

2026 年 08 月 17 日



华鑫证券
CHINA FORTUNE SECURITIES

DeepSeek-V4-Pro 正式版 API 发布，智谱发布 GLM-5.3

—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)	11.2	-10.7	-11.6
沪深 300	4.7	-2.4	12.8

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：亚马逊 (AMZN.O)：AWS 重回高增长通道，AI 原生需求与自研芯片构建长期壁垒》2026-08-17
- 2、《计算机行业点评报告：苹果 (AAPL.O)：Q3 业绩超预期，AI 研发投入蓄力长期成长》2026-08-13
- 3、《计算机行业点评报告：Meta (META.O)：Q2 收入稳定增长，AI 投入加码压缩短期利润率》2026-08-12

算力：阿里云灵骏真武 M890 超节点实例正式上线，DeepSeek-V4-Pro 正式版 API 发布

2026 年 8 月 13 日，DeepSeek 发布 DeepSeek-V4-Pro 正式版 API，模型版本更新为 DeepSeek-V4-Pro-0813。该版本 Agent 能力在预览版基础上大大提升，部分性能逼近甚至超越 Claude Fable 5。同日，DeepSeek 宣布对 API 价格进行更新调整，采用峰谷定价模式。其中，高峰时段包括北京时间 9:00-12:00、14:00-18:00，其余时间均为空闲时段，价格为高峰时段的一半。新价格将于北京时间 2026 年 8 月 17 日 00:00 正式生效。

AI 应用：Perplexity 周时长环比+2.32%，GLM-5.3 正式发布

2026 年 8 月 14 日，智谱公司正式发 GLM-5.3，并宣布该模型将在两周内开源。这款模型拥有 7430 亿个参数，在多项主流基准测试的评估中，GLM-5.3 在目前已开源或即将开源的模型中均位列第一。

AI 融资动向：River AI 完成 11 亿美元种子轮及 A 轮融资

2026 年 8 月 11 日，AI 初创公司 River AI 宣布完成 11 亿美元种子轮及 A 轮融资。该轮融资由 General Catalyst、AMP PBC 联合领投，NVIDIA、AMD Ventures 提供战略投资，Y Combinator、Temasek 参投。该公司由 xAI 联合创始人 Igor Babuschkin 创立，总部位于美国加州帕洛阿尔托，2026 年 4 月 20 日在内华达州注册成立，距离公司创立到本轮融资公布仅约四个月时间。

投资建议

近日，Lumentum (LITE) 与 Coherent (COHR) 相继发布 2026 财年第四财季财报。Lumentum 当季营收 10.06 亿美元，单季首次突破 10 亿美元，同比增长 109.3%；non-GAAP 每股收益 3.23 美元（上年同期 0.88 美元）；non-GAAP 毛利率 50.4%（上年同期 37.8%）；公司指引 FY27Q1 营收 12.25 - 12.75 亿美元，non-GAAP 每股收益 4.05 - 4.35 美元。Coherent 当季营收 20.5 亿美元，同比增长 34%；non-GAAP 每股收益 1.74 美元（同比增长 74%）；non-GAAP 毛利率 40.2%；公司指引 FY27Q1

营收 22 - 24 亿美元、non-GAAP 每股收益 1.85 - 2.05 美元。电话会层面，CPO 产业化信号显著强化：Lumentum 第一大 CPO 客户量产计划不变且需求信号较上季增强，超高功率激光芯片预计 2027 年下半年放量，公司已获首个 ELS 模块订单（ASP 显著高于单激光芯片），管理层称超高功率激光“出货远落后于需求”，并已与 AXT 签订磷化铟衬底供应协议；OCS 出货环比翻倍，公司预计 FY27Q1OCS 收入将首次单季突破 1 亿美元。Coherent 管理层明确表示 CPO 需求“没有任何推迟、反而在提前”，CPO 收入预计自 2026 年 12 月季度开始贡献；公司将于 ECOC 发布集成光学平台 PhotonLink，覆盖激光到探测全链路并支持 CPO/NPO，相关产品同样自 12 月季度起贡献收入；产能端，6 英寸磷化铟产能预计本季末同比翻倍（较原计划提前一个季度）；订单端，FY2027 已基本订满，客户采购订单延伸至 2028 日历年，长期协议（LTA）更延伸至本十年末。

两家财报印证了 CPO 产业已从预期验证迈入订单落地与产能扩张阶段，光通信景气能见度显著拉长。需求端看，Lumentum 第一大 CPO 客户量产计划不变且需求信号增强，Coherent 明确表示 CPO 节奏较原计划提前，直接证伪产业链内的推迟论，叠加 ELS 模块首单落地、Coherent 给出 CPO 收入的具体起点（2026 年 12 月季度）与平台化载体（PhotonLink），2027 年下半年 CPO 放量的可见度正在强化而非证伪。技术路线层面，Lumentum 强调 NPO 机会对其而言为完全增量、是迈向 CPO 的中间一步，Coherent 表示正与客户同步推进 NPO 与 CPO，技术路线摇摆不改光环节价值量。供需层面，超高功率激光等环节供不应求，Lumentum 锁定磷化铟衬底供应、Coherent 产能提前翻倍且订单排至 2028 年并含 take-or-pay 条款，表明未来两年行业增长的约束在供给而非需求，量价支撑扎实。我们认为，两家光通信龙头的业绩与指引相互印证，光通信是当前 AI 算力硬件产业链中供需格局最清晰、能见度最长的环节之一；对国内市场而言，海外龙头的订单能见度与产能扩张为国内光通信产业提供了直接的景气映射，建议关注国内光通信产业链的投资机会。

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	2026-08-17	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	531.00	-0.30	0.30	0.60	-1770.00	1770.00	885.00	买入
301196.SZ	唯科科技	95.35	2.53	3.34	3.98	37.69	28.55	23.96	买入
301369.SZ	联动科技	146.69	0.48	1.08	2.62	305.60	135.82	55.99	买入
603859.SH	能科科技	42.30	0.92	1.05	1.32	45.98	40.29	32.05	买入
688615.SH	合合信息	120.71	3.24	4.22	5.25	37.26	28.60	22.99	买入

