2026 年 07 月 06 日

华鑫证券 CHINA FORTUNE SECURITIES

Anthropic 发布 Claude Sonnet 5，DSpark 与 JetSpec 共探投机解码效率新边界

—计算机行业周报

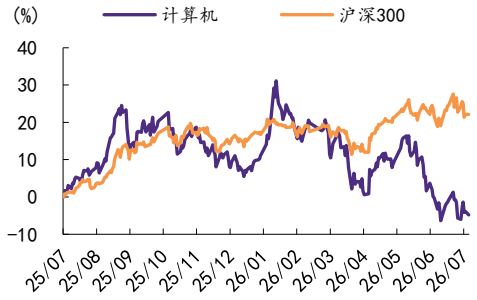
推荐(维持)

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)	-4.5	-5.4	-4.7
沪深 300	0.5	9.0	21.6

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：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

2、《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

3、《计算机行业点评报告：EverPure（P）：Q1 收入与 RPO 高增，数据云与 AI 存储需求支撑指引上修》2026-06-30

投资要点

算力：算力租赁价格平稳，Anthropic 发布 Claude Sonnet 5

2026 年 7 月 1 日，Anthropic 正式发布 Claude Sonnet 5。该模型能够自主规划任务、调用浏览器和终端工具，相较于上一代模型 Sonnet 4.6，该模型在推理、工具使用、编程和知识工作任务等方面性能显著提升，官方称其为迄今 Agent 能力最强的 Sonnet 模型。测评结果显示，该模型在多项基准测试中显著超越 Sonnet 4.6，部分指标逼近甚至反超旗舰模型 Opus 4.8。与此同时，该模型主打高性能、低成本路线，标准价格仅为旗舰模型 Opus 4.8 定价的六成。

AI 应用：Character.AI 周时长环比+9.94%，DSpark 与 JetSpec 共探投机解码效率新边界

近期，继 DeepSeek 推出其投机解码框架 DSpark 后，阶跃星辰也同期发布了名为 JetSpec 的技术方案。尽管两者都致力于解决大模型高频调用的问题，但其技术路径和侧重点存在显著差异：DSpark 更侧重于在服务系统层面优化验证环节，减少无效计算，而 JetSpec 则从 Draft 生成入手，通过因果并行树，旨在单次验证步骤中接受更多有效 Token。

AI 融资动向：Together AI 完成 8 亿美元 C 轮融资，投后估值达 83 亿美元

2026 年 7 月 2 日，开源 AI 云平台 Together AI 宣布完成 8 亿美元 C 轮融资，投后估值达 83 亿美元，较上一轮翻倍。本轮融资由 Aramco Ventures 领投，NVIDIA、General Catalyst 等多家机构参投，所得资金将主要用于拓展推理服务与扩充基础设施。该公司成立于 2022 年，核心产品为一套针对开源 AI 模型优化的训练与推理平台，技术上搭载自研的 ATLAS 引擎，推理速度最高可提升 400%。业务方面，公司第二季度年化订单额已突破 11.5 亿美元，客户群体Cursor、Cognition、Decagon 等 AI 原生企业，公司预计未来五年内计算容量与基础设施规模将扩展约 50 倍。

投资建议

DeepSeekV4 正式版计划于 7 月中旬上线，API 定价将引入结构性调整，首次采用"峰谷定价"机制。新版本推出 deepseek-v4-pro 与 deepseek-v4-flash 两款主力模型，按北京时间划

分平时段与高峰时段：工作日 9:00-12:00 及 14:00-18:00 为高峰时段，定价为平时段两倍；晚间、夜间及周末全天执行平时段价格。具体来看，v4-pro 平时段百万 tokens 输入（缓存命中/未命中）0.025 元/3 元、输出 6 元，高峰时段对应 0.05 元/6 元/12 元；v4-flash 平时段输入 0.02 元/1 元、输出 2 元，高峰时段对应 0.04 元/2 元/4 元。公司表示 V4 将带来更多功能优化和性能提升，此次定价调整旨在更合理配置资源、提升服务稳定性。

此次 DeepSeek 推出峰谷定价，前提是推理调用量已具备显著规模且存在真实波峰，标志着国产大模型从技术验证走向规模化商用。国产算力已从可选配置转为稀缺资源，产业链核心逻辑从替代性价比切换为供需定价。进一步看，高峰时段的算力调度对集群互联效率提出刚性约束，低延迟、高带宽互联方案从性能优化项升级为产能释放瓶颈。在推理需求持续放量、Token 消耗量快速增长的背景下，我们认为 CPO 作为高密度推理集群的核心互联路径，是 Token 经济学下的核心解。

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	2026-07-06	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	494.30	-0.30	0.30	0.60	-1647.67	1647.67	823.83	买入
301196.SZ	唯科科技	114.35	2.53	3.34	3.98	45.20	34.24	28.73	买入
301369.SZ	联动科技	224.18	0.48	1.08	2.62	467.04	207.57	85.56	买入
603859.SH	能科科技	48.50	0.92	1.21	1.50	52.72	40.08	32.33	买入
688615.SH	合合信息	105.50	3.24	4.22	5.25	32.56	25.00	20.10	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，Anthropic 发布 Claude Sonnet 5

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 6 月 29 日至 7 月 5 日，周度 Token 消耗量保持平稳，消耗量为 46.7T，环比上周无增减。在 Tokens 规模 Leaderboard 前五名中，DeepSeek 的 DeepSeek V4 Flash 以 5.34T tokens 位居榜首；Xiaomi 的 MiMo-V2.5 以 4.38T tokens 位列第二；MiniMax 的 MiniMax M3 以 4.11T tokens 位居第三；Tencent 的 Hy3 preview 以 3.13T tokens 位居第四；Z.AI 旗下的 GLM 5.2 以 2.58T tokens 位居第五。

文本类模型的市场份额，Google 以 897M requests 占据 28.2%的份额，位居第一；DeepSeek 以 652M requests 占据 20.5%，位列第二；OpenAI、Qwen、Anthropic 则分别以 516M、185M、167M requests，对应占据 16.2%、5.8%、5.2%的市场份额。

