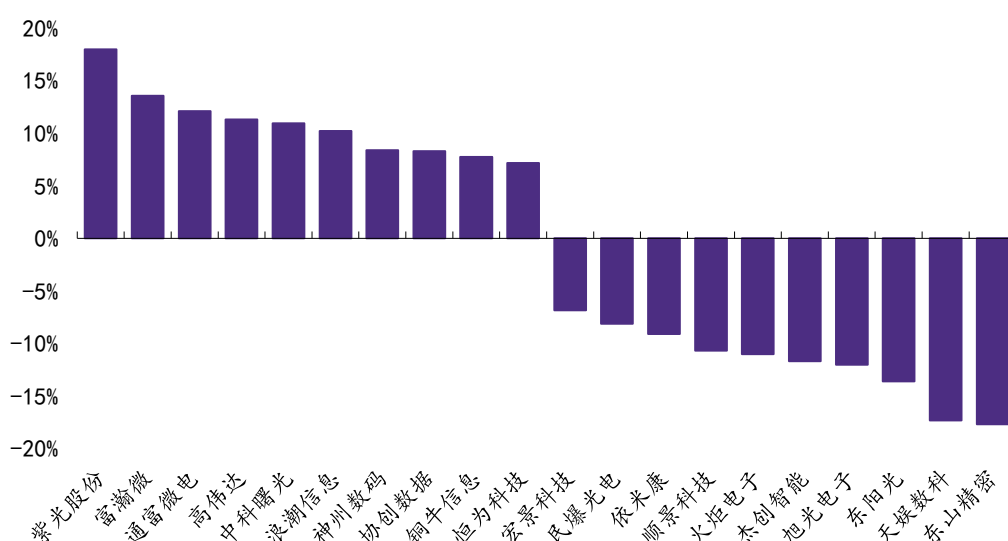
上周（2026. 7. 20-2026. 7. 24 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 3.16%/4.58%/3.44%/3.19%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-2.87%/-3.38%/-2.18%/-1.84%。AI 算力指数内部，紫光股份以 17.99% 录得上周最大涨幅，东山精密以-17.67%录得上周最大跌幅。AI 应用指数内部，创业黑马以 19.76%录得上周最大涨幅，格灵深瞳以-22.02%录得上周最大跌幅。

图表 13：上周（2026. 7. 20-2026. 7. 24 日）指数日涨跌幅



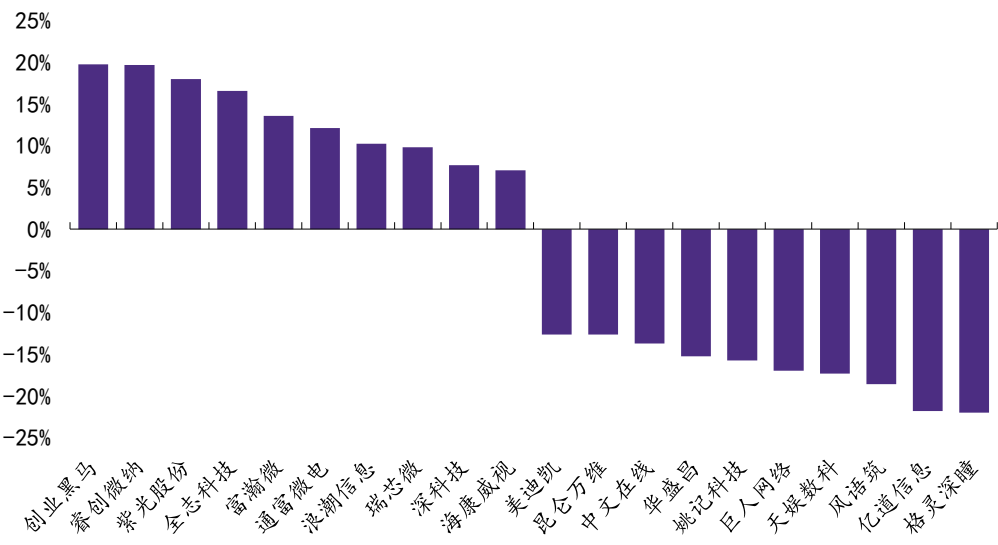
资料来源：wind, 华鑫证券研究

图表 14：上周（2026. 7. 20-2026. 7. 24 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 15：上周（2026. 7. 20-2026. 7. 24 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 22 日，谷歌发布 2026Q2 财报：当季实现总营收 1197.96 亿美元，同比增长 24%；净利润 1121.07 亿美元，摊薄后每股收益 9.11 美元。分业务看，谷歌云增长最为强劲，当季营收 247.68 亿美元，同比大幅增长 82%，营业利润率由 20.7%提升至 35.6%，在手积压订单由一季度的 4600 亿美元增至 5140 亿美元；公司在财报电话会上表示，旗下 GeminiEnterprise 已渗透近 90%的《财富》100 强企业。谷歌服务营收 945.40 亿美元（同比+15%），其中搜索业务营收 632.71 亿美元（同比+17%），YouTube 广告营收 110.55 亿美元（同比+13%）。资本开支方面，公司单季资本开支 449 亿美元，同比翻倍；管理层在电话会上表示，其中约 60%投向服务器采购、40%投向数据中心与网络设备，并将全年资本开支指引由 1800 - 1900 亿美元上调至 1950 - 2050 亿美元。模型层面，公司表示正全力推进新一代旗舰模型 Gemini4 的预训练并保持按月迭代节奏；据 CEO 皮查伊披露，API 每分钟 Token 处理量已由上季度的 160 亿提升至 220 亿，公司并于当季首次向客户数据中心交付 TPU 系统。

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，谷歌云 82%的同比增速、5140 亿美元积压订单及 Token 消耗的爆发式增长表明企业级 AI 需求仍在加速，且供给持续紧张——管理层表示 Q3 起将租用第三方算力作为过渡，进一步印证算力缺口为结构性而非周期性。从传导路径看，全年 2000 亿美元级资本开支将直接转化为上游采购需求：约 60%投向服务器、40%投向数据中心与网络设备，叠加 TPU 首次对外交付客户数据中心并开辟新收入模式。我们认为，本轮资本开支上行由真实需求驱动：订单积压与对外租算力并存，表明算力供给仍滞后于需求增长，资本开支具备扎实的基本面支撑与持续性；在此背景下，AI 服务器、光模块与高速互联、数据中心网络设备、IDC 及电力配套等上游硬件环节将按资本开支投向结构直接承接订单，需求传导的确定性较强。

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 16: 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 17：重点关注公司及盈利预测

公司代码	名称	2026-07-28 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	427.12	-0.30	0.30	0.60	-1423.73	1423.73	711.87	买入
301196.SZ	唯科科技	77.30	2.53	3.34	3.98	30.55	23.14	19.42	买入
301369.SZ	联动科技	140.20	0.48	1.08	2.62	292.08	129.81	53.51	买入
603859.SH	能科科技	36.58	0.92	1.05	1.32	39.76	34.84	27.71	买入
688615.SH	合合信息	93.03	3.24	4.22	5.25	28.71	22.05	17.72	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。

■ 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。