

2026 年 07 月 15 日



华鑫证券
CHINA FORTUNE SECURITIES

OpenAI 正式发布 GPT-5.6 系列大模型， SpaceXAI 发布 Grok 4.5

—计算机行业周报

推荐(维持)

投资要点

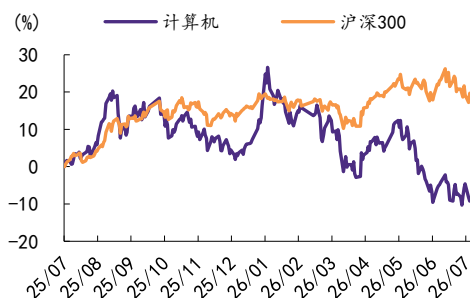
分析师：任春阳 S1050521110006

rency@cfsc.com.cn

行业相对表现

表现	1M	3M	12M
计算机(申万)	-3.7	-13.0	-9.3
沪深 300	-1.9	2.4	19.3

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：ClaudeSonnet5 发布，DSpark 与 JetSpec 共探投机解码效率新边界》2026-07-06
- 2、《计算机行业周报：火山引擎发布豆包 2.1 系列大模型，ClaudeTag 开启人机协作新范式》2026-07-02
- 3、《计算机行业点评报告：新思科技(SNPS)：Q2 收入高于指引，AI 芯片复杂度支撑 EDA 与仿真需求》2026-06-30

算力：算力租赁价格平稳，OpenAI 正式发布 GPT-5.6 系列大模型

2026 年 7 月 10 日，OpenAI 正式面向所有用户发布 GPT-5.6 系列模型。该系列采用天体命名体系，推出旗舰模型 Sol、均衡型模型 Terra 及轻量经济型模型 Luna，分别面向高要求智能体任务、高效日常工作及大批量任务场景。同时，该系列新增 max 与 ultra 两档推理模式，其中 max 模式在原有 xhigh 基础上进一步延长推理时间，ultra 模式则通过并行多 Agent 协同工作。此外，官方特别指出，GPT-5.6 具备加速 AI 研究的能力，内测期间研究员日均输出 Token 量达 GPT-5.5 最高水平两倍以上。

AI 应用：NotionAI 周时长环比+9.54%，SpaceXAI 发布 Grok 4.5

2026 年 7 月 9 日，SpaceXAI 正式发布了 Grok 4.5。该模型在数万块英伟达 GB300 上完成了训练，专为编码任务和自主智能体应用而生。其在工程方面的整体表现与 GPT-5.5 不相上下，并紧贴 Opus 4.8 的水平。在 AAAI 测试中，其取得了第四名的排位，并在 Harvey 智能体基准测试中拔得头筹。

AI 融资动向：Prime Intellect 完成 1.3 亿美元 A 轮融资，投后估值达 10 亿美元

2026 年 7 月 8 日，AI 基础设施公司 Prime Intellect 宣布完成 1.3 亿美元 A 轮融资，投后估值达 10 亿美元，正式跻身独角兽行列。该公司成立于 2024 年，致力于帮助企业摆脱对 OpenAI、Anthropic 等外部 AI 实验室的依赖，通过一站式全栈开发平台赋予企业自主训练和部署定制化 AI Agent 系统的能力。目前，公司年化收入运行率已达 1 亿美元，主要客户涵盖 Ramp、Zapier 等企业。本轮融资由 Radical Ventures 领投，Nvidia Ventures、Intel Capital、Dell Technologies Capital、Iconiq 等多家机构跟投，资金将用于扩大工程师团队、构建开放技术联盟及拓展全球市场。

投资建议

近期 CPO 板块产业验证持续加速。英伟达在近期与花旗的访谈中重申，Scalable-out 网络的 CPO 已通过 Spectrum-X 实现量产，并将于今年晚些时候进一步披露客户采用细节，公

司明确表示 CPO 在 Scale-out 网络中的客户接受度较高。供应链层面，大立光董事长林恩平在 6 月股东大会上宣布，计划 9 月前完成首条光纤阵列（FA）自动化试产线架设，并在 7 月 9 日法说会上证实，已有一家客户提出 FA 量产订单需求，预计本月送样，客户将以最快速度推进认证。同日，北美测试设备厂商 AeHR 披露获得核心硅光子学客户的追加量产订单，采购一套全自动化 FOX-XP® 晶圆级老化（WLBI）系统，用于支持 AI 光互连及超大规模数据中心硅光子器件的大规模量产老化测试，CPO 产业链正逐步进入实质性兑现阶段。

伴随海外云厂商 Q2 业绩陆续披露，其资本开支指引有望进一步明朗，叠加 CPO 产业验证持续落地，我们认为 CPO 作为光互连升级的核心方向，产业趋势明确，三季度产业验证节奏有望稳步推进。

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	2026-07-15 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	495.50	-0.30	0.30	0.60	-1651.67	1651.67	825.83	买入
301196.SZ	唯科科技	106.41	2.53	3.34	3.98	42.06	31.86	26.74	买入
301369.SZ	联动科技	204.99	0.48	1.08	2.62	427.06	189.81	78.24	买入
603859.SH	能科科技	41.72	0.92	1.21	1.50	45.35	34.48	27.81	买入
688615.SH	合合信息	107.21	3.24	4.22	5.25	33.09	25.41	20.42	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，OPENAI 正式发布 GPT-5.6 系列大模型4.....