资料来源：Wind，华鑫证券研究

正文目录

1、 算力动态：阿里云灵骏真武 M890 超节点实例正式上线，DEEPSEEK-V4-PRO 正式版 API 发布.....	5
1.1、 Tokens 跟踪	5
1.2、 数据跟踪：阿里云灵骏真武 M890 超节点实例正式上线	7
1.3、 产业动态：DeepSeek-V4-Pro 正式版 API 发布，DeepSeek 调整 API 定价模式	8
2、 AI 应用动态：PERPLEXITY 周时长环比+2.32%，GLM-5.3 正式发布	11
2.1、 周流量跟踪：Perplexity 周时长环比+2.32%	11
2.2、 产业动态：GLM-5.3 正式发布，后训练提升编程效率与网安能力.....	11
3、 AI 融资动向：RIVER AI 完成 11 亿美元种子轮及 A 轮融资	14
4、 行情复盘	16
5、 投资建议	18
6、 风险提示	20

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：DEEPSEEK-V4-PRO-0813 基准测试结果横向对比图	8
图表 4：DEEPSEEK-V4-FLASH VS DEEPSEEK-V4-PRO	9
图表 5：DEEPSEEK-V4 API 定价调整	9
图表 6：2026.8.7-2026.8.13 AI 相关网站流量	11
图表 7：GLM-5.3 在多项测试中的表现	12
图表 8：GLM-5.3 在模型效率方面的表现	12
图表 9：GLM-5.3 在 CYBERGYM 基准测试中的得分	13
图表 10：上周 AI 初创公司融资动态	14
图表 11：上周（2026.8.10-2026.8.14 日）指数日涨跌幅	16
图表 12：上周（2026.8.10-2026.8.14 日）AI 算力指数内部涨跌幅度排名	16
图表 13：上周（2026.8.10-2026.8.14 日）AI 应用指数内部涨跌幅度排名	17
图表 14：FICONTEC2025 年年中至今公告订单	18
图表 15：重点关注公司及盈利预测	19

1、算力动态：阿里云灵骏真武 M890 超节点实例正式上线，DeepSeek-V4-Pro 正式版 API 发布

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 8 月 10 日至 8 月 16 日，周度 Token 消耗量有所上涨，消耗量为 75.3T，环比上周增加 9.13%。在 Tokens 规模 Leaderboard 前五名中，DeepSeek 的 DeepSeek V4 Flash 0731 以 11.2T tokens 位列榜首；Tencent 的 Hy3 以 9.95T tokens 位居第二；OpenAI 的 GPT-5.6 Luna 以 5.31T tokens 位居第三；DeepSeek 的 DeepSeek V4 Flash 0423 以 4.83T tokens 位居第四；Z.ai 的 GLM 5.2 以 4.33T tokens 位居第五。

文本类模型的市场份额，DeepSeek 以 1.09B requests 占据 27.1%，位居第一；Google 以 956M requests 占据 23.8% 的份额，位列第二；OpenAI、Qwen、Anthropic 则分别以 736M、216M、155M requests，对应占据 18.3%、5.4%、3.9% 的市场份额。

图像生成类模型调用量略微下降，调用量为 5.78M requests，环比上周减少 1.70%。在 Image Leaderboard 中，Google 的 Nano Banana 包揽榜单前五名席位，其中，Nano Banana (Gemini 2.5 Flash Image) 以 1.32M requests 位居第一；Nano Banana 2 (Gemini 3.1 Flash Image)、Banana 2 Lite (Gemini 3.1 Flash Lite Image)、Nano Banana 2 (Gemini 3.1 Flash Image Preview) 分别以 986K、667K、569K requests 位居第二、第三和第四；Nano Banana Pro (Gemini 3 Pro Image Preview) 以 422K requests 位列第五。从市场份额来看，Google 以 4.37M requests 占据 75.7% 的份额，位居第一；OpenAI 以 624K requests 占据 10.8%，位列第二；Black Forest Labs、xAI、ByteDance Seed 则分别以 237K、202K、187K requests，对应占据 4.1%、3.5%、3.2% 的市场份额。

嵌入模型调用量有所上涨，调用量为 284M requests，环比上周增加 5.97%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 100M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 73M requests 位列第二；Google 的 Gemini Embedding 2 以 18.9M requests 位列第三；BAAI 的 bge-m3 以 18.6M requests 位列第四；OpenAI 的 Text Embedding 3 Large 以 16M requests 位列第五。从市场份额来看，OpenAI 以 119M requests 占据 41.8% 的份额，位居第一；Qwen 以 77.1M requests 占据 27.1%，位列第二；Google、BAAI、Perplexity 则分别以 36.3M、19.4M、17.7M requests，对应占据 12.8%、6.8%、6.2% 的市场份额。

重排序模型调用量大幅上涨，调用量为 6.27M requests，环比上周增加 30.90%。在 Rerank Leaderboard 中，VoyageAI 的 rerank-2.5 以 2.63M requests 位列榜首；NVIDIA 的 Llama Nemotron Rerank VL 1B V2 以 1.58M requests 位居第二；Cohere 的 Rerank v3.5 以 818K requests 位列第三；VoyageAI 的 rerank-2.5-lite 以 575K requests 位列第四；Cohere 的 Rerank 4 Pro 以 327K requests 位列第五。从市场份额来看，VoyageAI 以 3.2M requests 占据 51.1% 的份额，位居第一；NVIDIA 以 1.58M requests 占据 25.2% 的份额，位居第二；Cohere 分别以 1.46M、35K requests 占据 23.2%、0.6% 的份额。

视频生成类模型调用量略微下降，调用量为 173K requests，环比上周减少 1.14%。在

Video Leaderboard 前五名中，Google 的 Veo 3.1 Lite、Veo 3.1 Fast 分别以 31K、28K requests 位列第一、第二；ByteDance 的 Seedance 2.0 Fast、Seedance 2.0 分别以 21K、15K requests 位居第三、四；KwaiVGI 的 Video v3.0 Standard 以 10K requests 位居第五。从市场份额来看，Google 以 68K requests 占据 39.4% 的份额，位居第一；ByteDance 以 59K requests 占据 34.2%，位列第二；KwaiVGI、xAI、MiniMax 则分别以 16K、13K、8K requests，对应占据 9.0%、7.6%、4.5% 的市场份额。

