仅供内部参考，请勿外传

2026 年 10 月 13 日

华鑫证券 CHINA FORTUNE SECURITIES

OpenAI 上线 GPT-6.1 Sol，Anthropic 发布 Claude Haiku 5.5

—计算机行业周报

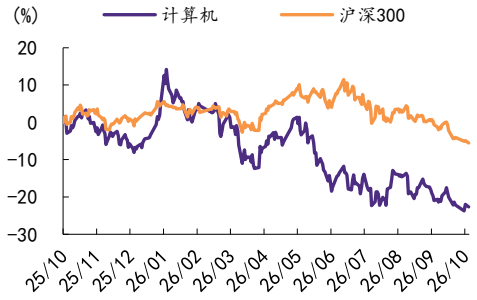
推荐(维持)

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)	-2.1	-6.5	-24.9
沪深 300	-4.9	-8.7	-6.7

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：GPT-6Sol 与 Luna 推出，Anthropic 发布 ClaudeOpus5.5》2026-09-30

2、《计算机行业点评报告：Oracle（ORCL.N）：算力交付带动 OCI 增长提速，AI 订单支撑收入扩张》2026-09-24

3、《计算机行业点评报告：Snowflake（SNOW.N）：产品收入增长提速，AI 应用扩展带动平台消费》2026-09-24

投资要点

算力：腾讯云开源 TeamAI，OpenAI 上线 GPT-6.1 Sol 极速版探索速度分层定价

腾讯云正式开源内部 AI 协作工具 TeamAI，基于 Git 打通跨 Agent 的 Skill 与配置代码化共享，支持按权限分发与知识图谱沉淀，测试显示低成本模型叠加团队知识可实现成本降 76%且评分提升 6.1%；OpenAI 上线 GPT-6.1 Sol 极速版（解码提速至 300 Token/秒）并推出即时转向功能，DeepSWE 基准对标旗舰 GPT-6 Astra 且单任务成本降至 1/5；同时建立标准、Fast 及 Ultrafast 三档速度阶梯定价，大模型产业正加速向组织级治理协同与细分场景差异化计费演进。

AI 应用：QuillBot 周时长环比+3.51%，Anthropic 发布 Claude Haiku 5.5

2026 年 10 月 8 日，Anthropic 正式推出模型 Claude Haiku 5.5，该模型被官方定位为迄今成本最低、响应速度最快且性能最强的小型人工智能系统。此次发布的核心亮点在于价格的大幅下调，其单位使用成本仅为前代 Haiku 4.5 的十分之一，同时也仅为 Sonnet 5.5 的二十分之一。Haiku 5.5 主要面向高容量、成本敏感型任务，能够高效处理摘要生成、文本压缩、数据库查询以及分类请求等重复性高、速度要求快的工作负载，并且能够作为 Opus 5.5 与 Sonnet 5.5 的辅助子代理，在编程等场景中发挥协同作用。

AI 融资动向：TypeSafe AI 完成 8.7 亿美元 A 轮融资，估值达 75 亿美元

TypeSafe 完成 8.7 亿美元 A 轮融资，估值攀升至 75 亿美元，由 a16z 领投，红杉资本参投。核心模型 Jev 基于 Transformer 但专攻非文本决策，直接返回类型化值与概率分布，分类推理速度达前沿 LLM 的 100－200 倍且成本低至 1/400；发布 3 天生成量破 1 万亿 token，财富 500 强中约三分之一已接入；其嵌入 Agent 循环替代高频 LLM 调用，推动 AI 以软件原生语言深度融入基础架构。

投资建议

10 月 6 日，Sierra 与 Meta 联合发布个人智能体协议（PersonalAgentProtocol，代号 Poppy），首批合作伙伴包括 Genesys、Instinct、Rocket、Shopify、Stripe 与

仅供内部参考，请勿外传

Walmart。该协议为开放标准，定义个人智能体代表用户与企业交互时的身份认证、权限与会话管理方式，基于 OAuth 等现有标准构建；企业可据此识别来访的智能体及其代表的用户，并选择通过网页、API 或自有智能体承接任务。10 月 9 日，Sierra 公布协议草案并新增 35 家设计伙伴，包括 OpenAI 、Visa 、Mastercard 、PayPal 、美国银行、Cloudflare、Target、美联航等，覆盖模型厂商、支付网络、银行、零售与航空多个环节；支付、消息推送等能力暂列为后续扩展方向。

本次协议发布的核心信号在于，智能体竞争已经延伸至交互标准的制定。先前公司 Muse 智能体的爆发式增长验证了消费级智能体的需求，而当智能体开始代表用户完成购物、改签、售后等真实交易，智能体与企业系统之间缺乏统一的交互规则就成为新的瓶颈，规则的制定权将成为入口之争以外的第二层竞争。从参与方结构看，发起方与新增伙伴覆盖模型厂商、支付网络、银行与零售龙头，头部模型 OpenAI 亦以设计伙伴身份加入，交易链条上的关键角色已齐聚同一协议框架。我们认为，个人智能体协议的落地标志着智能体经济开始完善规则层基础设施，智能体与支付、电商、客服系统标准化对接将成为 AI 应用落地的下一条主线。

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	2026-10-13	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	481.60	-0.40	2.13	4.49	-1204.00	226.10	107.26	买入
301196.SZ	唯科科技	75.38	2.05	1.92	2.39	39.28	39.23	31.50	买入
301369.SZ	联动科技	111.17	0.48	1.08	2.62	231.60	102.94	42.43	买入
603859.SH	能科科技	36.57	0.92	1.05	1.32	39.75	34.83	27.70	买入
688615.SH	合合信息	84.17	3.24	4.22	5.25	25.98	19.95	16.03	买入

资料来源：Wind，华鑫证券研究

仅供内部参考，请勿外传

正文目录

1、 算力动态：腾讯TEAMAI开源，OPENAI 上线 GPT-6.1 SOL 极速版探索速度分层定价 5

1.1、 Tokens 跟踪 5
1.2、 算力租赁动态：租金价格整体平稳 7
1.3、 云厂动态：腾讯云开源团队 AI 协作工具 TeamAI，打通跨 Agent 资产共享与代码化治理体系 9
1.4、 产业动态：OpenAI 上线 GPT-6.1 Sol 极速版拓展差异化定价，推出即时转向强化协同交互 9

2、 AI 应用动态：QUILLBOT 周时长环比+3.51%，ANTHROPIC 发布 CLAUDE HAIKU 5.5 12

2.1、 周流量跟踪：QuillBot 周时长环比+3.51% 12
2.2、 产业动态：Anthropic 发布 Claude Haiku 5.5，价格降低且性能大幅跃升 12