 1.1、 Tokens 跟踪..... 4

 1.2、 数据跟踪：阿里云百炼推出企业级 Agentic RAG，腾讯云发布原生存储底座 Agent Bucket.... 6

 1.3、 产业动态：OpenAI 正式发布 GPT-5.6 系列大模型..... 7

2、 AI 应用动态：NOTIONAI 周时长环比+9.54%，SPACEXAI 发布 GROK 4.5..... 11

 2.1、 周流量跟踪：NotionAI 周时长环比+9.54%..... 11

 2.2、 产业动态：SpaceXAI 发布 Grok 4.5，实现响应速度和成本效益双突破..... 11

3、 AI 融资动向：PRIME INTELLECT 完成 1.3 亿美元 A 轮融资，投后估值达 10 亿美元..... 14

4、 行情复盘 16

5、 投资建议 18

6、 风险提示 19

图表目录

图表 1：TOKENS 规模 LEADERBOARD（TOKEN 消耗量为统计口径）..... 6

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）..... 6

图表 3：ARTIFICIAL ANALYSIS CODING AGENT INDEX 测试结果横向对比图..... 7

图表 4：TERMINAL-BENCH 2.1 测试结果横向对比图..... 8

图表 5：BROWSECOMP、OSWORLD 测试结果横向对比图..... 8

图表 6：GPT-5.6 SOL 不同推理模式测试结果横向对比图..... 9

图表 7：递归自我改进（RSI）测试结果横向对比图 10

图表 8：2026.7.4-2026.7.10 AI 相关网站流量..... 11

图表 9：GROK 4.5 在各项测试中的表现..... 12

图表 10：SPACEXAI 对 GROK 4.5 的训练..... 12

图表 11：GROK 4.5 的 TOKEN 利用效率..... 13

图表 12：GROK 4.5 的定价..... 13

图表 13：上周 AI 初创公司融资动态 14

图表 14：上周（2026.7.6-2026.7.10 日）指数日涨跌幅..... 16

图表 15：上周（2026.7.6-2026.7.10 日）AI 算力指数内部涨跌幅度排名 16

图表 16：上周（2026.7.6-2026.7.10 日）AI 应用指数内部涨跌幅度排名 17

图表 17：FICONTEC2025 年年中至今公告订单..... 18

图表 18：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，OpenAI 正式发布 GPT-5.6 系列大模型

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 7 月 6 日至 7 月 12 日，周度 Token 消耗量显著上涨，消耗量为 52.6T，环比上周增加 12.63%。在 Tokens 规模 Leaderboard 前五名中，Tencent 的 Hy3 以 6.13 T tokens 位居榜首；Xiaomi 的 MiMo-V2.5 以 5.95T tokens 位列第二；DeepSeek 的 DeepSeek V4 Flash 以 5.22T tokens 位列第三；MiniMax 的 MiniMax M3 以 4.26T tokens 位居第四；Z.AI 旗下的 GLM 5.2 以 3.19T tokens 位居第五。

文本类模型的市场份额，Google 以 882M requests 占据 27.0% 的份额，位居第一；DeepSeek 以 661M requests 占据 20.2%，位列第二；OpenAI、Qwen、Anthropic 则分别以 542M、169M、166M requests，对应占据 16.6%、5.2%、5.1% 的市场份额。

图像生成类模型调用量小幅回落，调用量为 5.33M requests，环比上周减少 2.56%。在 Image Leaderboard 中，Google 的 Nano Banana 系列包揽榜单前五名，其中 Nano Banana (Gemini 2.5 Flash Image) 以 1.68M requests 位居第一；Nano Banana 2 (Gemini 3.1 Flash Image Preview)、Nano Banana 2 (Gemini 3.1 Flash Image) 分别以 837K、655K requests 位列第二、第三；Nano Banana Pro (Gemini 3 Pro Image Preview)、Nano Banana Pro (Gemini 3 Pro Image) 分别以 556K、361K requests 位列第四、第五。从市场份额来看，Google 以 4.32M requests 占据 81.0% 的份额，位居第一；OpenAI 以 399K requests 占据 7.5%，位列第二；xAI、Black Forest Labs、ByteDance Seed 则分别以 256K、163K、131K requests，对应占据 4.8%、3.1%、2.5% 的市场份额。

嵌入模型调用量有所上升，调用量为 223M requests，环比上周增加 5.19%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 77.9M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 65.2M requests 位列第二；Google 的 Gemini Embedding 001 以 13.2M requests 位列第三；OpenAI 的 Text Embedding 3 Large 以 11.9M requests 位列第四；BAAI 的 bge-m3 以 10.8M requests 位列第五。从市场份额来看，OpenAI 以 92.4M requests 占据 41.4% 的份额，位居第一；Qwen 以 68.4M requests 占据 30.7%，位列第二；Google、BAAI、Perplexity 则分别以 24M、12.5M、11.9M requests，对应占据 10.8%、5.6%、5.3% 的市场份额。

重排序模型调用量显著上升，调用量为 2.48M requests，环比上周增加 72.22%。在 Rerank Leaderboard 中，NVIDIA 的 Llama Nemotron Rerank VL 1B V2 以 1.38M requests 位列榜首；Cohere 的 Rerank v3.5、Rerank 4 Pro、Rerank 4 Fast 以 598K、286K、217K requests 位列第二、第三和第四。从市场份额来看，NVIDIA 以 1.38M requests 占据 55.6% 的份额，位居第一；Cohere 以 1.1M requests 占据 44.4% 的份额，位列第二。

视频生成类模型调用量保持稳定，调用量为 129K requests，环比上周无增减。在 Video Leaderboard 前五名中，Google 的 Veo 3.1 Fast 以 42K requests 位列第一；ByteDance 的 Seedance 2.0 以 21K requests 位居第二；Google 的 Veo 3.1、Veo 3.1 Lite 以 18K requests 并列第三；xAI 的 Grok Imagine Video 以 7.78K requests 位居第五。从市场份额来看，Google 以 78K requests 占据 60.6% 的份额，位居第一；ByteDance 以 32K requests 占据 25.0%，位列第二；xAI、KwaiVGI、Alibaba 则分别以 8K、6K、4K requests，对应占据 6.0%、

4.7%、3.3%的市场份额。

语音生成类模型调用量显著增长，调用量为 1.29M requests，环比上周增加 40.22%。在 Speech Leaderboard 前五名中，Google 的 Gemini 3.1 Flash TTS Preview 以 845K requests 位居第一；Hexgrad 的 Kokoro 82M 以 330K requests 位列第二；xAI 的 Grok Voice TTS 1.0 以 79K requests 位列第三；Microsoft 的 MAI-Voice-2 以 29K requests 位列第四；Mistral AI 的 Voxtral Mini TTS 以 4.25K requests 位列第五。从市场份额来看，Google 以 845K requests 占据 65.5% 的份额，位居第一；Hexgrad 以 330K requests 占据 25.6%，位列第二；xAI、Microsoft、Mistral AI 则分别以 79K、29K、4K requests，对应占据 6.1%、2.2%、0.3% 的市场份额。