图像生成类模型调用量有所上升，调用量为 5.47M requests，环比上周增加 3.01%。在 Image Leaderboard 中，Google 的 Nano Banana 系列包揽榜单前四名，其中 Nano Banana (Gemini 2.5 Flash Image)、Nano Banana 2 (Gemini 3.1 Flash Image Preview)分别以 2.03M、855K requests 位居前二；Nano Banana 2 (Gemini 3.1 Flash Image)、Nano Banana Pro (Gemini 3 Pro Image Preview)分别以 659K、498K requests 位列第三、第四；xAI 旗下的 Grok Imagine Image Quality 以 382K requests 位列第五。从市场份额来看，Google 以 4.51M requests 占据 82.6%的份额，位居第一；xAI 以 382K requests 占据 7.0%，位列第二；OpenAI、Black Forest Labs、ByteDance Seed 则分别以 235K、149K、131K requests，对应占据 4.3%、2.7%、2.4%的市场份额。

嵌入模型调用量有所上升，调用量为 212M requests，环比上周增加 7.61%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 70.9M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 54.9M requests 位列第二；BAAI 的 bge-m3 以 16.5M requests 位列第三；OpenAI 的 Text Embedding 3 Large 以 11.8M requests 位列第四；Google 的 Gemini Embedding 001 以 10.2M requests 位列第五。从市场份额来看，OpenAI 以 85.3M requests 占据 40.2%的份额，位居第一；Qwen 以 61.2M requests 占据 28.8%，位列第二；Google、BAAI、Perplexity 则分别以 21.6M、19.3M、11.9M requests，对应占据 10.2%、9.1%、5.6%的市场份额。

重排序模型调用量有所下降，调用量为 1.44M requests，环比上周减少 10%。在 Rerank Leaderboard 中，Cohere 的 Rerank v3.5 以 496K requests 位列榜首；NVIDIA 的 Llama Nemotron Rerank VL 1B V2 以 394K requests 位列第二；Cohere 的 Rerank 4 Fast 以 308K requests 位居第三； Cohere 的 Rerank 4 Pro 以 239K requests 位居第四。从市场份额来看，Cohere 以 1.04M requests 占据 72.6%的份额，位居第一；NVIDIA 以 394K requests 占据 27.4%，位列第二。

视频生成类模型调用量大幅回落，调用量为 129K requests，环比上周减少 24.12%。在 Video Leaderboard 前五名中，Google 的 Veo 3.1 Fast 以 41K requests 位列第一；Google 的 Veo 3.1 以 23K requests 位列第二；ByteDance 的 Seedance 2.0 以 18K requests 位居第三；Google 的 Veo 3.1 Lite 以 15K requests 位居第四；xAI 的 Grok Imagine Video 以 8.83K requests 位居第五。从市场份额来看，Google 以 79K requests 占据 61.2%的份额，

位居第一；ByteDance 以 29K requests 占据 22.4%，位列第二；xAI、KwaiVGI、Alibaba 则分别以 9K、7K、5K requests，对应占据 6.8%、5.3%、3.8%的市场份额。

语音生成类模型调用量显著增长，调用量为 920K requests，环比上周增加 31.81%。在 Speech Leaderboard 前五名中，Google 的 Gemini 3.1 Flash TTS Preview 以 387K requests 位居第一；Hexgrad 的 Kokoro 82M 以 374K 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 69K requests 位列第四；Canopy Labs 的 Orpheus 3B 以 5.63K requests 位列第五。从市场份额来看，Google 以 387K requests 占据 42.1%的份额，位居第一；Hexgrad 以 374K requests 占据 40.7%，位列第二；xAI、Microsoft、Canopy Labs 则分别以 79K、69K、6K requests，对应占据 8.6%、7.5%、0.6%的市场份额。

语音转写类模型调用量有所回落，调用量为 6.95M requests，环比上周减少 5.44%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3 以 3.73M requests 位居第一；OpenAI 的 Whisper Large V3 Turbo 以 975K requests 位列第二；OpenAI 的 GPT-4o Mini Transcribe 以 947K requests 位列第三；NVIDIA 的 Parakeet TDT 0.6B v3 以 794K requests 位列第四；Qwen 的 Qwen3 ASR Flash 以 189K requests 位列第五。从市场份额来看，OpenAI 以 5.8M requests 占据 83.4%的份额，位居第一；NVIDIA 以 794K requests 占据 11.4%，位列第二；Qwen、Mistral AI、Google 则分别以 189K、73K、70K requests，对应占据 2.7%、1.0%、1.0%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。

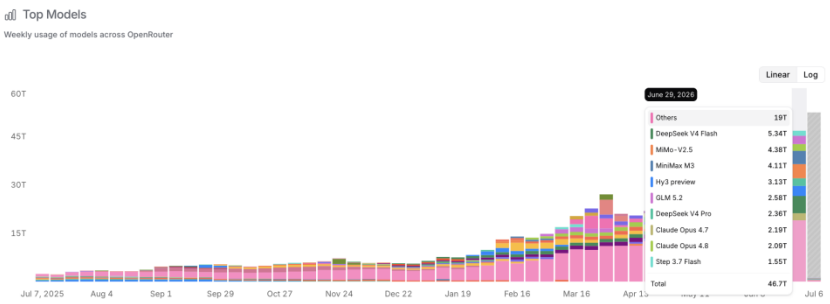
6 月 30 日，英伟达宣布在 Blackwell 平台上完成全栈推理优化，对于 DeepSeek V4 模型，相较 1 个月前上线初期，单 Token 处理成本已最多降至 1/5，创下行业最低水平。同时，英伟达已将单 Token 成本列为 AI 总拥有成本的核心考核指标。

7 月 1 日，摩根大通在最新发布的《数据中心观察》报告中表示，尽管 Token 价格有所下降，但大模型实际调用量仍在加速扩张，Token 支出已重新进入上行通道。同时，非超大规模云厂商市场的 GPU 租赁价格继续走高，DRAM 现货价格也保持坚挺。报告认为，模型降价并非需求疲软的信号，反而通过降低使用门槛、拓展应用场景，正进一步推动推理端需求扩张。