3、 AI 融资动向：TYPESAFE AI 完成 8.7 亿美元 A 轮融资，估值达 75 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） 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） 7

图表 3：2026.10.5 -2026.10.12 日 H100 价格（单位：美元/小时） 7

图表 4：2026.10.5 -2026.10.12 日 H200 价格（单位：美元/小时） 8

图表 5：2026.10.5 -2026.10.12 日 A100 价格（单位：美元/小时） 8

图表 6：2026.10.5 -2026.10.12 日 B200 价格（单位：美元/小时） 8

图表 7：GPT-6.1 SOL 极速版与标准版输出性能与响应延迟实测对比 10

图表 8：GPT-6.1 SOL 各档位与 GPT-6 ASTRA API 资费及生成速度对比 11

图表 9：2026.10.3-2026.10.9 AI 相关网站流量 12

图表 10：CLAUDE HAIKU 5.5 的定价 13

图表 11：CLAUDE HAIKU 5.5 在多个测试中的表现 14

图表 12：CLAUDE HAIKU 5.5 查询离首都机场最近的酒店示例 15

图表 13：上周 AI 初创公司融资动态 16

图表 14：上周（2026.10.5-2026.10.9 日）指数日涨跌幅 18

图表 15：上周（2026.10.5-2026.10.9 日）AI 算力指数内部涨跌幅度排名 18

图表 16：上周（2026.10.5-2026.10.9 日）AI 应用指数内部涨跌幅度排名 19

图表 17：FICONTEC2025 年年中至今公告订单 20

仅供内部参考，请勿外传

图表 18：重点关注公司及盈利预测 21

仅供内部参考，请勿外传

1、算力动态：腾讯云开源 TeamAI，OpenAI 上线 GPT-6.1 Sol 极速版探索速度分层定价

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 10 月 5 日至 10 月 11 日，周度 Token 消耗量小幅增长，消耗量达 172T，环比上周增加 3.61%。在 Tokens 规模 Leaderboard 前五名中，DeepSeek 的 DeepSeek V4.1 Flash 以 40.97T tokens 稳居榜首；Z-ai 的 GLM 5.3 Flash 以 12.24T tokens 位列第二；Xiaomi 的 MiMo-V2.6-Flash 以 11.46T tokens 位列第三；StepFun 的 Step 5 Preview 以 8.84T tokens 位列第四；Tencent 的 Hy4 preview 以 8.28T tokens 位列第五。

文本类模型的市场份额方面，deepseek 占据 24.5%的份额，位列第一；openai 占据 18.6%的份额，位列第二；google、z-ai、qwen 分别占据 17.8%、13.4%、5.4%的市场份额。

图像生成类模型调用量稳步增长，调用量为 8.03M requests，环比上周增加 6.64%。在 Image Leaderboard 中，Google 的 Gemini 3.1 Flash Image 以 1.35M requests 位居第一；OpenAI 的 GPT Image 2.5 Sunburst 以 1.09M requests 位列第二；Google 的 Gemini 3.1 Flash Lite Image 以 736K requests 位列第三；Meta 的 Muse Image 1.0 以 629K requests 位列第四；OpenAI 的 GPT Image 2.5 Flare 以 611K requests 位列第五。从市场份额来看，google 以 3.24M requests 占据 40.3%的份额，位列第一；openai 以 2.18M requests 占据 27.2%，位列第二；meta 以 629K requests 占据 7.8%的份额。

嵌入模型调用量稳步增长，调用量为 370M requests，环比上周增加 7.25%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 160.23M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 53.97M requests 位列第二；Google 的 Gemini Embedding 2 以 40.37M requests 位列第三；Google 的 Gemini Embedding 001 以 24.95M requests 位列第四；BAAI 的 bge-m3 以 20.61M requests 位列第五。从市场份额来看，openai 以 179.01M requests 占据 48.4%的份额，位居第一；google 以 65.31M requests 占据 17.7%，位列第二；qwen、baai、perplexity 分别以 53.97M、20.61M、8.72M requests，对应占据 14.6%、5.6%、2.4%的市场份额。

重排序模型调用量小幅增长，调用量为 12.3M requests，环比上周增加 6.03%。在 Rerank Leaderboard 前五名中，VoyageAI 的 rerank-2.5 以 4.32M requests 位居榜首；VoyageAI 的 rerank-2.5-lite 以 3.04M requests 位列第二；VoyageAI 的 rerank-3-lite 以 1.29M requests 位列第三；Cohere 的 Rerank v3.5 以 970K requests 位列第四；Qwen 的 Qwen3 Reranker 8B 以 898K requests 位列第五。从市场份额来看，voyageai 以 9.23M requests 占据 75.3%的份额，位列第一；cohere 以 2.03M requests 占据 16.5%，位居第二；qwen、nvidia 分别以 898K、101K requests，对应占据 7.3%、0.8%的市场份额。

视频生成类模型调用量显著增长，调用量为 336K requests，环比上周增加 21.74%。在 Video Leaderboard 前五名中，Heygen 的 Video 1 以 55K requests 位居第一；Google 的 Veo 3.1 Lite 以 49K requests 位列第二；ByteDance 的 Seedance 2.5 以 31K requests 位列第三；ByteDance 的 Seedance 2.0 以 22K requests 位列第四；ByteDance 的 Seedance 2.0 Fast 以 19K requests 位列第五。从市场份额来看，bytedance 以 85K requests 占据 26.3%的份额，位列第一；google 以 67K requests 占据 19.8%，位居第二；heygen、alibaba、