语音转写类模型调用量大幅回落，调用量为 3.51M requests，环比上周减少 49.5%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3 Turbo 以 1.02M requests 位列第一；OpenAI 的 Whisper Large V3 以 960K requests 位列第二；OpenAI 的 GPT-4o Mini Transcribe 以 493K requests 位列第三；NVIDIA 的 Parakeet TDT 0.6B v3 以 350K requests 位列第四；Qwen 的 Qwen3 ASR Flash 以 257K requests 位列第五。从市场份额来看，OpenAI 以 2.61M requests 占据 74.5% 的份额，位居第一；NVIDIA 以 350K requests 占据 10.0%，位列第二；Qwen、Mistral AI、Microsoft 则分别以 257K、176K、58K requests，对应占据 7.3%、5.0%、1.7% 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。

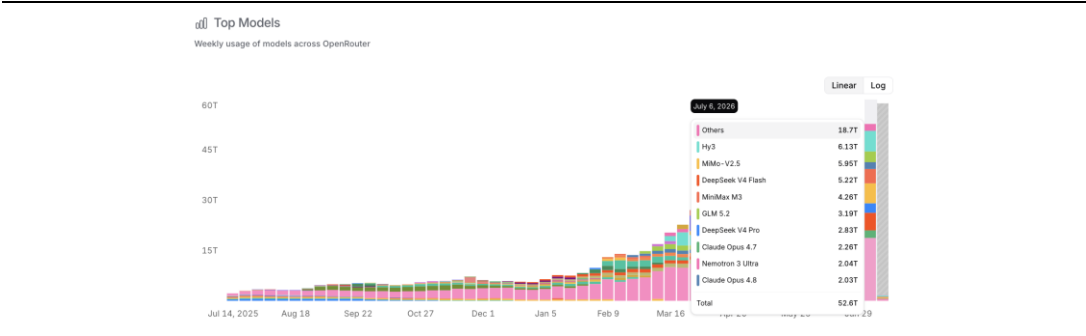
7 月 6 日，腾讯正式发布混元大模型 Hy3。该模型采用 MoE 架构，最大支持上下文长度达 256K Token。价格方面，该模型延续了高性价比的普惠定位：输入 1 元/百万 Tokens、输出 4 元/百万 Tokens，输入命中缓存 0.25 元/百万 Tokens。同时，该模型已接入 WorkBuddy、CodeBuddy、元宝等多款产品，并将陆续于 OpenRouter、Hermes 等海外平台上线。

近期，华为技术有限公司副总裁、华为数据通信产品线总裁王雷于行业分享大会上指出，当前国内 Token 的实际流通效率仅为 60%，其余 40% 的价值在传输过程中因丢包、时延抖动和数据损耗而被无谓消耗，产业整体效率仍有极大的优化空间。

7 月 8 日，中国工程院院士邬贺铨于第二十五届中国互联网大会上指出，大模型虽然开启了 AI 时代，但智能体才是补齐精准度短板的关键所在。同时，邬贺铨院士预测，到 2030 年，国内对话类 Token 流量占总量比例将从 2025 年的约 50% 下降至 12%，而智能体及智能体相关服务的流量占比将攀升至 75%。

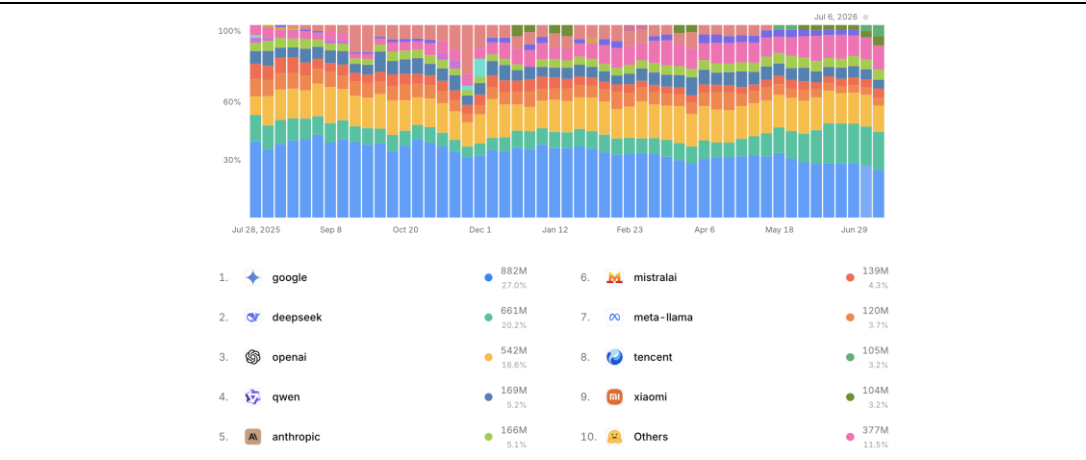
近期，中国电信、中国移动、中国联通三大运营商负责 Token 经营分别发表主题演讲，三家运营商在研判中一致认为，人工智能正引发全方位、革命性的产业变革，正逐步成为智能时代重构价值分配的核心计量单位，并深刻重塑着传统产业的分工格局与经营逻辑。

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：阿里云百炼推出企业级 Agentic RAG，腾讯云发布原生存储底座 Agent Bucket

2026 年 7 月 7 日，阿里云百炼知识库推出企业级 Agentic RAG 服务 Knowledge Studio。该服务提供多模态搜索回答、Agentic Search 及多库混合检索问答等能力，支持最多 15 个知识库联合检索。

核心能力方面，该服务通过 Agentic Search 实现智能化检索，突破了传统 RAG 方案在复杂问题推理方面准确率较低的局限，能够循环完成意图理解、任务拆解、知识检索与结果验证的全流程。同时，Agentic Search 内置六种检索工具，包括语义搜索、章节浏览、章节精读、页面浏览、SQL 执行及文件获取，并具备 15 个知识库联合检索的能力。用户可根据业务需求为每个知识库设置权重，系统会根据检索结果自动更换策略、拆分子问题、切换检索工具进行重试，直至获得可靠结果。

在回答生成阶段，该服务可通过 Agentic Generation 反复验证检索结果，思考与检索过程全程可见。此外，该服务支持引用图片、表格、文档等多模态内容，所有信息均标注来源文档、段落及页码，用户可一键查看原文，确保答案可追溯、可核查。

当前，百炼RAG已同吉利汽车、友邦人寿进行深度合作，其核心能力已支持MCP、Skill、

CLI 调用，Qoder、Codex、Dify 等 Agent 工具均可直接接入。

与此同时，腾讯云于 7 月 9 日正式发布 Agent Bucket，该产品面向 AI Agent 场景打造原生存储底座，旨在为亿级 Agent 提供独立的云端空间，统一存放用户上传的资料以及 Agent 生成的报告、图片、代码等各类工作产物。