7 月 2 日，北京举办的“Token 普惠：交付可见价值 重塑 AI 效能”主题沙龙上，业内人士指出，当前 Token 市场存在计费不透明、计量标准混乱、价值匹配脱节三大核心痛点，呼吁建立全行业统一的 Token 计量基准，参照移动流量体系制定标准化核算规则，保障不同主体的 Token 计量具备可比性。

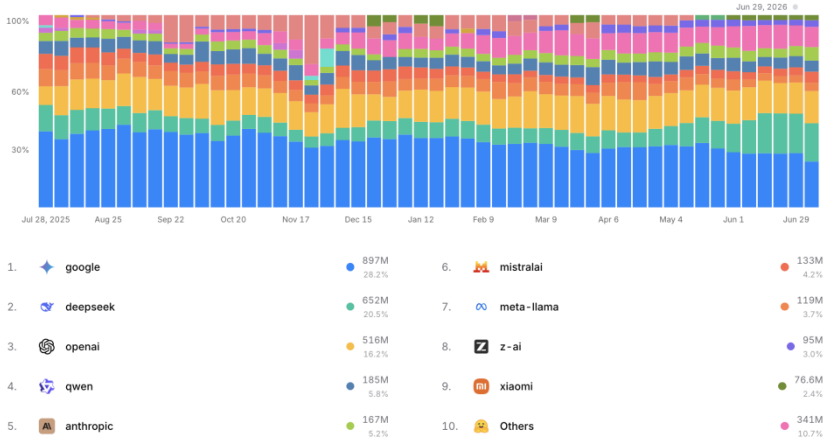
截至 7 月 3 日，硅谷数据 LLM Token 支出指数已从 5 月份的高点下跌近 20%。该指数于去年 12 月创立，此后经历近乎翻倍的增长，主要用于追踪 AI Token 付费情况。该指标的回落表明，AI 行业定价权正在承压，市场或将重新审视巨额资本开支的实际投资回报率。

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 Qoder 企业版，支持企业知识库与 Credits 灵活分配

2026 年 7 月 3 日，阿里云正式发布 Qoder 企业版，并围绕企业知识库与灵活计费两大核心进行了升级。

在企业知识库方面，企业版为开发者提供个人云端知识库 QMind，支持跨产品、跨设备、跨人员的知识共享。QMind 集成了 RepoWiki、本地文件、URL 等多种数据源，可将分散资料整合为个人或团队的可复用资产。其中，RepoWiki 可处理十万级文件代码库（如企业沉淀近十年的超大代码库），从模块功能、核心业务逻辑等维度快速建立代码理解。在此基础上，开发者在 Qoder Desktop 中进行跨库开发时，可直接调用其他仓库的代码语义和文档上下文，大幅提升开发效率。当前，QMind 已上架官方 Skill 市场，企业版用户可通过对话调用 Skill，对知识库进行增删改查。

在安全能力方面，企业版覆盖传输加密、身份认证与访问控制、AI 运行时、数据存储、审计合规五个安全领域，并通过 ISO/IEC 27001:2022 认证。同时，针对 AI Agent 的安全风险，Qoder 具备执行从命令拦截、语义分析、AI 风险判定到沙箱隔离逐层防护的能力，能够有效应对提示词注入、代码注入等攻击。

在计费模式方面，Qoder 企业版推出资源池化的 Credits 付费模式。企业以共享资源包方式持有 Credits，管理员可按需为成员或计费组动态分配额度，也可按群组或代码库配置可用模型范围，为不同敏感度的代码库匹配不同模型，实现精细化管理。

目前，Qoder 全系产品在全球拥有超 500 万用户，服务对象包括中国一汽、中信证券、亚信科技等企业。

1.3、 产业动态：Anthropic 发布 Claude Sonnet 5

2026 年 7 月 1 日，Anthropic 正式发布 Claude Sonnet 5。该模型能够自主规划任务、调用浏览器和终端工具，相较于上一代模型 Sonnet 4.6，该模型在推理、工具使用、编程和知识工作任务等方面性能显著提升，官方称其为迄今 Agent 能力最强的 Sonnet 模型。

测评结果显示，新模型在多项基准测试中显著超越 Sonnet 4.6，部分指标逼近甚至反超旗舰模型 Opus 4.8：

智能体编程方面，该模型在 SWE-bench Pro 基准上获得 63.2%的评分，超过前代模型 58.1%的评分，同时高于 GPT-5.5 的 58.6%以及 Gemini 3.5 的 55.4%，低于 Opus 4.8 的 69.2%。与此同时，在 Terminal-Bench 2.1 基准上，该模型取得 80.4%的评分，远高于前代模型 67.0%的评分，略低于 Opus 4.8 的 82.7%；

跨学科推理方面，该模型带工具在 Humanity's Last Exam 测试上获得 57.4%的评分，同一测试下评分同时超过 GPT-5.5 的 52.2%和 Gemini 3.1 Pro 的 51.4%，并且与 Opus 4.8 的 57.9%仅剩 0.5 个百分点的差距；

电脑操控能力方面，该模型在 OSWorld-Verified 基准上得分 81.2%，评分超过 GPT-5.5（78.7%）并且直追 Opus 4.8（83.4%）；