仅供内部参考，请勿外传

kwaivgi 分别以 55K、19K、12K requests，对应占据 16.2%、5.7%、3.7%的市场份额。

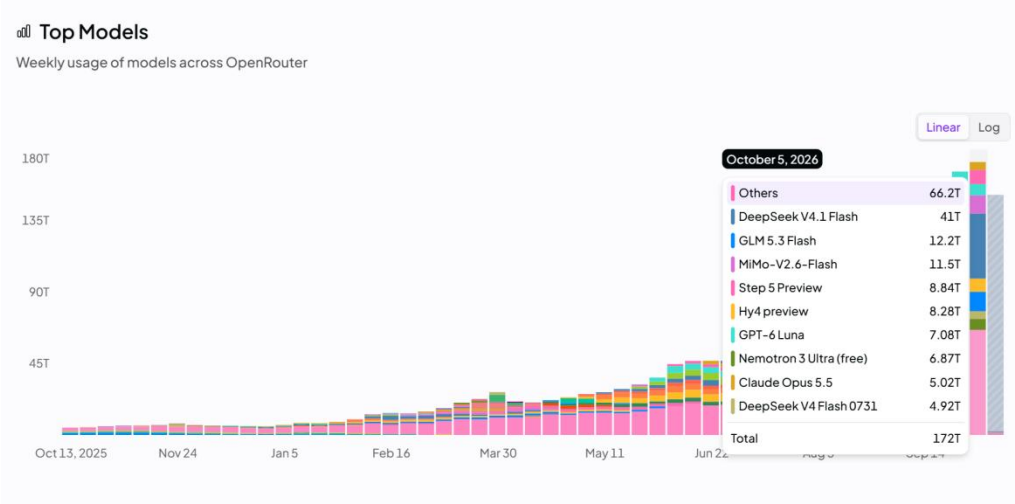
语音生成类模型调用量显著增长，调用量为 4.92M requests，环比上周增加 35.91%。在 Speech Leaderboard 前五名中，Google 的 Gemini 3.8 Flash TTS 以 1.54M requests 位居第一；Google 的 Gemini 3.1 Flash TTS Preview 以 607K requests 位列第二；Hexgrad 的 Kokoro 82M 以 535K requests 位列第三；Google 的 Gemini 3.8 Flash Lite TTS 以 521K requests 位列第四；xAI 的 Grok Voice TTS 1.0 以 321K requests 位居第五。从市场份额来看，google 以 2.67M requests 占据 54.3%的份额，位居第一；hexgrad 以 535K requests 占据 10.9%，位列第二；fish-audio、x-ai、elevenlabs 分别以 399K、321K、306K requests，对应占据 8.1%、6.5%、6.2%的市场份额。

语音转写类模型调用量显著增长，调用量为 28.09M requests，环比上周增加 37.75%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3 以 10.86M requests 位居第一；OpenAI 的 Whisper Large V3 Turbo 以 5.46M requests 位列第二；Microsoft 的 MAI-Transcribe 2 以 3.10M requests 位列第三；ElevenLabs 的 Scribe v2 以 2.99M requests 位列第四；OpenAI 的 GPT-4o Mini Transcribe 以 1.01M requests 位列第五。从市场份额来看，openai 以 18.41M requests 占据 65.6%的份额，位居第一；microsoft 以 3.10M requests 占据 11.0%，位列第二；elevenlabs、nvidia、x-ai 分别以 2.99M、623K、580K requests，对应占据 10.6%、2.2%、2.1%的市场份额。

* 统计口径说明：以上数据均来源于 OpenRouter。当前 Rankings 采用 Token 与 Request 两套统计维度：仅 Tokens 规模 Leaderboard 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。

10 月 11 日，润建股份牵头组建的“广西词元经济及智能应用概念验证中心”成功入选首批广西科技成果转化基地。该中心联合广西大学、东盟技术转移中心等多家单位共建，立足区位优势，依托五象云谷词元工厂智算底座与 Token 服务，聚焦解决 AI 应用验证成本高、转化链条长与跨境落地难等痛点，打造面向东盟的词元与智能应用跨境流通枢纽，开辟中国—东盟数字经济合作新通道。

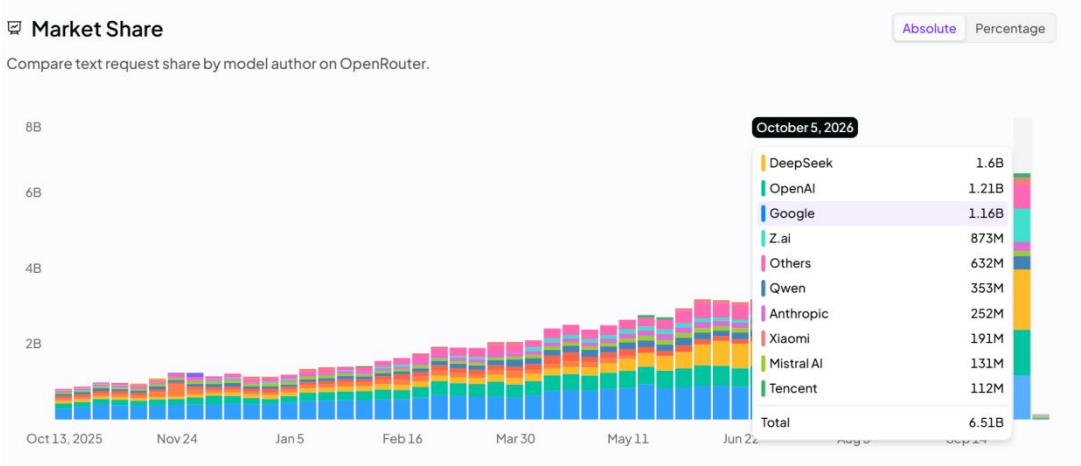
图表 1：Tokens 规模 Leaderboard（Token 消耗量为统计口径）



资料来源：OpenRouter，华鑫证券研究

仅供内部参考，请勿外传

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）



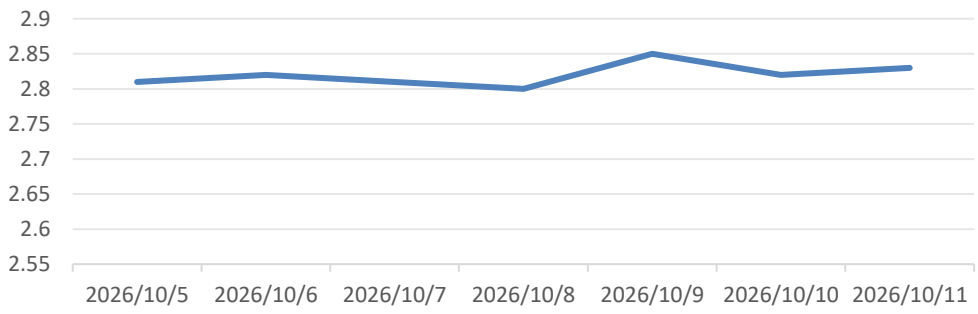
资料来源：OpenRouter，华鑫证券研究

1.2、 算力租赁动态：租金价格整体平稳

本周（2026.10.5-2026.10.11 日），算力租赁市场整体表现平稳。H100 指数/H200 指数/A100 指数/B200 指数日租金（美元/小时）最大值分别为 2.85/3.44/1.53/5.93，H100 指数/H200 指数/A100 指数/B200 指数日租金（美元/小时）最小值分别为 2.80/3.29/1.52/5.87，未出现明显波动。