针对传统存储方案在 Agent 场景中的短板，Agent Bucket 在 Bucket 之下引入 Space 作为独立单元，为每个 Agent 或用户配备专属访问凭证、容量配额和读写限速，使资源隔离从依赖代码逻辑升级为系统原生能力，从而支持精细到单个 Agent 的资源管理，并在高并发场景下自动实现负载均衡，避免单点热点瓶颈。

在体验层面，Agent Bucket 将目录管理、搜索、分享、回收站等通用能力前置到存储层，支持基于 SHA-256 的秒传去重、POSIX 文件系统语义以及多模态检索能力，开发者无需额外构建文件管理系统即可实现完整功能。

针对 Agent 高频读写场景，Agent Bucket 引入 GooseFS 数据加速能力，通过分布式缓存减少远端访问，提高响应速度，同时采用冷热分层策略在性能与成本之间取得平衡。

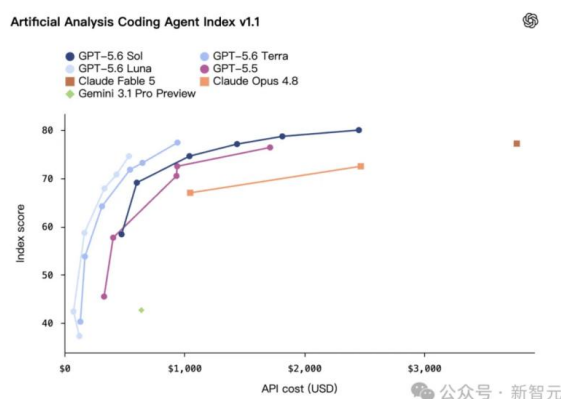
目前，该服务已搭载于 WorkBuddy、QClaw、ClawPro 等腾讯 Agent 应用中，并服务于华硕和东风日产等客户。

1.3、产业动态：OpenAI 正式发布 GPT-5.6 系列大模型

2026 年 7 月 10，OpenAI 正式面向所有用户发布 GPT-5.6 系列模型。该系列采用天体命名体系，共包含三个版本，其中，旗舰模型 Sol（太阳）专攻硬核编程、知识工作、网络安全和科学领域，适用于高要求的智能体任务，均衡型模型 Terra（大地）主打高效日常工作，轻量经济型模型 Luna（月亮）则以高性价比满足大批量任务需求。

在编程能力方面，官方将 Sol 称为迄今最强编程模型。数据显示，在 Artificial Analysis Coding Agent Index 测试中，Sol 在 max 推理模式下取得 80 分，刷新 SOTA 纪录，较 Claude Fable 5 高出 2.8 分，同时输出 Token 不到 Fable 5 半数、耗时缩减逾半、预估成本降低约三分之一。与此同时，该系列其他模型，Terra 表现优于 Fable 5，Luna 则显著领先于 Opus 4.8，两者均以远低于竞争对手的算力与成本投入，实现了同等乃至更优的编程任务表现。

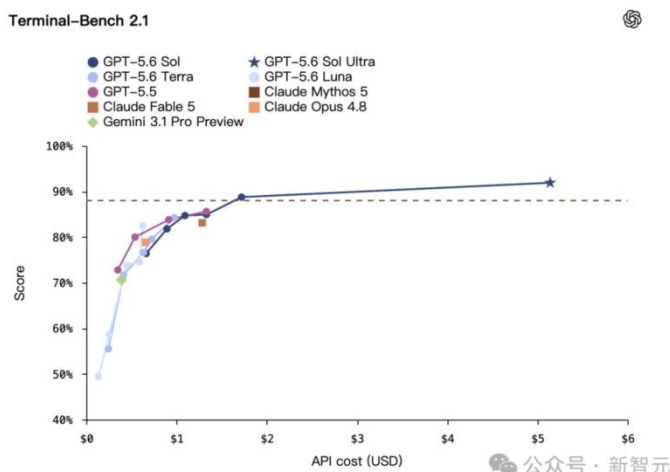
图表 3：Artificial Analysis Coding Agent Index 测试结果横向对比图



资料来源：新智元，华鑫证券研究

在考验命令行长程工程的 Terminal-Bench 2.1 评测中，Sol 同样刷新了行业最佳成绩。具体来看，Sol 标准模式取得 88.8% 的得分，开启 Ultra 模式后进一步提升至 91.9%，而 Claude Fable 5 仅为 83.1%。

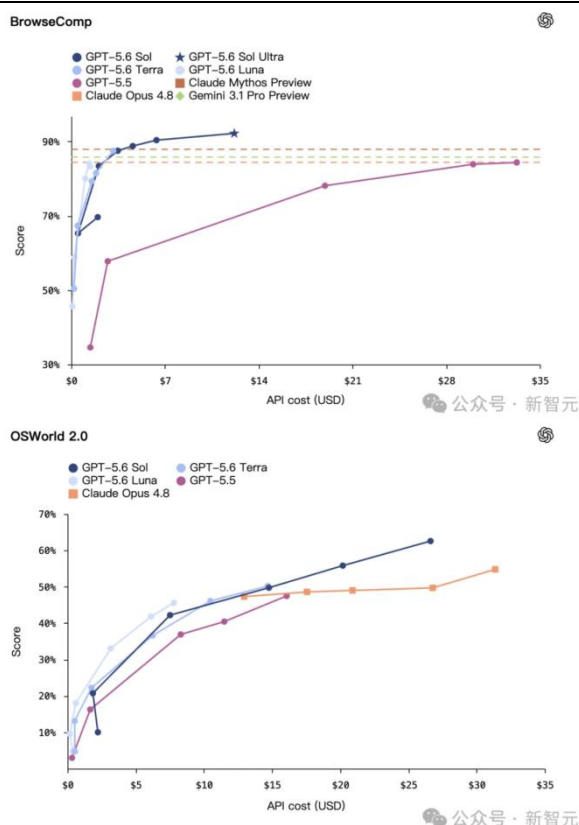
图表 4：Terminal-Bench 2.1 测试结果横向对比图



资料来源：新智元，华鑫证券研究

在端到端知识工作领域，Sol 同样表现亮眼。该模型在 BrowseComp、OSWorld 测评中分别取得 92.2%、62.6% 的得分，两项成绩均刷新 SOTA。其中，在 OSWorld 2.0 测试中，Sol 较 Opus 4.8 节省了 85% 的输出 Token。