知识工作任务方面，该模型在 GDPval-AA v2 基准上拿到了 1618 的评分，超过了 Opus 4.8 的 1615。

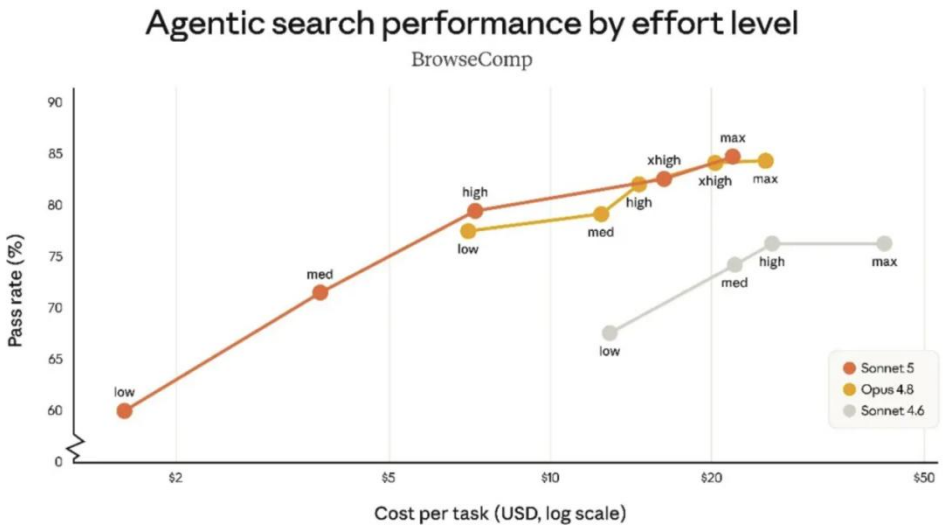
图表 3：Claude Sonnet 5 基准测试结果横向对比图

	Sonnet 5	Sonnet 4.6	Opus 4.8 For reference
Agentic coding SWE-bench Pro	63.2%	58.1%	69.2%
Agentic coding Terminal-Bench 2.1	80.4%	67.0%	82.7%
Multidisciplinary reasoning Humanity's Last Exam	43.2% no tools	34.6% no tools	49.8% no tools
	57.4% with tools	46.8% with tools	57.9% with tools
Computer use OSWorld-Verified	81.2%	78.5%	83.4%
Knowledge work GDPval-AA v2	1618	1395	1615

资料来源：智东西，华鑫证券研究

此外，在不同工作量水平下，Sonnet 5 相较 Sonnet 4.6 均实现了显著的性能提升。同时，与旗舰级的 Opus 4.8 相比，Sonnet 5 覆盖了更广泛的性价比区间：在中等工作量场景下，其成本效益更为突出；在特定条件下，高工作量下的性能表现可逼近 Opus 4.8。

图表 4：不同投入水平下性价比曲线示意图

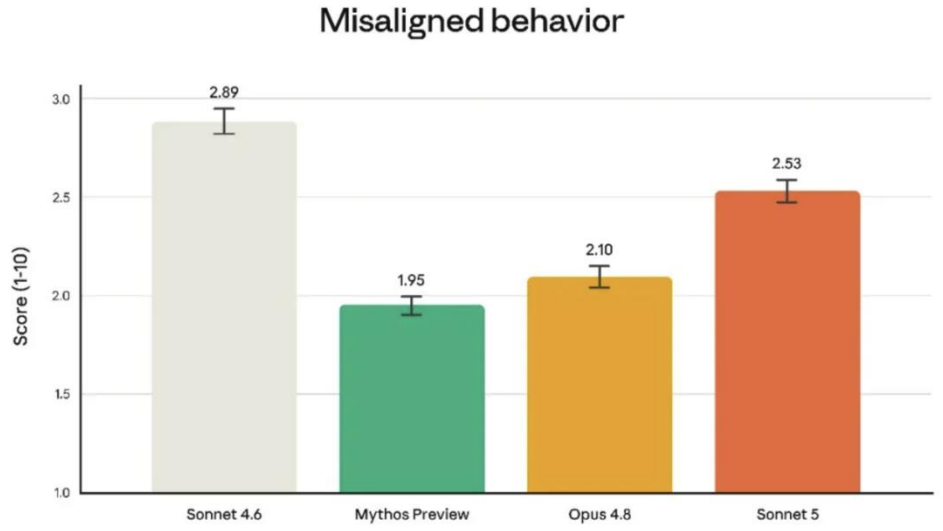


资料来源：智东西，华鑫证券研究

在安全能力方面，Sonnet 5 在防御恶意请求和即时注入攻击方面表现突出。数据显示，该模型提示注入攻击成功率仅为 0.19%，与 Opus 4.8 持平，显著优于 GPT-5.5 的 3.08%和 Gemini 3.5 Flash 的 6.66%；在浏览器注入防御上，该模型的攻击成功率低至 0.93%，而 Mythos 5 和 Opus 4.8 则分别为 29.7%和 31.5%。

与 Sonnet 4.6 相比，新模型表现出更低的幻觉和奉承行为发生率。在 Anthropic 的自动化行为审查中（该审计测试各类不协调行为，如滥用和欺骗等），Sonnet 5 的总体得分更低，即更安全。但与此同时，与 Opus 4.8 和 Mythos Preview 相比，Sonnet 5 在该评估中的不协调行为发生率略高。

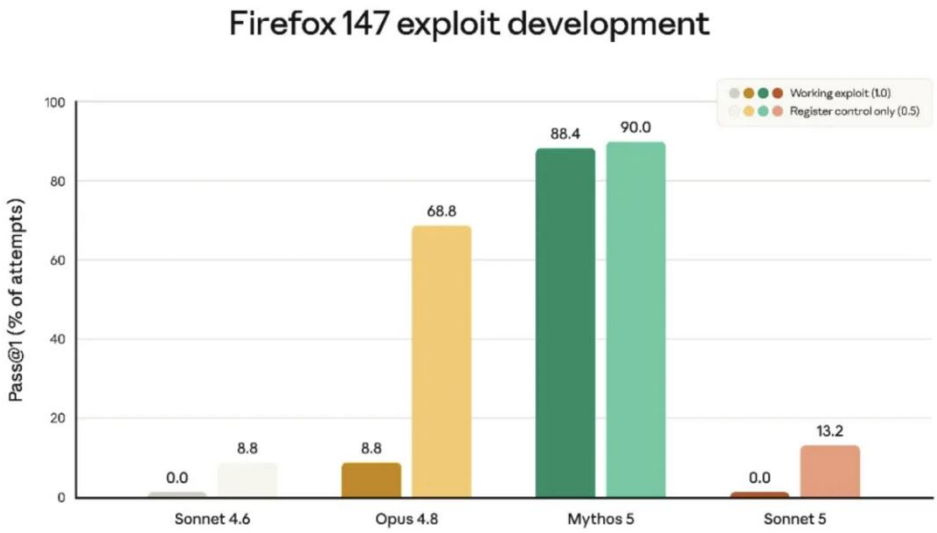
图表 5：不一致行为的发生率横向对比图



资料来源：智东西，华鑫证券研究

不仅如此，在测试潜在危险技能（如开发软件漏洞利用程序）的评估中，Sonnet 5 的表现远逊于 Opus 4.8 和 Mythos 5 等模型。在未进行过针对性网络安全任务执行训练的情况下，该模型仅能够执行一些常规的、无害的网络安全任务。