* 计算方法说明：以上数据均来源于 silicon data。每条 GPU 租赁记录首先会被转换为标准化的每 GPU 每小时费率——即单价等于总租赁价格除以（GPU 数量乘以租赁时长）。预处理后的数据会经过多阶段过滤，以剔除上报错误，同时不会忽略真实的结构性价格变动。由于单纯汇总价格会随每日配置组合的变化而波动，因此我们采用基差调整（basis adjustment）将每条记录标准化为基准等价合约，在此过程中会综合考虑租赁类型、地理位置、CPU 平台以及 GPU 型号等因素。经基差调整后的价格会在提供商层面进行汇总，最终的指数值是各提供商数据的加权平均值。

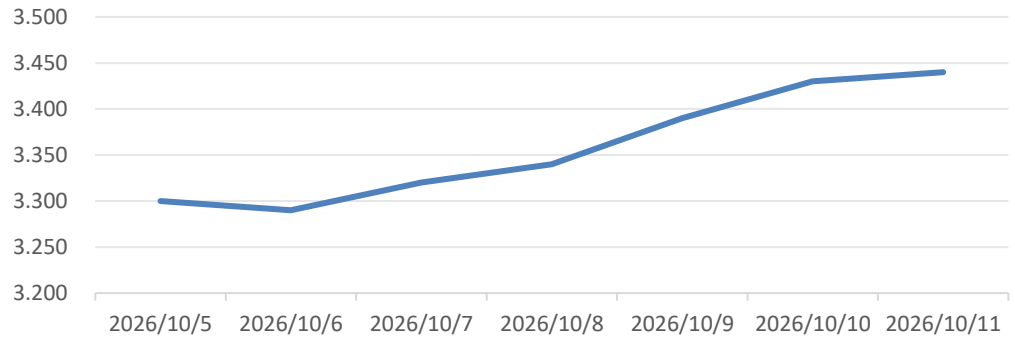
图表 3：2026.10.5 -2026.10.12 日 H100 价格（单位：美元/小时）



资料来源：silicon data，华鑫证券研究

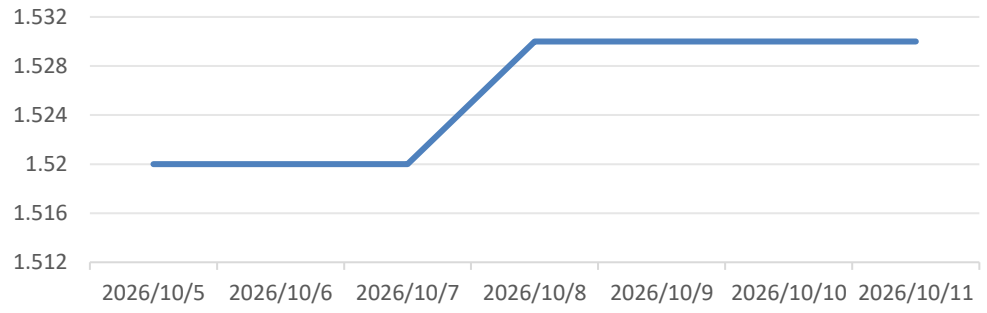
仅供内部参考，请勿外传

图表 4：2026.10.5 -2026.10.12 日 H200 价格（单位：美元/小时）



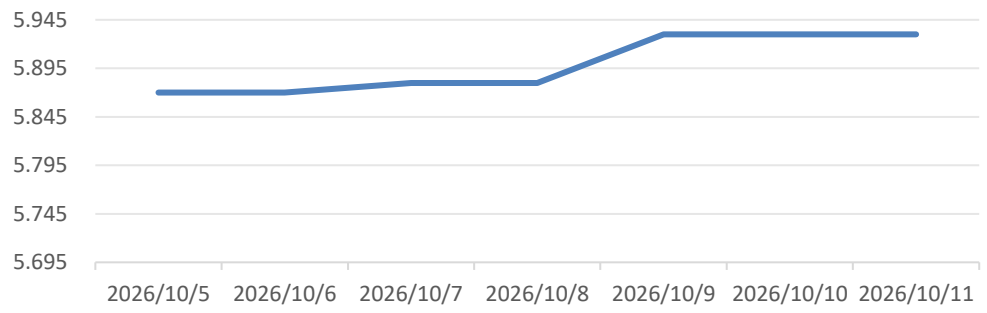
资料来源：silicon data，华鑫证券研究

图表 5：2026.10.5 -2026.10.12 日 A100 价格（单位：美元/小时）



资料来源：silicon data，华鑫证券研究

图表 6：2026.10.5 -2026.10.12 日 B200 价格（单位：美元/小时）



资料来源：silicon data，华鑫证券研究

仅供内部参考，请勿外传

1.3、 云厂动态：腾讯云开源团队 AI 协作工具 TeamAI，打通跨 Agent 资产共享与代码化治理体系

2026 年 10 月 10 日，腾讯云正式开源其内部自用的团队 AI 协作工具 TeamAI。针对企业研发中多成员使用不同 Agent 导致技能碎片化、配置无法跨端复用的痛点，TeamAI 基于 Git 构建，将 Skill、工作规范、MCP 工具配置、Rules、子 Agent 及项目知识统一纳入 Git 仓库，实现代码化评审、版本管理与回滚，并自动同步至各成员本地 Agent，达成“一人配置，全员复用”。项目已登顶 GitHub Trending，支持 GitHub、CNB 等代码平台并提供官方模板 teamai-hub，目前已适配 WorkBuddy、CodeBuddy、Claude Code、Codex、Cursor 等 16 种主流 Agent；同时，TeamAI 已集成至腾讯云智能体管理平台 ClawPro，支持管理员通过管控台对已接入 Agent 进行统一配置管理与分发。

协同机制与安全管控方面，TeamAI 沿用开发者熟悉的 Git 协作流程。成员提交或修改规范时会自动发起合并请求交由团队评审，合入后支持启动 Hook 的 Agent 将在新会话开始时自动拉取更新，确保修改全留痕且可快速回滚，避免配置版本分裂。平台支持按角色、项目或标签进行分发，满足不同岗位差异化配置需求；针对内部数据安全，TeamAI 新增模型配置统一下发功能，管理员可统一指定符合安全要求的模型服务，防范成员自行接入引发的安全风险。在交互体验上，管理员与普通成员安装 Skill 后，均可在适配工具内通过自然语言完成建仓初始化、团队加入以及经验技能一键分享。