语音生成类模型调用量大幅上涨，调用量为 2.68M requests，环比上周增加 74.03%。在 Speech Leaderboard 前五名中，Google 的 Gemini 3.1 Flash TTS Preview 以 944K requests 位居第一；Fish Audio 的 S2.1 Pro Free (free) 以 727K requests 位列第二；Hexgrad 的 Kokoro 82M 以 348K requests 位列第三；xAI 的 Grok Voice TTS 1.0 以 333K requests 位列第四；MiniMax 的 Speech 2.8 HD 以 92K requests 位列第五。从市场份额来看，Google 以 944K requests 占据 35.2% 的份额，位居第一；Fish Audio 以 800K requests 占据 29.8%，位列第二；Hexgrad、xAI、MiniMax 则分别以 348K、333K、101K requests，对应占据 13.0%、12.4%、3.8% 的市场份额。

语音转写类模型调用量有所上涨，调用量为 11.9M requests，环比上周增加 12.26%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3、Whisper Large V3 Turbo、GPT-4o Mini Transcribe 分别以 6.23M、2.73M、643K requests 位列第一、第二和第三；xAI 的 Grok STT 1.0 以 484K requests 位列第四；OpenAI 的 GPT Transcribe 以 396K requests 位列第五。从市场份额来看，OpenAI 以 10.5M requests 占据 88.5% 的份额，位居第一；xAI 以 484K requests 占据 4.1%，位列第二；Qwen、Mistral AI、Microsoft 则分别以 404K、224K、89K requests，对应占据 3.4%、1.9%、0.8% 的市场份额。

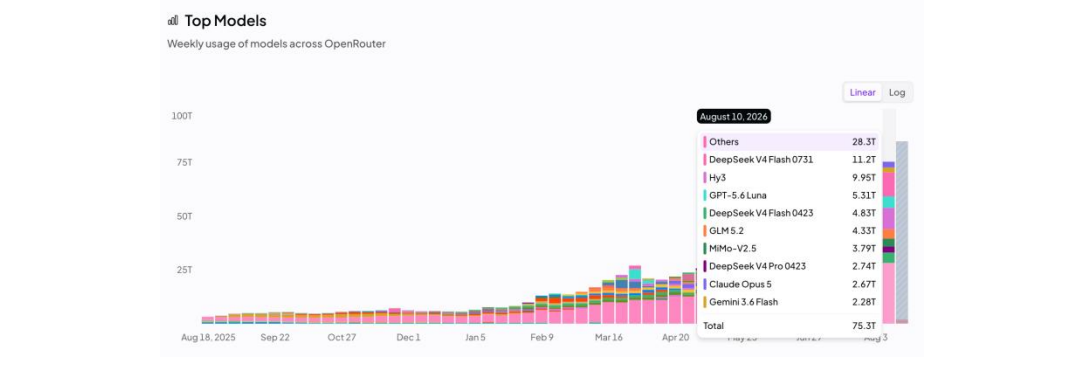
* 统计口径说明：以上数据均来源于 OpenRouter。基于官方统计口径调整，当前 Rankings 采用 Token 与 Request 两套统计维度：其中，仅 Tokens 规模 Leaderboard 仍以模型 Token 消耗量为统计口径，其余类目榜单及市场份额板块则统一采用 Request 调用量为统计口径，两者统计维度不同，不宜直接对应或进行横向比较。

8 月 11 日，蚂蚁百灵宣布开源轻量级混合推理 MoE 模型 Ling-3.0-tiny。该模型总参数量为 79 亿，推理时仅激活 13 亿参数，专为高效本地部署而设计，同步提供 BF16、FP8 和 INT4 三个版本。数据显示，FP8 精度下，Ling-3.0-tiny 在 DGX Spark 上的推理吞吐量可达 100 - 105tokens/s，在 M4 Pro MacBook 上推理吞吐量则为 86 - 90tokens/s，且在 8K 上下文长度时峰值内存占用仅为 8.34GiB，为开发者提供了兼具性能与资源效率的本地部署选择。

8 月 12 日，SpaceXAI 宣布推出新一代旗舰模型 Grok 4.6。相比上一代 Grok 4.5，新模型在保持相同定价的前提下实现了性能的显著优化。多项基准测试结果 Grok 4.6 high 版本的综合性能已与 GPT-5.6 Sol Max、Claude Fable 5 Max 等前沿模型相当。定价方面，Grok 4.6 的价格为每百万输入 Token 2 美元，每百万输出 Token 6 美元。

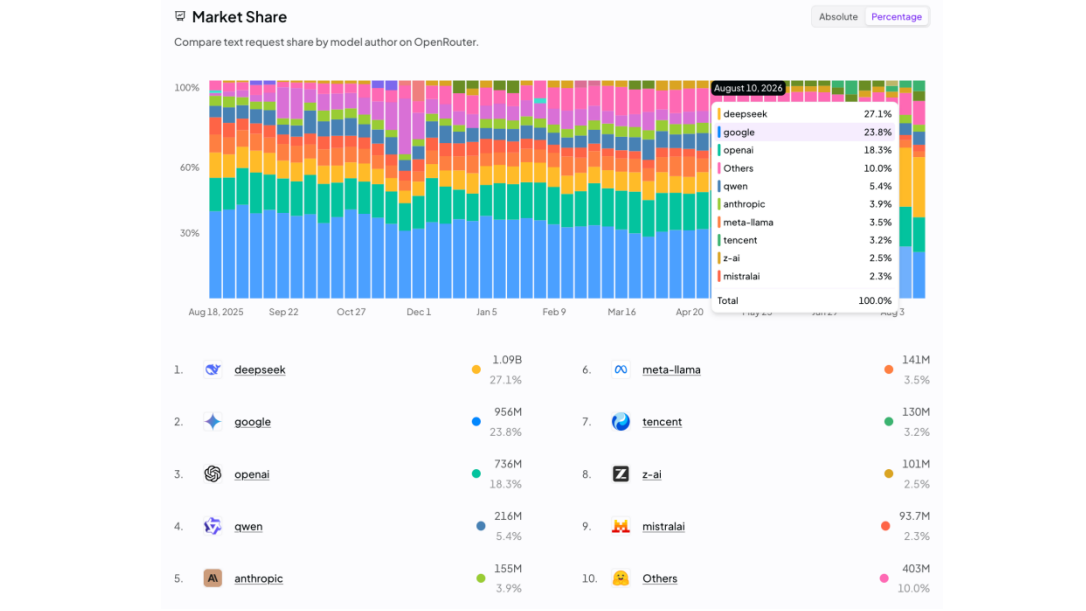
8 月 13 日，千问大模型正式开放 Qwen3.8-2.4T-A95B 模型权重，这是阿里千问首次将 Max 级别的旗舰模型对外开放。该模型基于 Qwen3.5 底层架构，总参数达 2.4 万亿，每个 Token 激活 950 亿参数，原生支持 262144 Token 上下文（可扩展至 101 万 Token），在编程、办公、科研工作及长周期智能体任务等方面均实现了性能提升。