图表 6：软件漏洞利用程序开发成功率对比图



资料来源：智东西，华鑫证券研究

定价方面，Sonnet 5 主打高性能、低成本的性价比路线。从发布之日起至 2026 年 8 月 31 日，Claude Platform 上提供首发优惠价：每百万输入 token 2 美元（约合人民币 13.6 元），每百万输出 token 10 美元（约合人民币 67.9 元）。优惠期结束后，价格将调整为每百万输入 token 3 美元（约合人民币 20.4 元），每百万输出 token 15 美元（约合人民币 101.8 元），仅为 Opus 4.8 标准定价的 60%。

当前，Sonnet 5 已成为所有 Free 和 Pro 用户的默认模型，Max 版、团队版和企业版用户也可使用，同时，模型已上线 Claude Code 和 Claude Platform，开发者可通过 API 调用。

2、 AI 应用动态：Character.AI 周时长环比+9.94%，DSpark 与 JetSpec 共探投机解码效率新边界

2.1、 周流量跟踪：Character.AI 周时长环比+9.94%

本期（2026.6.26-2026.7.2）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1226.0M）、Bing（833.4M）和 Gemini（625.4M），访问量环比增速第一为 Discord（3.82%）；平均停留时长前三位分别为 Character.AI（00:15:40）、Discord（00:11:13）和 Kimi（00:08:00）;平均停留时长环比增速第一为 Character.AI（9.94%）。

图表 7：2026.6.26-2026.7.2 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	1226.0	-2.85%	6:15	3.02%
Bing	搜索	微软	833.4	0.97%	7:15	-0.91%
Gemini	聊天机器人	谷歌	625.4	-8.77%	7:06	1.43%
Canva	在线设计	Canva	184.6	-12.68%	5:55	1.43%
Discord	游戏社区	微软	149.6	3.82%	11:13	0.30%
Github	代码托管	微软	139	-4.20%	6:21	-0.78%
Character.AI	聊天机器人	Character.AI	38.75	-2.74%	15:40	9.94%
Deepl	翻译工具	DeepL	24.45	-6.00%	2:27	-0.68%
Perplexity	AI 搜索	Perplexity	24.11	-16.23%	4:24	0.76%
Kimi	聊天机器人	Moonshot AI	8.80	-4.07%	8:00	-4.76%
QuillBot	释义工具	QuillBot	7.373	-11.59%	2:52	-1.71%
NotionAI	文本/笔记	Notion	3.585	-38.55%	6:07	-19.52%
文心一言	聊天机器人	百度	0.38	-25.59%	2:51	-5.52%

资料来源：similarweb,华鑫证券研究

2.2、 产业动态：DSpark 面向服务级吞吐量精准控费，JetSpec 瞄准端到端解码延迟有效加速

近期，继 DeepSeek 推出其投机解码框架 DSpark 后，阶跃星辰也同期发布了名为 JetSpec 的技术方案。尽管两者都致力于解决大模型高频调用的问题，但其技术路径和侧重点存在显著差异：DSpark 更侧重于在服务系统层面优化验证环节，减少无效计算，而 JetSpec 则从 Draft 生成入手，通过因果并行树，旨在单次验证步骤中接受更多有效 Token。

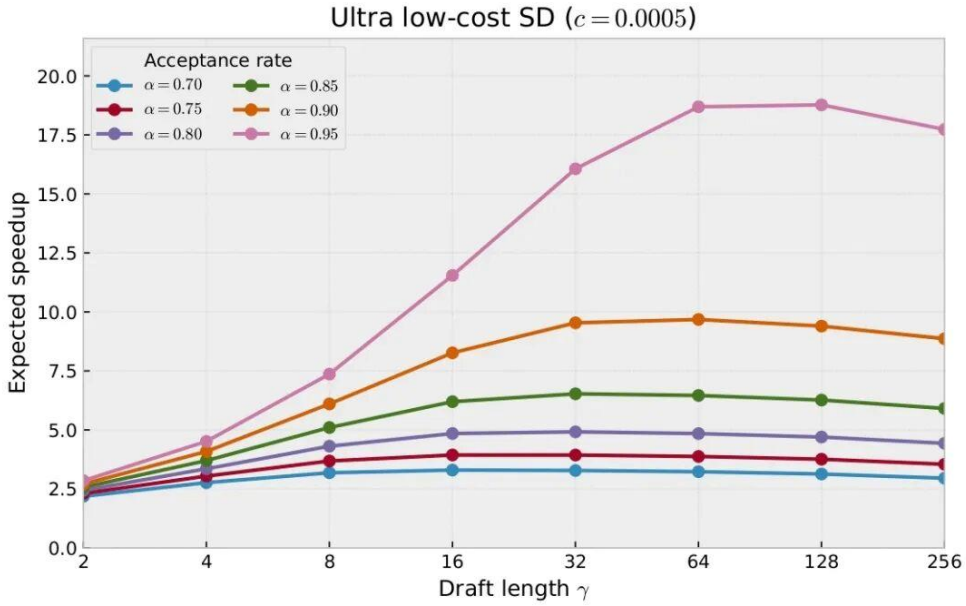
投机解码的核心理念在于利用轻量级草稿模型生成候选 token，再由目标大模型进行一次性并行验证，以此减少自回归生成所需的串行步骤。然而，草稿预算的增加并不必然带来速度的提升，真正的关键在于“接受率”。在低草稿生成成本的场景下，保持较高的逐 token 接受率尤为重要。理论公式表明，在低草稿成本但中等逐 token 接受率的情况下（例如当 c=0.05%，α≤0.8），最大理论加速比仍然低于 5 倍。

图表 8：投机解码的理论公式

$$\mathbb{E}[\#\text{tokens}] = \frac{1-\alpha^{\gamma+1}}{1-\alpha}, \quad \text{Speedup} = \frac{1-\alpha^{\gamma+1}}{(1-\alpha)(\gamma c+1)}$$