经验沉淀与降本增效表现方面，TeamAI 构建了经验闭环沉淀机制。在实际交互中，系统可根据反复纠正、工具调用失败等触发条件提示成员沉淀经验，由 Agent 将排障方法整理成文档入库，供后续任务检索参考；面向研发场景，TeamAI 支持构建代码知识图谱，辅助 Agent 定位相关模块源码。团队可基于采纳数据持续优化经验库，将高频方案沉淀为正式文档、Skill 或规则。官方内部测试数据显示，在涉及 13 个研发任务的测试中，接入团队知识的较低成本模型相比未接入知识的更高规格模型，综合评分提高 6.1%，模型调用成本显著降低 76%。

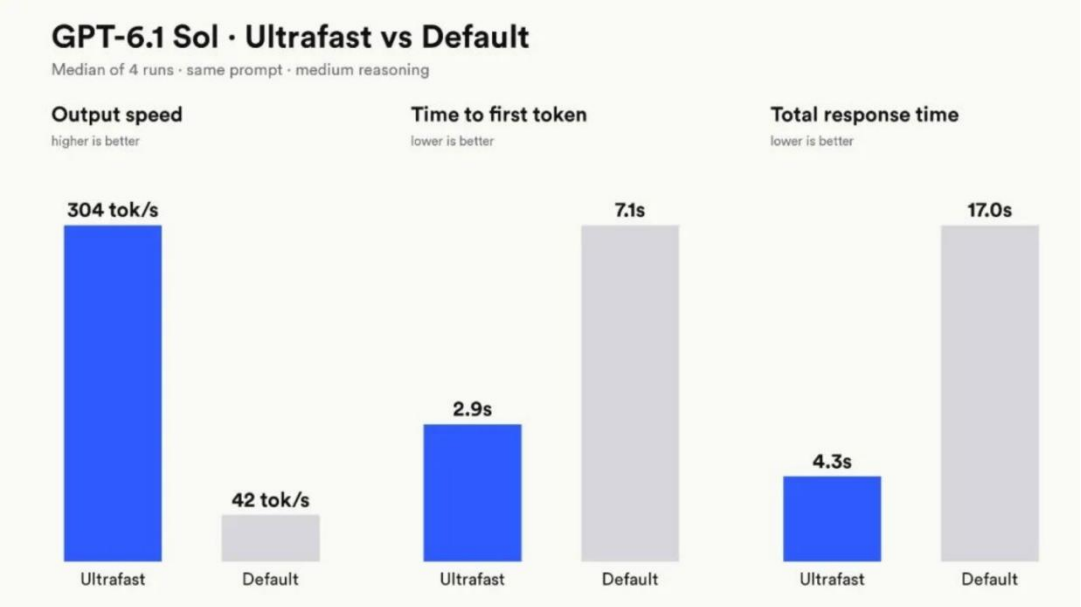
1.4、 产业动态：OpenAI 上线 GPT-6.1 Sol 极速版拓展差异化定价，推出即时转向强化协同交互

2026 年 10 月 9 日，OpenAI 正式上线 Ultrafast 推理模式（极速版）并下放至 GPT-6.1 Sol 模型，通过 API、Codex 与 ChatGPT Work 三端同步推送。在模型性能与架构指标方面，GPT-6.1 Sol 极速版输出解码速度达到标准模式的 8 倍（约每秒 300 个 Token）。在第三方评测机构 Artificial Analysis 的智能指数测试中，GPT-6.1 Sol 开满推理档位取得 52 分，仅比旗舰模型 GPT-6 Astra 低 1 分；在长程软件工程基准 DeepSWE v1.1 测试中，GPT-6.1 Sol 与旗舰模型 GPT-6 Astra 打成平手，单项任务成本约为 Astra 的五分之一；在电脑操作基准 OSWorld 2.0 测试中，Sol 得分仅比 Astra 低 2.1 分，单个任务的花费约为 Astra 的七分之一。在此前基准测试中，标准版 Sol 每秒输出约 67 个 Token，低于同类模型中位数水平，极速版的推出补齐了该模型在实时响应速度上的短板。在第三方开发者实测中，针对相同提示词测试显示，标准版 Sol 输出速度为 42 Token/秒，极速版输出速度达 304 Token/秒（提速超 7 倍），首字延迟从标准版的 7.1 秒压降至 2.9 秒，整段回答耗时由 17 秒缩减至 4.3 秒。伴随极速版发布，OpenAI 同步上线“即时转向”（Instant Steering）功能，