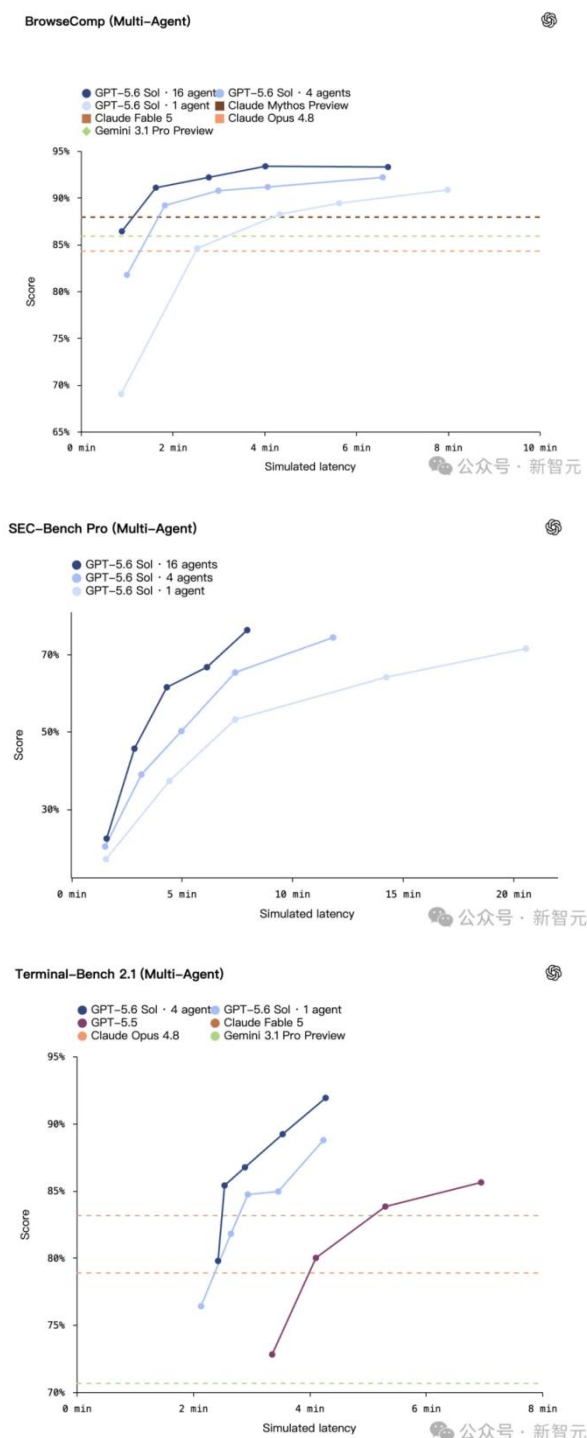
图表 5：BrowseComp、OSWorld 测试结果横向对比图



资料来源：新智元，华鑫证券研究

此次，GPT-5.6 新增了两档推理设置：max 档在原有的 xhigh 模式之上进一步延长模型的思考时间；ultra 档则默认同时启动四个 Agent 协同处理任务，面对复杂问题时最多可扩展至十六个，通过投入更多算力换取更优结果与更快完成速度。在 BrowseComp、SEC-Bench Pro、Terminal-Bench 2.1 三项评测中，ultra 模式表现尤为突出——启用并行 Agent 后，“分数—耗时”曲线整体向左上方移动，得分更高的同时速度更快。

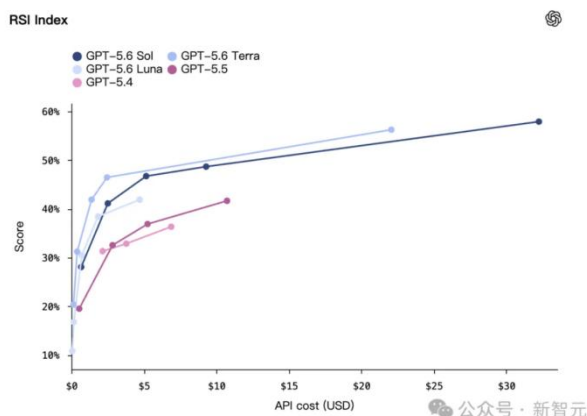
图表 6：GPT-5.6 Sol 不同推理模式测试结果横向对比图



资料来源：新智元，华鑫证券研究

除此之外，官方特别指出，GPT-5.6 具备加速 AI 研究的能力，研究员可借助它完成故障诊断、训练系统优化、实验执行与结果解读等全流程工作。内测数据显示，每位活跃研究员的日均输出 Token 量已达到 GPT-5.5 最高水平的两倍以上。过去半年间，OpenAI 内部编程推理研究算力份额增长了 100 倍，内部 Agent 的 Token 用量则增长了约 22 倍。在针对递归自我改进能力的内部评测中，Sol 的得分较 GPT-5.5 高出 16.2 分。

图表 7：递归自我改进（RSI）测试结果横向对比图



资料来源：新智元，华鑫证券研究

在定价策略上，OpenAI 采用了分层定价方案，按每百万 Token 计费。旗舰模型 Sol 的输入价格为 5 美元、输出价格为 30 美元；均衡型模型 Terra 为输入 2.5 美元、输出 15 美元；轻量型模型 Luna 为输入 1 美元、输出 6 美元。横向对比来看，Claude Fable 5 的输入价格为 10 美元、输出价格为 50 美元，整体定价水平显著高于 GPT-5.6 全系产品。

2、 AI 应用动态：NotionAI 周时长环比+9.54%，SpaceXAI 发布 Grok 4.5

2.1、 周流量跟踪：NotionAI 周时长环比+9.54%

本期（2026.7.4-2026.7.10）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1202.0M）、Bing（834.0M）和 Gemini（609.9M），访问量环比增速第一为 Github（2.30%）；平均停留时长前三位分别为 Character.AI（00:16:05）、Discord（00:11:13）和 Kimi（00:07:55）；平均停留时长环比增速第一为 NotionAI（9.54%）。