资料来源：机器之心，华鑫证券研究

图表 9：在不同逐 token 草稿成本和接受率下，投机解码的期望加速比随着草稿长度变化而变化

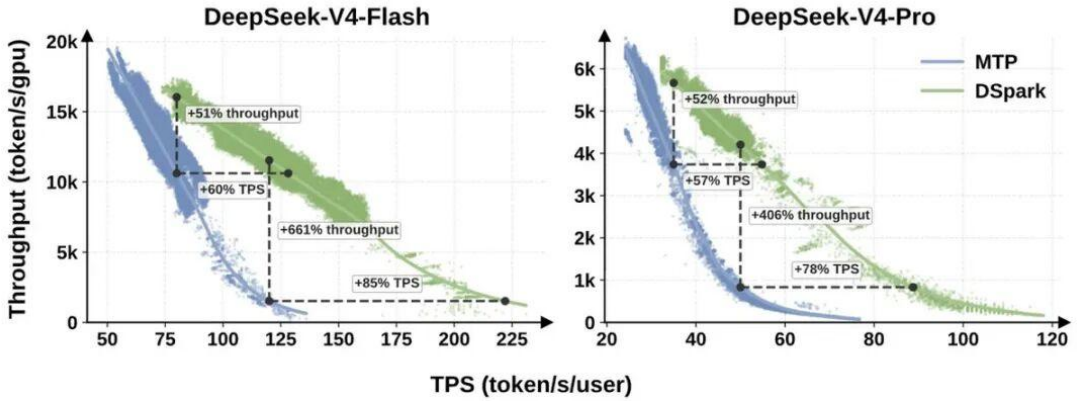


资料来源：机器之心，华鑫证券研究

当前投机解码面临着一个因果一致性与并行效率的两难困境。传统的自回归草稿方法，如 EAGLE 系列，虽能保证良好的因果一致性和高质量的候选令牌，但其串行生成步骤限制了扩展性，成本随树深度增长。反之，以 DFlash 为代表的块并行草稿方法，虽通过一次前向传播预测多个未来位置大幅降低了草稿成本，却因缺乏分支级的因果约束，导致生成的 token“局部合理、整体冲突”，接受率迅速被稀释，浪费了计算预算。DSpark 与 JetSpec 正是针对这一困境的两类互补性解法。

DSpark 面向的是高并发、吞吐量导向的生产服务场景。其策略是在保持并行草稿主干低成本优势的同时，引入轻量级的串行头和置信度估计。具体而言，对于每个草稿位置，DSpark 会生成基础 logits($z_i^{(0)}$)与对应的隐藏状态，并通过置信度头评估前缀相关性，随后仅将满足置信度阈值的最长前缀送入目标模型验证。这使得 DSpark 适用于预算感知的服务场景，提升整体吞吐量。在高并发场景下，DSpark 相比 MTP-1 基线，在吞吐量与每用户生成速度的权衡曲线上实现了明显改善，在 Flash 模型上展现出 60%至 85%的速度提升空间，在 Pro 模型上亦达到 57%至 78%。

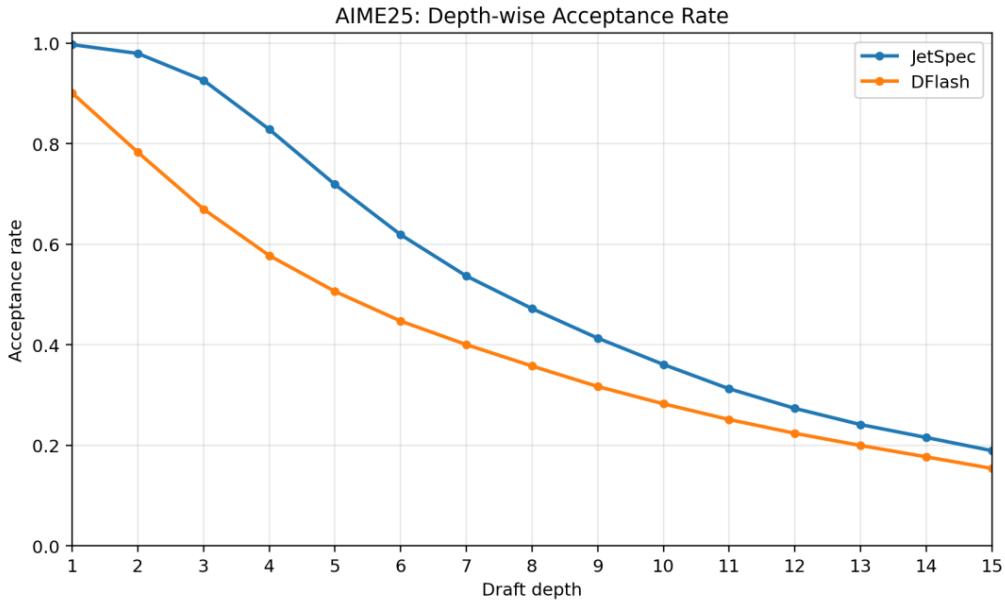
图表 10：DSpark 的吞吐量与每用户生成速度（TPS）关系曲线



资料来源：机器之心，华鑫证券研究

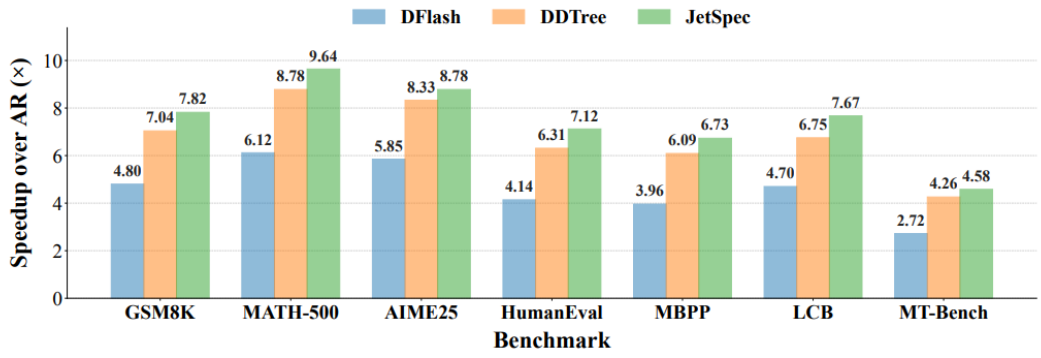
与 DSpark 不同，JetSpec 则集中在低并发、延迟导向的场景。在此类情境下，系统通常拥有更多的 FLOPs 预算，因此核心问题转变为如何把更高的计算预算转化为每次草稿—验证步骤中更多被接受的 token。JetSpec 的答案是通过设计因果并行草稿头，生成路径条件化的草稿树。这一设计让更深层的节点生成能够依赖同一分支上更早生成的 token。其效果是显著的：在 AIME25 上，JetSpec 在草稿深度 1 处的逐位置接受率几乎达到完美水平（约 99%），即便到深度 8 处，生存概率仍维持在约 50%。凭借此优势，JetSpec 在 Qwen3-8B 模型上实现了端到端解码最高达 9.64 倍的加速。在 MATH-500 任务上，其单次验证平均可接受高达 10.76 个 token，在 HumanEval、LiveCodeBench 及 MT-Bench 等代码与对话任务上，也分别实现了 7.12 倍、7.67 倍和 4.58 倍的显著加速。通过拟合，其有效逐 token 接受率约为 93%，显著高于 DFlash。