仅供内部参考，请勿外传

允许用户在模型执行与生成过程中实时介入修正指令，模型无需等待当前批次完成即可当场调整执行路径，显著减少无效生成与等待损耗。

图表 7：GPT-6.1 Sol 极速版与标准版输出性能与响应延迟实测对比

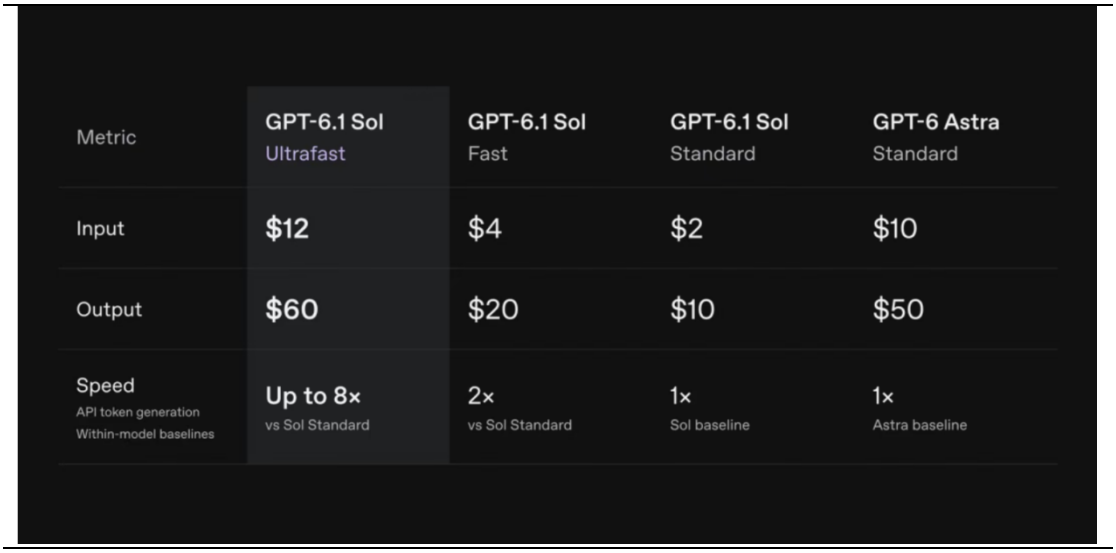


资料来源：新智元，华鑫证券研究

商业化定价与商业模式探索方面，GPT-6.1 Sol 极速版标志着大模型商业化由单纯的“能力智力阶梯”进一步衍生出以“推理速度”为维度的差异化价格带。具体资费方面，GPT-6.1 Sol 根据速度划分三档定价：标准版输入价格为每百万 Token 2 美元、输出价格为每百万 Token 10 美元；Fast 档速度提升一倍，输入价格为每百万 Token 4 美元、输出价格为每百万 Token 20 美元；Ultrafast 极速版每百万 Token 输入定价 12 美元、输出定价 60 美元（价格为标准版的 6 倍，但单价仅为旗舰 GPT-6 Astra 极速版的五分之一）。长文本计费方面，单次输入超过 27.2 万 Token 时，极速版资费上调至输入每百万 Token 24 美元、输出每百万 Token 90 美元。在接入与订阅权限上，API 端向全球所有支持地区全量开放，并支持美欧地区用户配置数据驻留，官方建议开发者通过 WebSocket 长连接调用以规避 Agent 多次工具交互中网络握手损耗；在端侧应用方面，Codex 与 ChatGPT Work 仅向每月 500 美元的 Pro 500、用量计费的企业版（须管理员手动开启）以及点数计费的教育版开放，Pro 500 套餐内运行极速版按标准额度的 8 倍扣除（额度耗尽后扣除点数按 6 倍计费），其余自助套餐均无法使用。实际任务成本测试显示，在 Codex 代码构建任务中，标准版 Sol 花费 1.02 美元，相同任务转换为极速版花费达 6.14 美元（略高于 Astra 标准模式的 5.96 美元）。速度分层计费机制精准切入了生产线应急运维、高频跨应用 Agent 交互及超低延迟实时体验等对响应时效高度敏感的企业级刚需场景。

仅供内部参考，请勿外传

图表 8：GPT-6.1 Sol 各档位与 GPT-6 Astra API 资费及生成速度对比



Metric	GPT-6.1 Sol Ultrafast	GPT-6.1 Sol Fast	GPT-6.1 Sol Standard	GPT-6 Astra Standard
Input	$12	$4	$2	$10
Output	$60	$20	$10	$50
Speed API token generation Within-model baselines	Up to 8× vs Sol Standard	2× vs Sol Standard	1× Sol baseline	1× Astra baseline

资料来源：新智元，华鑫证券研究

仅供内部参考，请勿外传

2、AI 应用动态：QuillBot 周时长环比+3.51%，Anthropic 发布 Claude Haiku 5.5

2.1、周流量跟踪：QuillBot 周时长环比+3.51%

本期（2026.10.3-2026.10.9）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1379.0M）、Bing（844.7M）和 Gemini（603.5M），访问量环比增速第一为文心一言（133.70%）；平均停留时长前三位分别为 Character.AI（00:15:02）、Discord（00:10:36）和 NotionAI（00:07:52）;平均停留时长环比增速第一为 QuillBot（3.51%）。

图表 9：2026.10.3-2026.10.9 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	1379.0	1.25%	6:39	-1.24%
Bing	搜索	微软	844.7	5.59%	7:05	0.00%
Gemini	聊天机器人	谷歌	603.5	-0.77%	6:24	-1.79%
Canva	在线设计	Canva	218.4	-0.50%	5:37	-1.46%
Github	代码托管	微软	159.7	5.27%	6:27	0.78%
Discord	游戏社区	微软	143.8	-1.44%	10:36	-1.40%
Character.AI	聊天机器人	Character.AI	33.89	-3.83%	15:02	-3.22%
DeepI	翻译工具	DeepL	21.86	-0.18%	2:19	-1.42%
Perplexity	AI 搜索	Perplexity	20.34	-2.63%	4:21	0.38%
QuillBot	释义工具	QuillBot	8.46	0.23%	2:57	3.51%
Kimi	聊天机器人	Moonshot AI	4.48	-16.83%	5:32	-4.60%
NotionAI	文本/笔记	Notion	2.701	-13.23%	7:52	-10.44%
文心一言	聊天机器人	百度	0.01	133.70%	0:00	-100.00%

资料来源：similarweb,华鑫证券研究

2.2、产业动态：Anthropic 发布 Claude Haiku 5.5，价格降低且性能大幅跃升

2026 年 10 月 8 日，Anthropic 正式推出模型 Claude Haiku 5.5，该模型被官方定位为迄今成本最低、响应速度最快且性能最强的小型人工智能系统。此次发布的核心亮点在于价格的大幅下调，其单位使用成本仅为前代 Haiku 4.5 的十分之一，同时也仅为 Sonnet 5.5 的二十分之一。Haiku 5.5 主要面向高容量、成本敏感型任务，能够高效处理摘要生成、文本压缩、数据库查询以及分类请求等重复性高、速度要求快的工作负载，并且能够作为

仅供内部参考，请勿外传

Opus 5.5 与 Sonnet 5.5 的辅助子代理，在编程等场景中发挥协同作用。

在价格结构方面，Haiku 5.5 的计费方式根据上下文长度有所区分。对于输入提示在十万 Token 以内的请求，每百万 Token 的输入价格为 0.10 美元，输出价格为 0.50 美元；一旦超过十万 Token 的阈值，输入与输出价格分别调整为 0.50 美元和 2.50 美元。缓存读取费用为每百万 Token 0.01 美元，若采用批处理模式还可享受五折优惠。官方测算显示，在同等任务量下，Haiku 5.5 的平均运行成本较 Haiku 4.5 降低约百分之七十五，这意味着相同的预算大约可以支持四倍的调用次数。与 GPT-6 Luna 相比，在十万 Token 以内两者的输入、输出及缓存价格完全一致，但在长上下文场景下 Luna 更具价格优势，其涨价阈值约为二十七万 Token，且涨价后的输入与输出价格分别为 0.20 美元和 0.75 美元。

图表 10：Claude Haiku 5.5 的定价

Price per 1 million tokens	Haiku 5.5 prompts up to / over 100k	Haiku 4.5	Sonnet 5.5
Cache reads	$0.01 / $0.05	$0.10	$0.10
Cache writes	$0.125 / $0.625	$1.25	$2.50
Input tokens	$0.10 / $0.50	$1.00	$2.00
Output tokens	$0.50 / $2.50	$5.00	$10.00