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：阿里云灵骏真武 M890 超节点实例正式上线

2026 年 8 月 12 日，阿里云灵骏真武 M890 超节点实例正式上线。该超节点实例最高可承载十万亿参数级 MoE 大模型推理，企业级客户无需自建机房，在云上即可开通 64 卡高速互联算力单元。

真武 M890 依托于阿里云统一底座——灵骏智算平台。得益于 HPN 8.0 训推一体网络，该平台单集群最高支持 13 万卡异构算力混布，可扩展至百万卡级，从而具备对超节点集群的全面支撑能力。

该超节点实例支持 FP8/FP4 低精度计算，通过 ICN Switch 1.0 芯片将 Scale-up 互联规模由 16 卡拉升至 64 卡，卡间互联带宽达 800GB/s。相较上一代真武 810E，该超节点实例在智能驾驶、具身智能等训练场景中，训练性能可提升至 3 倍。

与此同时，真武 M890 为国内首个成功运行超 2 万亿参数大模型的超节点形态算力，当前，Kimi K3 和 Qwen3.8-Max 均已通过该实例对外提供服务。

本次超节点实例首发地域为乌兰察布，是阿里云五大超级数据中心之一。该地绿电占比约 90%，能够为高密度算力提供更加低碳的运行环境。不仅如此，该数据中心首创全模块设计架构，供电、冷却、安防、智能化以及消防的整体模块化率高达 90%，能够大幅缩短交付周期，最大化兼容主流异构芯片与计算架构，同时能够支持至少三代的芯片迭代。

此外，阿里云计划将模块化数据中心的全球产能提升两倍以上，相关超节点实例将在其他地域陆续开服。

1.3、产业动态：DeepSeek-V4-Pro 正式版 API 发布，DeepSeek 调整 API 定价模式

2026 年 8 月 13 日，DeepSeek 发布 DeepSeek-V4-Pro 正式版 API，模型版本更新为 DeepSeek-V4-Pro-0813。该版本 Agent 能力在预览版基础上大大提升，部分性能逼近甚至超越 Claude Fable 5。

基准测试结果显示，相较预览版，正式版在性能表现方面得到了全面的提升。其中，在针对长周期、高复杂度真实软件工程的编程智能体测试 DeepSWE 中，DeepSeek-V4-Pro 正式版的得分从预览版的 12.8 大幅跃升至 62.7。除此之外，该模型在 AI 安全智能体基准测试套件 Cybergym 中，获得 83.3 分的评分，超越了 Claude Fable 5 的 83.1 分，同时在工作流智能体 AutomationBench 中，以 31.8 的得分显著超越了 Fable 5 的 29.1。

图表 3：DeepSeek-V4-Pro-0813 基准测试结果横向对比图

Benchmarks	DeepSeek-V4-Pro-0813	DeepSeek-V4-Flash-0731	DeepSeek-V4-Pro-Preview	DeepSeek-V4-Flash-Preview	Opus-4.8	Fable 5 (w/ fallback)
HLE (wo/w tools)	42.7/60.0	37.8/51.5	37.7/48.2	34.8/45.1	49.8/57.9	53.3/63.0
Terminal Bench 2.1	87.9	82.7	72.1	61.8	85.0	88.0
NL2Repo	61.5	54.2	38.5	39.4	69.7	-
Cybergym	83.3	76.7	52.7	38.7	78.3	83.1
DeepSWE	62.7	54.4	12.8	7.3	58.0	70.0
Toolathlon-Verified	74.1	70.3	55.9	49.7	76.2	77.9
Agents' Last Exam	25.7	25.2	16.5	15.8	25.7	-
AutomationBench (Public)	31.8	25.1	12.8	10.8	27.2	29.1
DSBench-FullStack	71.1	68.7	41.8	37.0	71.6	77.2
DSBench-Hard	67.2	59.6	31.1	25.8	71.7	68.3

DeepSeek V4 Pro 正式版模型在 Agent 相关评测集上的表现以及与其他前沿模型的对比

deepseek

资料来源：DeepSeek，华鑫证券研究

与此同时，在功能与调用方面，DeepSeek-V4-Pro-0813 的 Agent 能力也得到了大幅增强，具备支持 JSON 结构化输出、工具调用、Responses API 以及 Anthropic API 的能力。此外，该模型的并发限制为 500，远低于 DeepSeek-V4-Flash-0731 的 2500。

图表 4: DeepSeek-V4-Flash VS DeepSeek-V4-Pro

模型		deepseek-v4-flash	deepseek-v4-pro
BASE URL (OpenAI 格式)		https://api.deepseek.com	
BASE URL (Anthropic 格式)		https://api.deepseek.com/anthropic	
模型版本		DeepSeek-V4-Flash-0731	DeepSeek-V4-Pro-0813
思考模式		支持非思考与思考模式（默认） 切换方式详见 思考模式	
上下文长度		1M	
输出长度		最大 384K	
功能	Json Output	支持	支持
	Tool Calls	支持	支持
	Responses API	支持	支持
	Anthropic API	支持	支持
	对话前缀续写 (Beta)	支持	支持
	FIM 补全 (Beta)	仅非思考模式支持	仅非思考模式支持
价格 ⁽¹⁾	百万tokens输入（缓存命中）	0.02元	0.025元
	百万tokens输入（缓存未命中）	1元	3元
	百万tokens输出	2元	6元
并发限制 ⁽²⁾		2500	500

资料来源：智东西，华鑫证券研究

同日，DeepSeek 宣布对 API 价格进行更新调整，采用峰谷定价模式。其中，高峰时段包括北京时间 9:00-12:00、14:00-18:00，其余时间均为空闲时段，价格为高峰时段的一半。新价格将于北京时间 2026 年 8 月 17 日 00:00 正式生效。

调价前，DeepSeek-V4-Pro 定价为缓存命中输入 0.025 元/百万 tokens，缓存未命中输入 3 元/百万 tokens，输出 6 元/百万 tokens；DeepSeek-V4-Flash 定价为缓存命中输入 0.02 元/百万 tokens，缓存未命中输入 1 元/百万 tokens，输出 2 元/百万 tokens。