图表 8：2026.7.4-2026.7.10 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	1202.0	-1.96%	6:22	1.87%
Bing	搜索	微软	834.0	0.07%	7:14	-0.23%
Gemini	聊天机器人	谷歌	609.9	-2.48%	7:05	-0.23%
Canva	在线设计	Canva	183.4	-0.65%	5:58	0.85%
Discord	游戏社区	微软	149.4	-0.13%	11:13	0.00%
Github	代码托管	微软	142.2	2.30%	6:20	-0.26%
Character.AI	聊天机器人	Character.AI	38.43	-0.83%	16:05	2.66%
DeepL	翻译工具	DeepL	23.84	-2.49%	2:28	0.68%
Perplexity	AI 搜索	Perplexity	23.37	-3.07%	4:23	-0.38%
Kimi	聊天机器人	Moonshot AI	8.47	-3.77%	7:55	-1.04%
QuillBot	释义工具	QuillBot	7.00	-5.00%	2:53	0.58%
NotionAI	文本/笔记	Notion	2.60	-27.48%	6:42	9.54%
文心一言	聊天机器人	百度	0.29076	-23.56%	2:40	-6.43%

资料来源：similarweb, 华鑫证券研究

2.2、 产业动态：SpaceXAI 发布 Grok 4.5，实现响应速度和成本效益双突破

2026 年 7 月 9 日，SpaceXAI 正式发布了 Grok 4.5。该模型在数万块英伟达 GB300 上完成了训练，专为编码任务和自主智能体应用而生。其在工程方面的整体表现与 GPT-5.5 不相上下，并紧贴 Opus 4.8 的水平。在 AAAI 测试中，其取得了第四名的排位，并在 Harvey 智能体基准测试中拔得头筹。

从公布的基准测试成绩来看，Grok 4.5 展现出了一流的工程实力。在 SWE Bench Pro 中，它取得了 64.7% 的解决率，超越了 Opus 4.7 的 64.3%。在 Terminal Bench 2.1 测试中，其得分为 83.3%，几乎与 GPT-5.5 持平。而在 DeepSWE 1.0 基准上，它同样以 62.0% 的成绩超过了 Opus 4.8。马斯克在评价该模型时坦言，其内部评估显示 Grok 4.5 的综合能力大致与 Opus 4.7 相当，但核心竞争力在于其远超对手的响应速度。

图表 9: Grok 4.5 在各项测试中的表现



资料来源：新智元，华鑫证券研究

Grok 4.5 卓越的性能表现，根植于其独特且高强度的训练范式。研发团队在数据工程层面投入了巨大精力，对海量的原始语料执行了极为严格的过滤、去重与质量筛选流程，确保模型吸收的每一份数据都具备高度的信息密度和专业性。与此同时，RL 的重心被聚焦于一个常被忽视的指标——“单 Token 智能度”。与 Cursor 的深度整合成为了整个训练飞轮中的关键转折点，模型在训练过程中消化了数以万亿计的来自 Cursor 的真实开发交互数据。这些数据不仅包含了代码本身，更记录了人类开发者与代码库、开发工具及 AI 助手之间复杂的互动模式。同时，其训练栈针对高度异步任务进行了专门优化，能够支持智能体连续数小时执行长周期、多步骤的复杂工程任务。

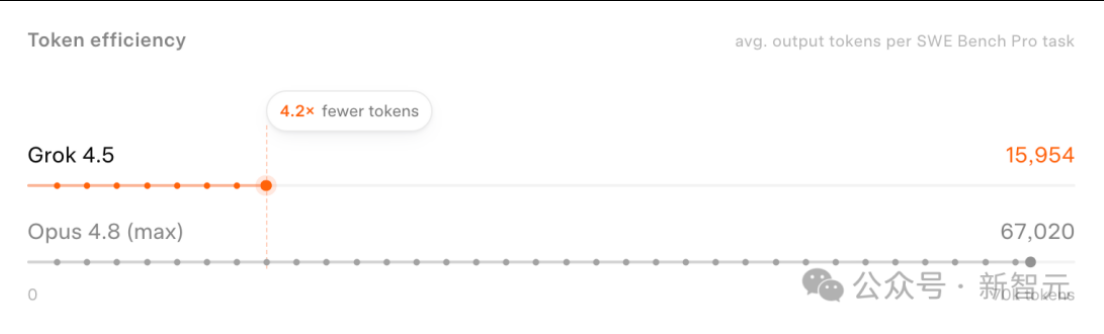
图表 10: SpaceX AI 对 Grok 4.5 的训练



资料来源：新智元，华鑫证券研究

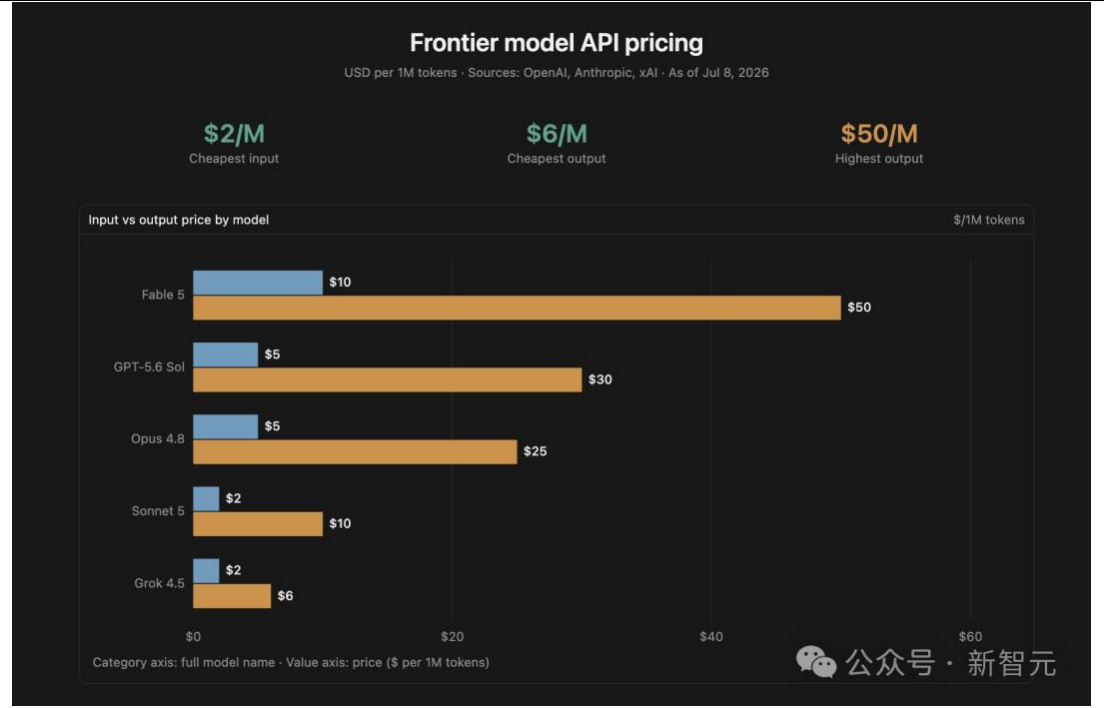
然而，Grok 4.5 真正的突破在于其高效的推理速度和 Token 利用效率。该模型的推理速度高达每秒 80 个 Token，官方宣称其超越了 Flash 类模型。在解决 SWE Bench Pro 中的同等任务时，Grok 4.5 平均仅需输出 15954 个 Token 即可完成，而 Opus 4.8 处理同一问题则平均需要消耗高达 6.7 万个 Token，前者仅为后者的四分之一不到。这种效率直接转化为成本优势，其输入 Token 定价为每百万个 2 美元，输出 Token 为每百万个 6 美元，定价低于市场上十几美元的同类顶级模型。