公众号 · 新智元

资料来源：新智元，华鑫证券研究

在性能表现上，Haiku 5.5 相较于前代实现了显著跃升。在操作电脑的 OSWorld 基准测试中，其得分从上一代的 15.7%大幅提升至 72.4%，与 Sonnet 5.5 的差距缩小至 11.5 个百分点。在终端编程的 Terminal-Bench 4.0 测试中，上一代产品未能得分，而 Haiku 5.5 则取得了 39.2%的成绩。在知识工作领域的 GDPval-AA 评估中，该模型获得 1620 分，达到上一代的两倍以上。在“人类最后的考试”HLE 不带工具的测试条件下，其得分从 10.2%上升至 45.9%。特别是在电脑操作方面，同价位的 Luna 在 OSWorld 上仅获得 48.9%，Haiku 5.5 领先其 23.5 个百分点；若考虑成本因素，Haiku 5.5 在 Medium 档位下的表现已超过 Luna 满档成绩，且花费更少。在 GDPval-AA 知识工作测试中，Haiku 5.5 的 1620 分高于 Luna 的 1437 分，但距离 Sonnet 5.5 的 1840 分仍有差距，其默认 Medium 档位得分为 1277 分。在长线项目评估 AA-Briefcase 中，该模型从上一代的 614 分提升至 1578 分，高于 Luna 的 1336 分。在推理与读图能力方面，HLE 结合搜索与代码工具时，Haiku 5.5 得分为 57.4%，Sonnet 5.5 为 64.5%；在专业图表解读测试 Chartography 中，其得分从上一代的 6.4%跃升至 46.4%，Luna 为 29.1%，Sonnet 5.5 为 61.6%。编程能力是 Haiku 5.5 与 Sonnet 差距最大的领域，Terminal-Bench 4.0 中其 39.2%的成绩是 Luna 16.4%的两倍多，但 Sonnet 5.5 达到了 70.6%；在 FrontierCode 测试中，Haiku 5.5 的 46.4%优于 Luna 的 42.4%，但落后于 Sonnet 5.5 最佳成绩 52.1%近六个百分点。官方也明确建议，复杂的 Agent 编程任务仍应交由 Sonnet 5.5 与 Opus 5.5 处理。

仅供内部参考，请勿外传

图表 11：Claude Haiku 5.5 在多个测试中的表现

	Haiku 5.5	Haiku 4.5	GPT-6 Luna	Sonnet 5.5 For reference
Knowledge work GDPval-AA v2.1	1620	735	1437	1840
Knowledge work AA-Briefcase v1.1	1578	614	1336	1824
Computer use OSWorld 2.1	72.4% Offline subset	15.7% Offline subset	48.9% Offline subset	83.9% Offline subset
Multidisciplinary reasoning Humanity's Last Exam	45.9% no tools	10.2% no tools	—	56.9% no tools
	57.4% with tools	18.7% with tools	—	64.5% with tools
Agentic coding Terminal-Bench 4.0	39.2%	0.0%	16.4%	70.6%
Agentic coding FrontierCode 1.1 (Main)	46.4%	—	42.4%	52.1% Xhigh
Visual reasoning Chartography	46.4% no tools	6.4% no tools	29.1% no tools	61.6% no tools

公众号 · 新智元

资料来源：新智元，华鑫证券研究

在应用场景方面，官方将 Haiku 5.5 定位为处理量大、要求快速且注重成本节约的任务，例如撰写摘要、压缩上下文、查询数据库以及执行分类操作。它也是目前速度最快的 Claude 模型，仅慢于开启快速模式的 Opus，非常适合实时客服与浏览器操作。另一个主要用途是作为 Opus 或 Sonnet 的子 Agent，由大模型负责拆解任务与决策，Haiku 负责并行执行。Cognition 的 Devin Fusion 让 Opus 5.5 担任主力、Haiku 5.5 作为副手，在 FrontierCode 上稳定保持 66.2 分的顶级水平，同时降低了成本与延迟。

在使用指南方面，Haiku 5.5 已在 Claude 客户端、Claude Code 以及 AWS、谷歌云、微软 Azure、Cursor、OpenRouter 等平台上线，模型 ID 为 claude-haiku-5-5，上下文窗口为 100 万 Token，单次最多可输出 12.8 万 Token。它是首个支持推理档位调节的 Haiku 模型，提供从 Low 到 Max 共五档选择，API 默认使用 Medium 档。Python 与 TypeScript SDK 也已加入电脑操作与浏览器操作功能，目前处于 Beta 阶段。从 Haiku 4.5 迁移时，相同文本会产生约三成额外的 Token，因此 max_tokens 与成本预估需要重新计算；budget_tokens 被自适应思考与 Effort 档位取代，temperature、top_p、top_k 参数需删除，预填回复功能不再可用，格式要求应改用结构化输出；电脑操作需更换为新工具集 computer_toolset_20250801，fastCityTier 暂不支持。若嫌迁移麻烦，可在 Claude Code 中运行命令/claude-api migrate this project to claude-haiku-5-5 让其自动完成调整。对于非开发者而言，现在即可直接在客户端中与 Haiku 5.5 对话。官方建议，简单问题无需再使用 Fable、Opus 或 Sonnet，Haiku 5.5 即使仅开启 Low 档推理强度，也能准确且迅速地给出答案，可完美替代搜索引擎。例如在查询离首都机场最近的酒店时，该模型通过网页搜索给出空港逸栈休息室酒店等选项，并提示信息来自携程等平台的历史页面，价格与营业状况可能变化，建议订房前再次确认。

仅供内部参考，请勿外传

图表 12：Claude Haiku 5.5 查询离首都机场最近的酒店示例



资料来源：新智元，华鑫证券研究

仅供内部参考，请勿外传

3、AI 融资动向：TypeSafe AI 完成 8.7 亿美元 A 轮融资，估值达 75 亿美元

2026 年 10 月 10 日，开发出非文本生成 AI 模型 Jev 的初创公司 TypeSafe AI 近日宣布完成 8.7 亿美元（约合 58.6 亿元人民币）的 A 轮融资，投后估值达到 75 亿美元（约合 505 亿元人民币），由 a16z 领投，红杉资本、DCVC 及多位天使投资人参投，a16z 合伙人马丁·卡萨多（Martin Casado）将加入 TypeSafe 董事会。团队背景方面，TypeSafe AI 由前 OpenAI 研究员迪奥戈·阿尔梅达（Diogo Almeida）、前 Meta 研究工程师盛莎莎（Sasha Sheng）以及工程师兼创业者埃里克·加夫尼（Erik Gafni）于 2024 年共同创立。业务增长与落地速度方面，旗舰模型 Jev 于 9 月 15 日发布后数周内推动公司估值暴涨 37 倍；a16z 披露 Jev 发布仅 3 天生成量即达 1 万亿 token，财富 500 强企业中已有约三分之一（a16z 博客称 25%）接入使用；开发者社区第一周涌现出涵盖生成式 UI、互动游戏、数据分析等数千个落地用例。