图表 5: DeepSeek-V4 API 定价调整

DeepSeek-V4 API 新定价				
(单位：¥ / 百万Tokens)				
模型		输入（缓存命中）	输入（缓存未命中）	输出
DeepSeek-V4-Flash	空闲时段	0.05 元	1.5 元	4.5 元
	高峰时段	0.10 元	3.0 元	9.0 元
DeepSeek-V4-Pro	空闲时段	0.15 元	4.5 元	13.5 元
	高峰时段	0.30 元	9.0 元	27.0 元
高峰时段：北京时间 9:00 - 12:00，14:00 - 18:00（其余为空闲时段） 北京时间 2026 年 8 月 17 日 00:00 开始生效				
				

资料来源：DeepSeek，华鑫证券研究

调价后，DeepSeek-V4-Pro 在高峰时段每百万 Tokens 输出最高价格达 27 元。以代码审查场景为例，假设单日消耗 20 万输出 Token：调价前单日原成本约 1.2 元；调价后空闲时段调用成本约 2.7 元，高峰时段调用成本约 5.4 元。仅对比输出 Token 单价，高峰时段定价为调价前原价的 4.5 倍，空闲时段为调价前原价的 2.25 倍。

针对企业大规模 API 调用场景进行测算，若每日在高峰时段消耗 1 亿输出 Tokens：调价前单日总成本为 600 元；调价后单日成本将升至 2700 元。若企业错峰至空闲时段调用同等 Token 量，单日成本为 1350 元，仍较调价前上涨 1.25 倍。

2、AI 应用动态：Perplexity 周时长环比+2.32%，GLM-5.3 正式发布

2.1、周流量跟踪：Perplexity周时长环比+2.32%

本期（2026.8.7-2026.8.13）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1263.0M）、Bing（832.3M）和 Gemini（579.0M），访问量环比增速第一为 QuillBot（2.30%）；平均停留时长前三位分别为 Character.AI（00:16:20）、Discord（00:11:08）和 NotionAI（00:08:48）；平均停留时长环比增速第一为 Perplexity（2.32%）。

图表 6：2026.8.7-2026.8.13 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	1263.0	1.28%	6:34	-0.25%
Bing	搜索	微软	832.3	-3.72%	7:22	0.23%
Gemini	聊天机器人	谷歌	579.0	-1.55%	6:58	-0.71%
Canva	在线设计	Canva	192.2	1.48%	5:56	-0.28%
Discord	游戏社区	微软	152.6	-0.52%	11:08	-0.45%
Github	代码托管	微软	146.6	0.62%	6:25	-0.26%
Character.AI	聊天机器人	Character.AI	36.41	-0.05%	16:20	-0.51%
DeepL	翻译工具	DeepL	22.45	-1.36%	2:27	-1.34%
Perplexity	AI 搜索	Perplexity	21.86	-0.50%	4:25	2.32%
Kimi	聊天机器人	Moonshot AI	13.46	-12.31%	7:07	0.00%
QuillBot	释义工具	QuillBot	6.901	2.30%	2:51	0.00%
NotionAI	文本/笔记	Notion	2.59	-4.22%	8:48	-0.94%
文心一言	聊天机器人	百度	0.02	-33.99%	2:16	-11.69%

资料来源：similarweb, 华鑫证券研究

2.2、产业动态：GLM-5.3 正式发布，后训练提升编程效率与网安能力

2026 年 8 月 14 日，智谱公司正式发布了 GLM-5.3，并宣布该模型将在两周内开源。这款模型拥有 7430 亿个参数，在多项主流基准测试的评估中，GLM-5.3 在目前已开源或即将开源的模型中均位列第一。