图表 11: Grok 4.5 的 Token 利用效率



资料来源：新智元，华鑫证券研究

图表 12: Grok 4.5 的定价



资料来源：新智元，华鑫证券研究

3、AI 融资动向：Prime Intellect 完成 1.3 亿美元 A 轮融资，投后估值达 10 亿美元

2026 年 7 月 8 日，AI 基础设施公司 Prime Intellect 宣布完成 1.3 亿美元 A 轮融资，投后估值达 10 亿美元，正式跻身独角兽行列。本轮融资由知名风投机构 Radical Ventures 领投，Nvidia Ventures、Intel Capital、Dell Technologies Capital、Iconiq 等多家机构跟投。此外，Perplexity、Box、Harvey、Cognition、Mercor 等多家知名公司创始人以天使投资人身份参与其中。

该公司成立于 2024 年，致力于帮助企业摆脱对 OpenAI、Anthropic 等外部 AI 实验室的依赖，赋予其自主训练和部署定制化 AI Agent 系统的能力。在传统模式下，企业若要将 AI 深度融入核心业务流程，通常需将自有数据托管至第三方模型平台，这不仅存在数据泄露的风险，也可能因服务商模型迭代滞后或服务突然下线等策略调整而陷入被动。Prime Intellect 正是瞄准这一痛点，通过提供全栈开发平台，让企业能够安全、自主地掌控 AI 技术栈，避免受制于外部供应商的不可控变动。

Prime Intellect 的独特之处，在于其将顶级 AI 实验室所具备的能力进行系统整合，构建为一套一站式开发平台。该平台本身为一套全栈 AI Agent 开发体系，集成了三类核心能力，其中包括计算资源接入、强化学习框架和模型评估工具，同时采用模块化设计，使企业客户能够根据实际需求灵活选用组件，无需从零搭建复杂的技术架构。

当前，Prime Intellect 年化收入运行率已达 1 亿美元，主要客户包括金融科技公 司 Ramp、自动化工作流工具提供商 Zapier，以及 Flapping Airplanes 等企业。其中，Ramp 借助该平台构建了一款可在电子表格中检索信息的 AI Agent，其准确率超越了前 沿模型，运行速度更快，成本仅为几分之一。

公开信息显示，本轮融资资金将主要用于三大方向：一是扩大工程师团队，持续 优化平台性能；二是加强与芯片厂商及云服务提供商的合作，构建开放技术联盟；三 是拓展全球市场，重点布局亚太和欧洲地区。

图表 13：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	
Prime Intellect	AI 基础设施	Radical Ventures	A 轮	1.3 亿美元	超 1.5 亿 美元	10 亿美元

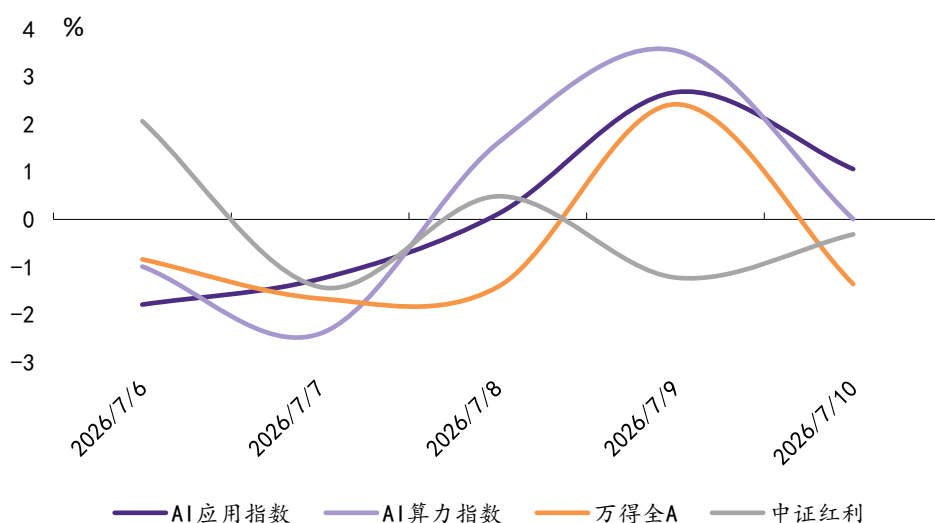
Norm	AI 法律	Khosla Ventures	C 轮	1.2 亿美元	超 2 亿美 元	12 亿美元
Bespoke Labs	AI 智能体	Wing VC	A 轮	3175 万美 元	4000 万美 元	未披露