图表 11：DFlash 和 JetSpec 在 AIME25 上不同草稿深度位置的逐位置接受率



资料来源：机器之心，华鑫证券研究

图表 12：在 H100 GPU 上，跨数学、代码和对话基准测试中，相较于标准自回归解码的端到端解码加速比



资料来源：机器之心，华鑫证券研究

3、AI 融资动向：Together AI 完成 8 亿美元 C 轮融资，投后估值达 83 亿美元

2026 年 7 月 2 日，开源 AI 云平台 Together AI 宣布完成 8 亿美元 C 轮融资，投后估值达 83 亿美元，较 2025 年 2 月的上一轮融资相比，投后估值再度翻倍。本轮融资由 Aramco Ventures 领投，Vista Equity Partners、General Catalyst、Emergence Capital、NVIDIA、March Capital 等多家机构共同参投。此轮融资所得资金将主要用于拓展推理服务能力、大规模扩充基础设施容量，持续深化市场定位。

该公司成立于 2022 年，以 AI 原生云为自身定位，致力于大幅降低闭源模型成本，其核心产品为一套针对开源 AI 模型优化的云平台，用于支持企业在 DeepSeek、MiniMax、Kimi 等主流开源模型的训练与推理。具体来说，公司对外提供 Serverless 推理、专属基础设施推理、批量推理（Batch Inference）三种服务形态，其中，Serverless 推理服务的环境性能约为同类最快方案的两倍，开发者无需自行配置 GPU 和网络设备，即可直接调用开源 AI 模型，与此同时，专属基础设施推理服务专注于提供更高可靠性保障与定制化选项，批量推理服务则以高性价比优势区别于另外两种服务。

技术方面，平台底层由 NVIDIA 芯片驱动，搭载自研 ATLAS 引擎。该引擎核心采用推测解码（Speculative Decoding）技术，通过轻量模型预生成草稿、主模型校验修正的方式提升推理速度，部分推理工作负载最高可提速 400%。与此同时，产品方案自身具备差异化优势，其 ATLAS 技术能够根据用户需求变化自动调整轻量模型配置，有效避免精度下降。

此外，在模型训练方面，其训练集群可接入数千张 GPU，支持 Kubernetes 和 Slurm 两种管理工具，以分别满足易用性和高度定制化需求，同时内置自动故障检测修复功能，以降低因芯片异常导致训练流程中断或引入错误的风险。

业务方面，公司第二季度年化订单额已突破 11.5 亿美元，客户群体涵盖 Cursor、Cognition、Decagon 等 AI 原生企业。客户反馈显示，与闭源模型相比，在同等甚至更优性能下，成本节省幅度可达 6 倍至 60 倍：以 Decagon 为例，迁移至 Together AI 后推理成本降低六倍。麦肯锡数据显示，近四分之三企业计划加大开源 AI 使用。与此同时，公司预计未来五年计算容量扩展约 50 倍。

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	目前累计融资额	目前估值
Together AI	AI 原生云	Aramco Ventures	C 轮	8 亿美元	超 13.2 亿美元	83 亿美元

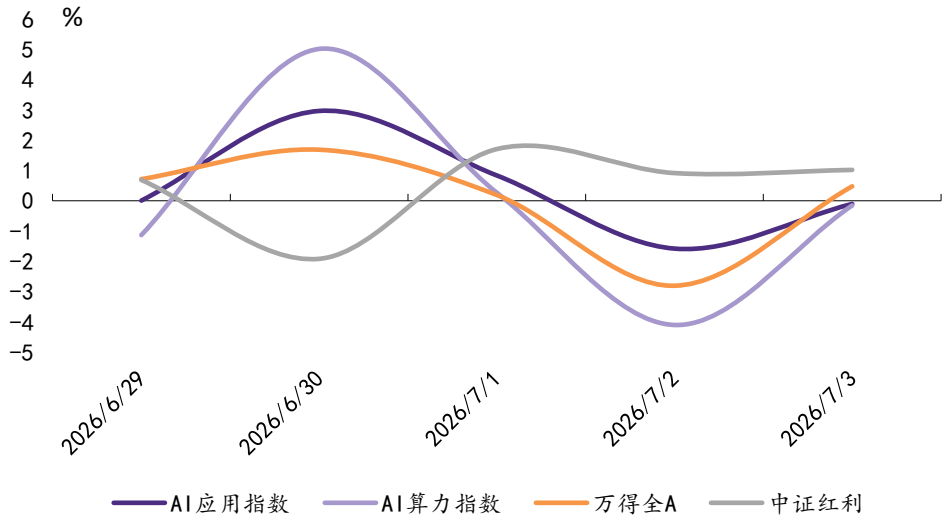
Tripo AI	AI 3D 大模型	未披露	A3 轮	1.5 亿美元	近 4 亿美元	未披露
Higharc	AI 原生建筑科技	Insight Partners	C 轮	9500 万美元	1.74 亿美元	未披露