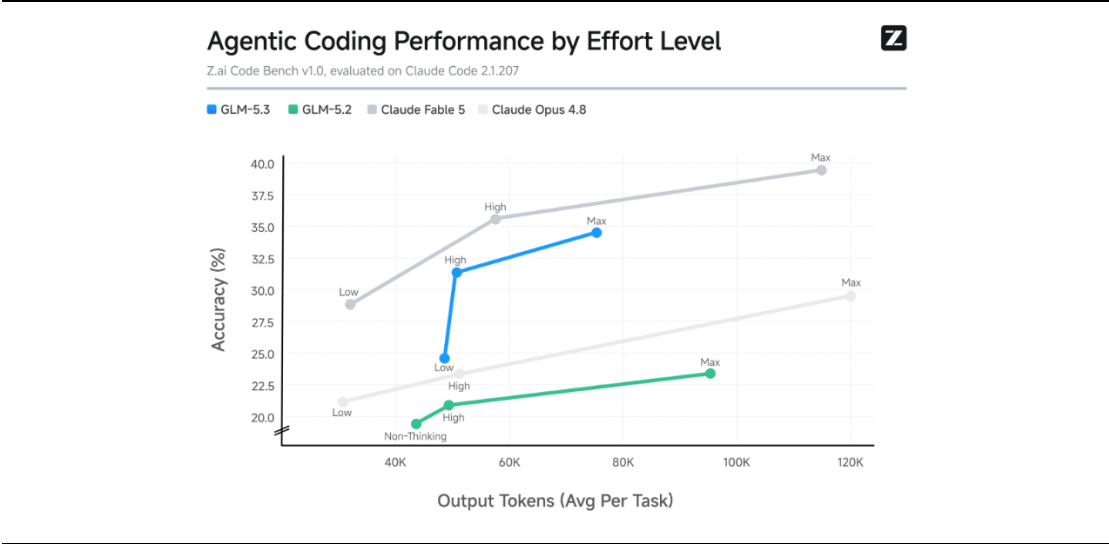
图表 7：GLM-5.3 在多项测试中的表现

Benchmark	GLM-5.3	GLM-5.2	Kimi K3	DeepSeek-V4 Pro-0813	Qwen3.8-Max	Opus 4.8	Fable 5 (w/ fallback)	GPT-5.6 Sol
CODING								
Terminal Bench 2.1	88.2	81.0	88.3	87.9	86.6	85.0	88.0	88.8
Terminal Bench 3.0	28.3	4.6	17.4	-	-	21.1	33.7	34.6
DeepSWE v1.1	66.9	46.2	67.5	62.7	56.6	58.0	69.7	72.7
NL2Repo	58.0	48.9	58.0	61.1	55.9	69.7	-	-
ProgramBench Almost Solved	19.0	9.5	17.5	-	10.5	15.5	33.0	23.0
FrontierSWE	78.1	67.5	-	-	-	66.5	88.2	-
SWE-Marathon v1.1	42.5	19.4	48.1	-	-	48.8	33.1	42.5
PostTrainBench	39.8	31.7	32.0	-	-	32.9	41.8	36.2
CYBER								
CyberGym	84.5	77.2	80.0	83.3	78.5	78.1	83.8	83.6
ExploitGym 2h / 6h	105 / 130	29 / 39	36 / 70	-	14 / 26	80 / 120	181 / 247	216 / 293
ExploitBench	54.4	24.4	32.2	-	28.8	40.0	78.0	76.5
AGENTIC								
Toolathon Verified	73.0	59.9	76.5	74.1	72.5	76.2	74.7	74.9
AutomationBench v1.0.8	48.2	26.2	46.7	43.2	39.8	41.0	46.2	45.8
Agents' Last Exam ALE-CLJ	28.5	23.8	27.6	25.7	27.0	25.7	23.8	28.6
HLE w/ Tools	62.5	54.7	59.8	60.0	56.2	57.9	63.9	64.5
GDPval-AA v2	1769	1508	1682	1590	1739	1588	1743	1730

资料来源：智东西，华鑫证券研究

在模型效率方面，GLM-5.3 在能力与 Token 使用效率之间实现了更优的平衡。在相同的高档位思考预算条件下，GLM-5.3 的准确率达到 31.5%，超越了 Claude Opus 4.8 的 29.5%，而任务平均输出 Token 数约为 5 万个，不到 Opus 4.8 平均输出量的一半。

图表 8：GLM-5.3 在模型效率方面的表现



资料来源：智东西，华鑫证券研究

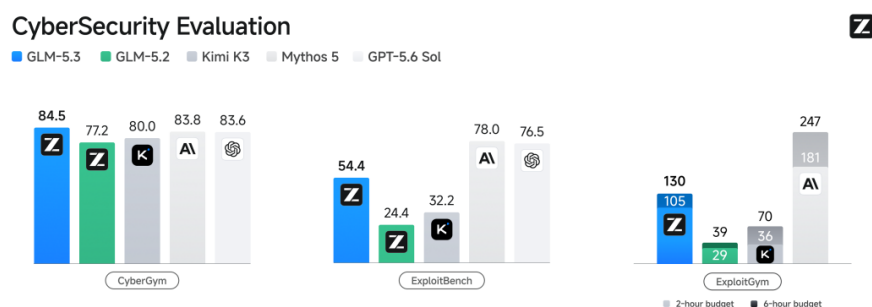
在实际开发场景的深度体验中，GLM-5.3 配合智谱的 AI 编程工具 ZCode 工具，展现了从零到一构建复杂应用系统的完整能力。测试团队以真实的 OpenStreetMap 数据为基础，要求模型开发一款 1:1 复刻上海陆家嘴至外滩区域的 3D 开放世界驾驶游戏。需求文档明确规定地图解析、坐标转换和游戏逻辑等所有关键模块必须由模型自行编写，仅允许使用 Three.js 的底层渲染能力，同时还暗中设置了关于建筑灯光定时熄灭和地图边缘软性阻挡

的两项隐蔽条款。GLM-5.3 首先通过编写脚本从 Overpass API 拉取了超过 8 万个 OSM 节点进行数据勘查，确认数据完整性后才正式开始开发，整个流程持续多轮，最终交付了约 4900 行代码，覆盖数据管线、车辆物理、音效、导航、昼夜循环和存档系统等完整功能模块，两项隐蔽条款也均被准确实现。在问题排查过程中，GLM-5.3 展现出清晰的诊断思路，当测试早期页面完全卡死时，它并未盲目修改代码，而是通过为本地服务器添加请求探针并利用面包屑日志精确定位问题根源。对于相机视角偶尔穿透建筑的模糊反馈，模型统计发现约四成建筑的轮廓环绕方向与墙面法线约定相反，从而以极小的代码改动解决了这一问题。由于测试环境无法直接目检画面，模型甚至自行编写了一个零依赖的 PNG 解码器，通过像素统计客观验证了水面、标线和夜间灯光等渲染元素的正确性。搭配 ZCode 使用时，对话的平均缓存命中率超过 99%，显著降低了 API 调用成本。不过，该模型在审美判断方面仍有提升空间，建筑贴图和道路标线的观感经过多轮人工反馈仍未达到理想效果，用户仍需提供一定程度的反馈才能打磨出更好的交付成果。

第二项实测任务聚焦于模型对陌生大型代码仓库的理解与改造能力，测试对象是 DeepSeek 刚刚开源的 DeepSeek Harness 框架。该仓库采用“一切皆插件”的架构，包含 40 多个顶层包、200 多个工作区项目和超过 7400 个文件，对于模型而言是一次毫无先验信息的“裸考”。面对这一体量的代码仓库，GLM-5.3 派出了 4 个并行探索子智能体，分别追查 CLI 入口、插件机制、模型契约与消息通路，并将结论汇总成链路报告，随后逐条核对关键代码。模型成功识别出插件装配依据 YAML 清单逐行登记而非目录扫描自动发现，锁定了所有模型提供方共同遵守的 LlmAdapter 契约，并确认启动失败采用快速失败策略。在此基础上，GLM-5.3 进一步为 DeepSeek Harness 开发了一个“人格插件”，使 AI 回复能够在文言文、东北话与猫语之间切换，且要求做到对原有框架的“零侵入”。模型派出多个探索子智能体并行通读仓库，累计消耗约 690 万 Token，最终选定系统提示注册表作为注入口，确保指令随系统层下发而不改动用户消息，插件卸载后注册表自行摘除。初版交付涉及 20 个文件、净增 560 行代码，配套的 11 条单元测试一次通过。在全量回归测试中，有 6 个文件测试失败，模型先将自身改动整体撤下并在干净基线上重跑，确认失败问题源于本机环境而非新增代码，体现了严谨的实验对照意识。解决环境问题后，模型复用 DeepSeek Harness 既有的命令机制，将人格切换接入网页端，28 项文档门禁与 937 对双语文档检查全部通过。

在技术层面，GLM-5.3 的基座模型与 GLM-5.2 保持一致，能力提升主要来自后训练阶段的 Scaling，包括更长任务环境、更丰富环境类型以及明显延长的后训练时间。随着任务环境不断丰富，模型还涌现出超预期的网络安全能力，在 CyberGym 基准测试中 GLM-5.3 得分 84.5%，超过 Claude Mythos 5 和 GPT-5.6 Sol。