资料来源：wind，Saasverse，华鑫证券研究

4、行情复盘

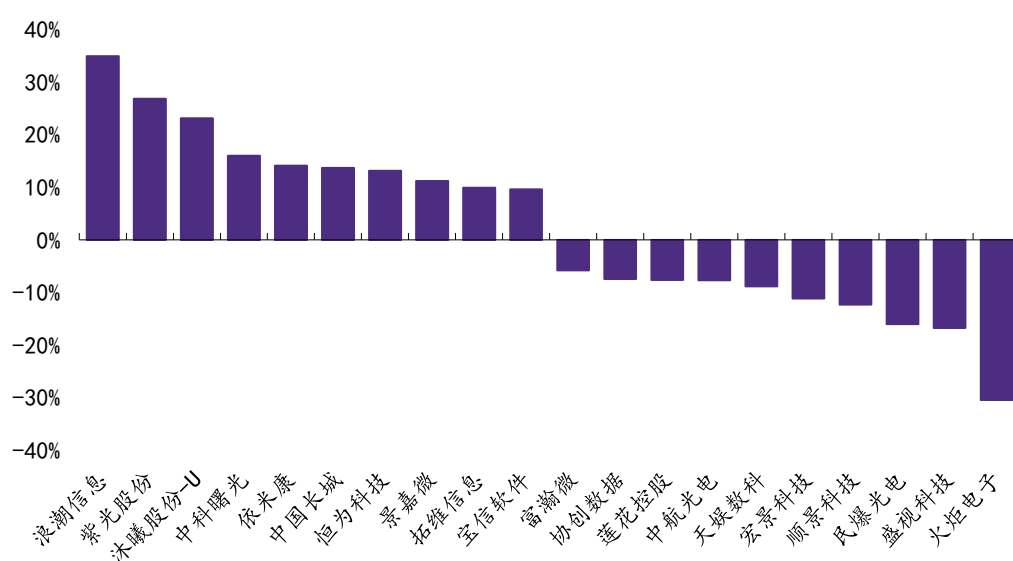
上周（2026.7.6-2026.7.10日），AI应用指数/AI算力指数/万得全A/中证红利日涨幅最大值分别为2.67%/3.55%/2.42%/2.06%，AI应用指数/AI算力指数/万得全A/中证红利日跌幅最大值分别为-1.79%/-2.4%/-1.66%/-1.42%。AI算力指数内部，浪潮信息以34.92%录得上周最大涨幅，火炬电子以-30.45%录得上周最大跌幅。AI应用指数内部，视源股份以36.8%录得上周最大涨幅，领益智造以-15.87%录得上周最大跌幅。

图表 14：上周（2026.7.6-2026.7.10日）指数日涨跌幅



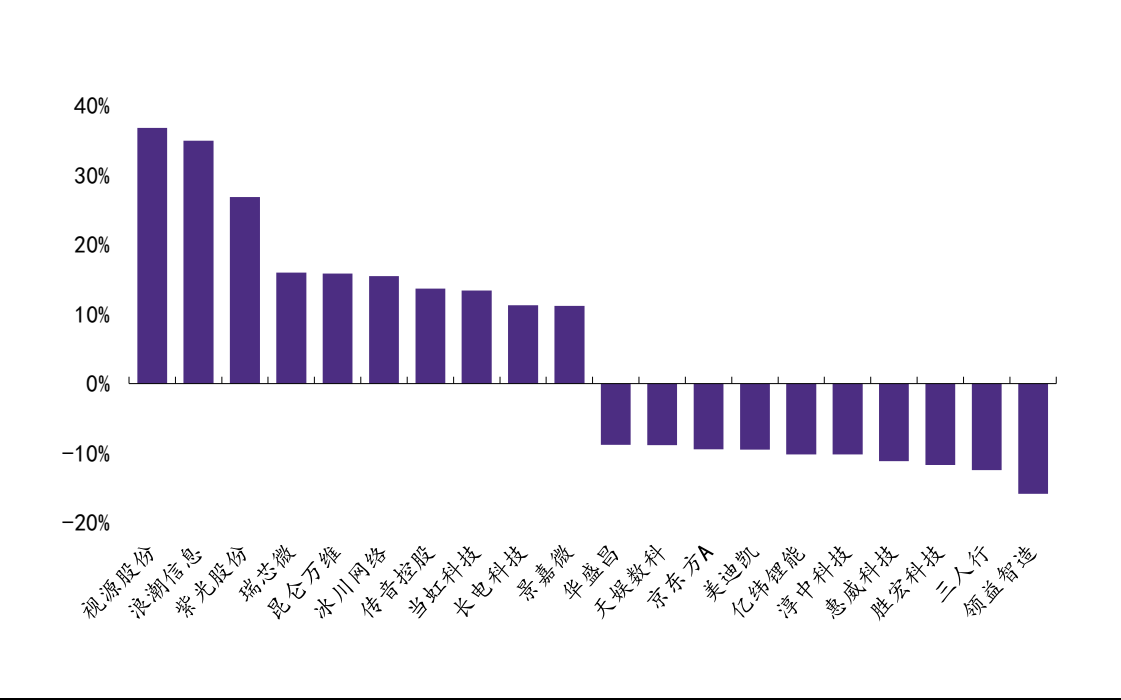
资料来源：wind, 华鑫证券研究

图表 15：上周（2026.7.6-2026.7.10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 16：上周（2026. 7. 6-2026. 7. 10 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

近期 CPO 板块产业验证持续加速。英伟达在近期与花旗的访谈中重申，用于 Scale-out 网络的 CPO 已通过 Spectrum-X 实现量产，并将于今年晚些时候进一步披露客户采用细节，公司明确表示 CPO 在 Scale-out 网络中的客户接受度较高。供应链层面，大立光董事长林恩平在 6 月股东大会上宣布，计划 9 月前完成首条光纤阵列（FA）自动化试产线架设，并在 7 月 9 日法说会上证实，已有一家客户提出 FA 量产订单需求，预计本月送样，客户将以最快速度推进认证。同日，北美测试设备厂商 Aehr 披露获得核心硅光子学客户的追加量产订单，采购一套全自动化 FOX-XP®晶圆级老化（WLBI）系统，用于支持 AI 光互连及超大规模数据中心硅光子器件的大规模量产老化测试，CPO 产业链正逐步进入实质性兑现阶段。

伴随海外云厂商 Q2 业绩陆续披露，其资本开支指引有望进一步明朗，叠加 CPO 产业验证持续落地，我们认为 CPO 作为光互连升级的核心方向，产业趋势明确，三季度产业验证节奏有望稳步推进。

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 17: 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 18：重点关注公司及盈利预测

公司代码	名称	2026-07-15 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	495.50	-0.30	0.30	0.60	-1651.67	1651.67	825.83	买入
301196.SZ	唯科科技	106.41	2.53	3.34	3.98	42.06	31.86	26.74	买入
301369.SZ	联动科技	204.99	0.48	1.08	2.62	427.06	189.81	78.24	买入
603859.SH	能科科技	41.72	0.92	1.21	1.50	45.35	34.48	27.81	买入
688615.SH	合合信息	107.21	3.24	4.22	5.25	33.09	25.41	20.42	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：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。