资料来源：wind，Saasverse，华鑫证券研究

4、 行情复盘

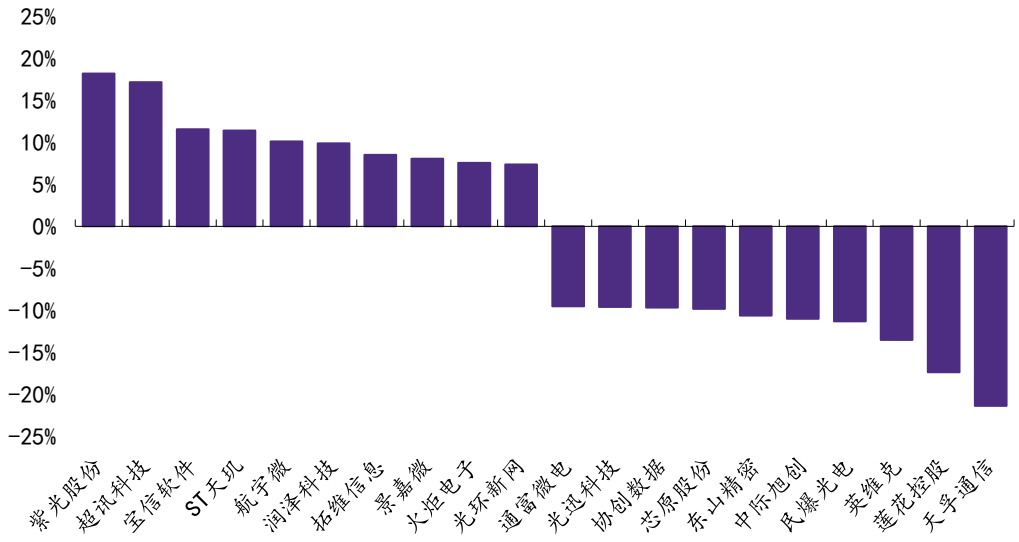
上周（2026.6.29-2026.7.3 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 2.97%/5.02%/1.69%/1.71%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-1.58%/-4.1%/-2.81%/-1.92%。AI 算力指数内部，紫光股份以 18.19%录得上周最大涨幅，天孚通信以-21.35%录得上周最大跌幅。AI 应用指数内部，格灵深瞳以 18.96%录得上周最大涨幅，美迪凯以-18.27%录得上周最大跌幅。

图表 14：上周（2026.6.29-2026.7.3 日）指数日涨跌幅



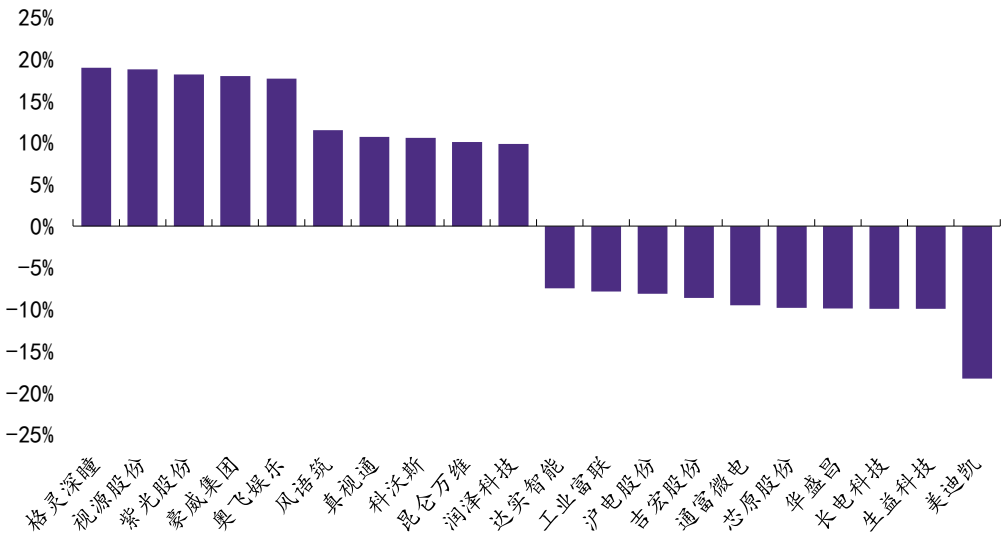
资料来源：wind，华鑫证券研究

图表 15：上周（2026.6.29-2026.7.3 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（2026.6.29-2026.7.3 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

DeepSeekV4 正式版计划于 7 月中旬上线，API 定价将引入结构性调整，首次采用"峰谷定价"机制。新版本推出 deepseek-v4-pro 与 deepseek-v4-flash 两款主力模型，按北京时间划分平时段与高峰时段：工作日 9:00-12:00 及 14:00-18:00 为高峰时段，定价为平时段两倍；晚间、夜间及周末全天执行平时段价格。具体来看，v4-pro 平时段百万 tokens 输入（缓存命中/未命中）0.025 元/3 元、输出 6 元，高峰时段对应 0.05 元/6 元/12 元；v4-flash 平时段输入 0.02 元/1 元2 元输出高峰时段对应 0.04 元/2 元/4 元。公司表示 V4 将带来更多功能优化和性能提升，此次定价调整旨在更合理配置资源、提升服务稳定性。

此次 DeepSeek 推出峰谷定价，前提是推理调用量已具备显著规模且存在真实波峰，标志着国产大模型从技术验证走向规模化商用。国产算力已从可选配置转为稀缺资源，产业链核心逻辑从替代性价比切换为供需定价。进一步看，高峰时段的算力调度对集群互联效率提出刚性约束，低延迟、高带宽互联方案从性能优化项升级为产能释放瓶颈。在推理需求持续放量、Token 消耗量快速增长的背景下，我们认为 CPO 作为高密度推理集群的核心互联路径，是 Token 经济学下的核心解。

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	2026-07-06	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	494.30	-0.30	0.30	0.60	-1647.67	1647.67	823.83	买入
301196.SZ	唯科科技	114.35	2.53	3.34	3.98	45.20	34.24	28.73	买入
301369.SZ	联动科技	224.18	0.48	1.08	2.62	467.04	207.57	85.56	买入
603859.SH	能科科技	48.50	0.92	1.21	1.50	52.72	40.08	32.33	买入
688615.SH	合合信息	105.50	3.24	4.22	5.25	32.56	25.00	20.10	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。