图表 9: GLM-5.3 在 CyberGym 基准测试中的得分



资料来源：智东西，华鑫证券研究

3、AI 融资动向：River AI 完成 11 亿美元种子轮及 A 轮融资

2026 年 8 月 11 日，AI 初创公司 River AI 宣布完成 11 亿美元种子轮及 A 轮融资。该轮融资由 General Catalyst、AMP PBC 联合领投，NVIDIA、AMD Ventures 提供战略投资，Y Combinator、Temasek 参投。该公司由 xAI 联合创始人 Igor Babuschkin 创立，总部位于美国加州帕洛阿尔托，2026 年 4 月 20 日在内华达州注册成立，距离公司创立到本轮融资公布仅约四个月时间。

当前，该公司的核心产品为 River API，面向企业提供基于前沿开放权重模型（open-weight models）的 LoRA 微调与强化学习服务。通过该 API，企业无需配备专门的基建团队，仅需 15 至 20 分钟，即可完成复杂的强化学习训练，相较于闭源方案，成本可降低二至四倍。与此同时，产品的计费方式则按照训练与推理所消耗的 Token 量计量，企业无需为闲置 GPU 容量付费。

River AI 致力于打造个人与公司的专属智能体，公司创始人 Babuschkin 认为，AI 应当是开放的、可自由获取的、可负担的，其服务对象应当为其使用者，而不是训练它的实验室。因而在长期规划的路径上，公司正在构建覆盖训练基础设施、个性化产品及新型硬件的完整技术栈，同时旨在打造一种能够学习用户习惯、由用户实际掌控，并足够了解用户以代表其利益行事的智能系统，并将其作为长期目标。现阶段，River AI 通过 API 将模型的掌控权赋予开发者和企业客户，未来，公司则计划将相同的能力进一步拓展至个人用户。

General Catalyst 董事总经理 Marc Bhargava 认为，当前 AI 技术的能力与多数企业在实际应用中的体验之间仍存在明显落差。许多企业缺乏一种兼具成本效益的方式，来根据自身需求训练、调优并真正拥有定制化的 AI 模型，而 River AI 的产品正是为了弥合这一鸿沟。

图表 10：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	
River AI	AI 全栈	General Catalyst、AMP PBC	种子轮及 A 轮	11 亿美元	11 亿美元	未披露
Lovable	AI 编程	Menlo Ventures、Scaleup Europe Fund	C 轮	4 亿美元	超 9.5 亿美元	133 亿美元

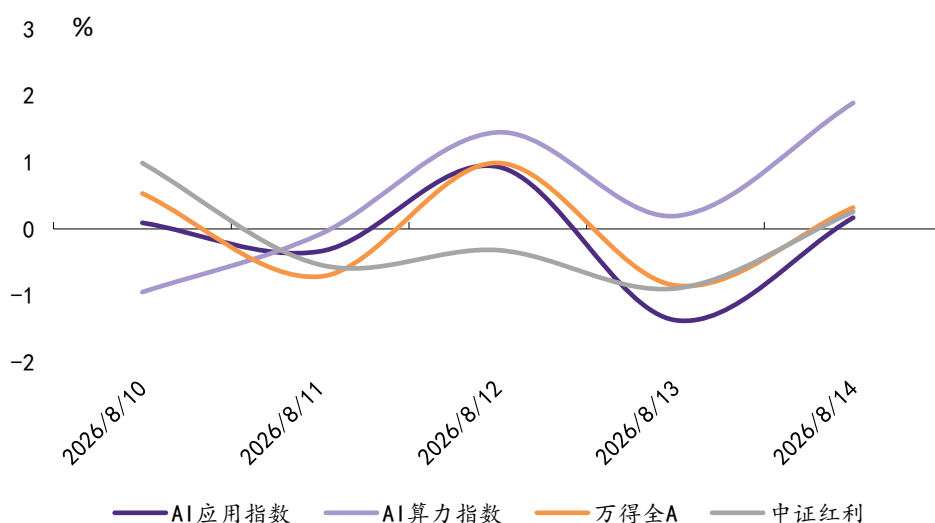
CodeRabbit	AI 代码审查	Atomico、Smash Capital	C 轮	1.43 亿美	超 2 亿美 元	15 亿美元
------------	---------	-----------------------	-----	---------	-------------	--------

资料来源：wind，Saasverse，华鑫证券研究

4、行情复盘

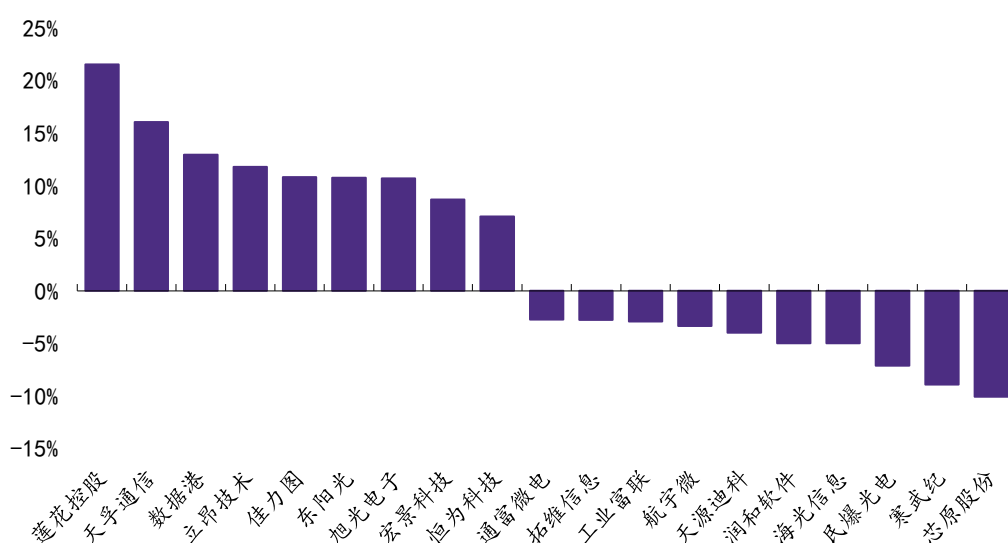
上周（2026.8.10-2026.8.14日），AI应用指数/AI算力指数/万得全A/中证红利日涨幅最大值分别为0.93%/1.89%/0.99%/0.99%，AI应用指数/AI算力指数/万得全A/中证红利日跌幅最大值分别为-1.37%/-0.95%/-0.85%/-0.89%。AI算力指数内部，莲花控股以21.54%录得上周最大涨幅，芯原股份以-10.07%录得上周最大跌幅。AI应用指数内部，鸿博股份以18.71%录得上周最大涨幅，中岩大地以-12.84%录得上周最大跌幅。

图表 11：上周（2026.8.10-2026.8.14日）指数日涨跌幅



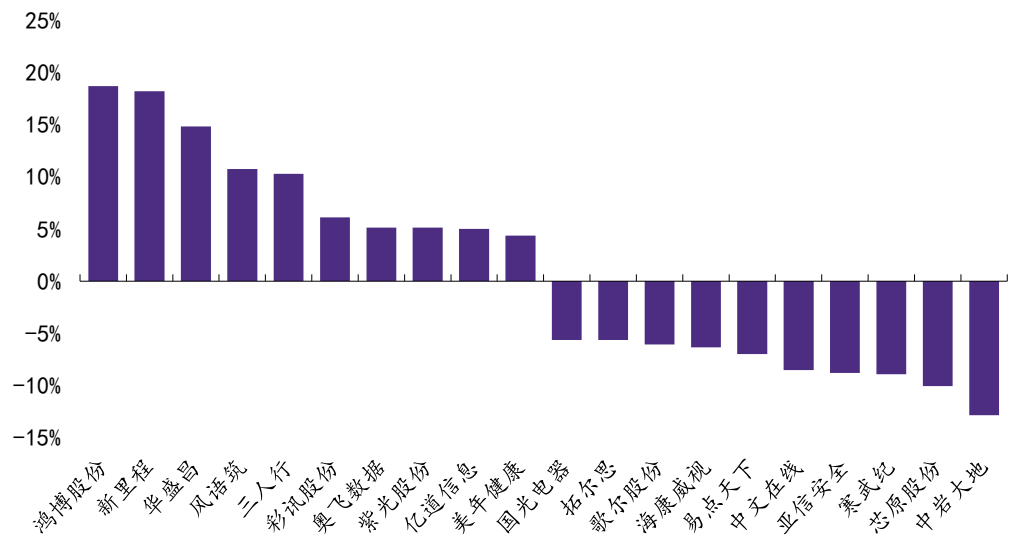
资料来源：wind, 华鑫证券研究

图表 12：上周（2026.8.10-2026.8.14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 13：上周（2026. 8. 10-2026. 8. 14 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

近日，Lumentum (LITE) 与 Coherent (COHR) 相继发布 2026 财年第四财季财报。Lumentum 当季营收 10.06 亿美元，单季首次突破 10 亿美元，同比增长 109.3%；non-GAAP 每股收益 3.23 美元（上年同期 0.88 美元）；non-GAAP 毛利率 50.4%（上年同期 37.8%）；公司指引 FY27Q1 营收 12.25 - 12.75 亿美元，non-GAAP 每股收益 4.05 - 4.35 美元。Coherent 当季营收 20.5 亿美元，同比增长 34%；non-GAAP 每股收益 1.74 美元（同比增长 74%）；non-GAAP 毛利率 40.2%；公司指引 FY27Q1 营收 22 - 24 亿美元、non-GAAP 每股收益 1.85 - 2.05 美元。电话会层面，CPO 产业化信号显著强化：Lumentum 第一大 CPO 客户量产计划不变且需求信号较上季增强，超高功率激光芯片预计 2027 年下半年放量，公司已获首个 ELS 模块订单（ASP 显著高于单激光芯片），管理层称超高功率激光“出货远落后于需求”，并已与 AXT 签订磷化铟衬底供应协议；OCS 出货环比翻倍，公司预计 FY27Q1OCS 收入将首次单季突破 1 亿美元。Coherent 管理层明确表示 CPO 需求“没有任何推迟、反而在提前”，CPO 收入预计自 2026 年 12 月季度开始贡献；公司将于 ECOC 发布集成光学平台 PhotonLink，覆盖激光到探测全链路并支持 CPO/NPO，相关产品同样自 12 月季度起贡献收入；产能端，6 英寸磷化铟产能预计本季末同比翻倍（较原计划提前一个季度）；订单端，FY2027 已基本订满，客户采购订单延伸至 2028 日历年，长期协议更延伸至本十年末。

两家财报印证了 CPO 产业已从预期验证迈入订单落地与产能扩张阶段，光通信景气能见度显著拉长。需求端看，Lumentum 第一大 CPO 客户量产计划不变且需求信号增强，Coherent 明确表示 CPO 节奏较原计划提前，直接证伪产业链内的推迟论，叠加 ELS 模块首单落地、Coherent 给出 CPO 收入的具体起点（2026 年 12 月季度）与平台化载体（PhotonLink），2027 年下半年 CPO 放量的可见度正在强化而非证伪。技术路线层面，Lumentum 强调 NPO 机会对其而言为完全增量、是迈向 CPO 的中间一步，Coherent 表示正与客户同步推进 NPO 与 CPO，技术路线摇摆不改光环节价值量。供需层面，超高功率激光等环节供不应求，Lumentum 锁定磷化铟衬底供应、Coherent 产能提前翻倍且订单排至 2028 年并含 take-or-pay 条款，表明未来两年行业增长的约束在供给而非需求，量价支撑扎实。我们认为，两家光通信龙头的业绩与指引相互印证，光通信是当前 AI 算力硬件产业链中供需格局最清晰、能见度最长的环节之一；对国内市场而言，海外龙头的订单能见度与产能扩张为国内光通信产业提供了直接的景气映射，建议关注国内光通信产业链的投资机会。

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 14: 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 15：重点关注公司及盈利预测

公司代码	名称	2026-08-17		EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E		
300757.SZ	罗博特科	531.00	-0.30	0.30	0.60	-1770.00	1770.00	885.00	买入	
301196.SZ	唯科科技	95.35	2.53	3.34	3.98	37.69	28.55	23.96	买入	
301369.SZ	联动科技	146.69	0.48	1.08	2.62	305.60	135.82	55.99	买入	
603859.SH	能科科技	42.30	0.92	1.05	1.32	45.98	40.29	32.05	买入	
688615.SH	合合信息	120.71	3.24	4.22	5.25	37.26	28.60	22.99	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：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。