产品定位与核心技术架构方面，Jev 同样基于谷歌 2017 年提出的 Transformer 神经网络架构，但不属于输出文本的大语言模型，核心定位于首个专为快速做出结构化决策而设计的“System One”模型。针对传统 LLM 强行输出文本再由代码解析为结构化决策的流程错配问题，用户向 Jev 传入上下文状态与类型化问题后，模型不生成文本亦无需解析，直接返回供代码使用的类型化值与概率分布，用于驱动分支判断、排序或路由。Jev 使用专为其设计的“校准决策强化学习”（RLCD）进行训练，并提供三种 AI 原语（Choice、Score、Noul），支持单次请求中针对同一状态并行评估多个问题且几乎不增加响应时间。性能与成本开销方面，a16z 数据显示 Jev 在分类任务上的速度为前沿模型的 100 倍，成本约为前沿模型的 1/100 至 1/500，精度相当；TypeSafe 官方数据称其在分类任务上的推理速度最高可达同类 LLM 的 200 倍，成本低至 1/400。在应用场景上，Jev 可配合 LangChain 等智能体框架使用，直接嵌入“模型决策-工具执行-评估反馈”的智能体循环，替代高频聊天 LLM 调用，推动 AI 以软件原生语言深度融入代码。

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
TypeSafe AI	AI 模型开发公司	Andreessen Horowitz	A 轮	8.7 亿美元	暂未披露	75 亿美元
Arena	AI 评测	Lightspeed & Khosla Ventures	B 轮	2 亿美元	暂未披露	31 亿美元

仅供内部参考，请勿外传

Nous Research	开源 AI 智能体与去中心化算力基础设施	Robot Ventures	B 轮	0.9 亿美元	1.58 亿美元	15 亿美元

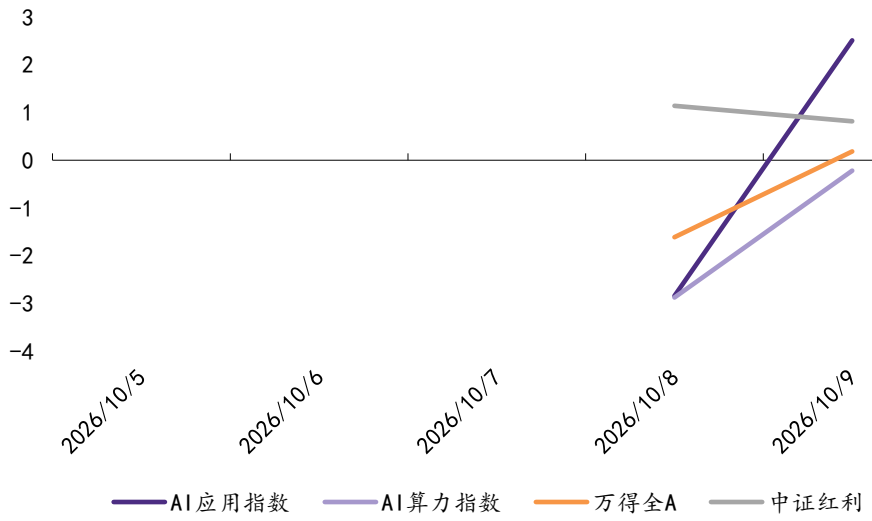
资料来源：Saasverse，智东西，华鑫证券研究

仅供内部参考，请勿外传

4、 行情复盘

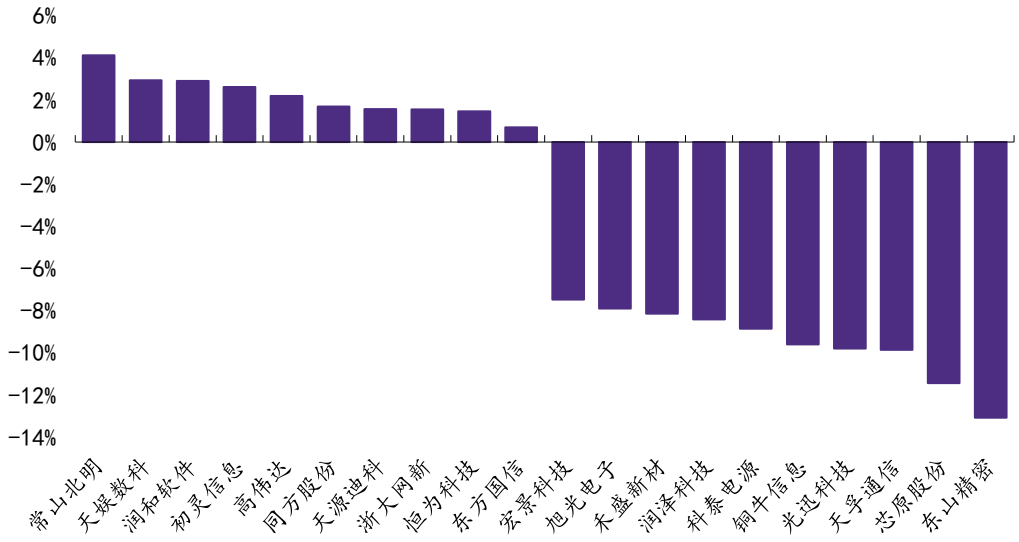
上周（2026.10.5-2026.10.9 日），AI 应用指数/万得全 A/中证红利日涨幅最大值分别为 2.51%/0.18%/1.14%，AI 应用指数/AI 算力指数/万得全 A 日跌幅最大值分别为-2.84%/-2.88%/-1.61%。AI 算力指数内部，常山北明以 4.11%录得上周最大涨幅，东山精密以-13.09%录得上周最大跌幅。AI 应用指数内部，华策影视以[illegible]%录得上周最大涨幅，美迪凯以-13.55%录得上周最大跌幅。

图表 14：上周（2026.10.5-2026.10.9 日）指数日涨跌幅



资料来源：wind，华鑫证券研究

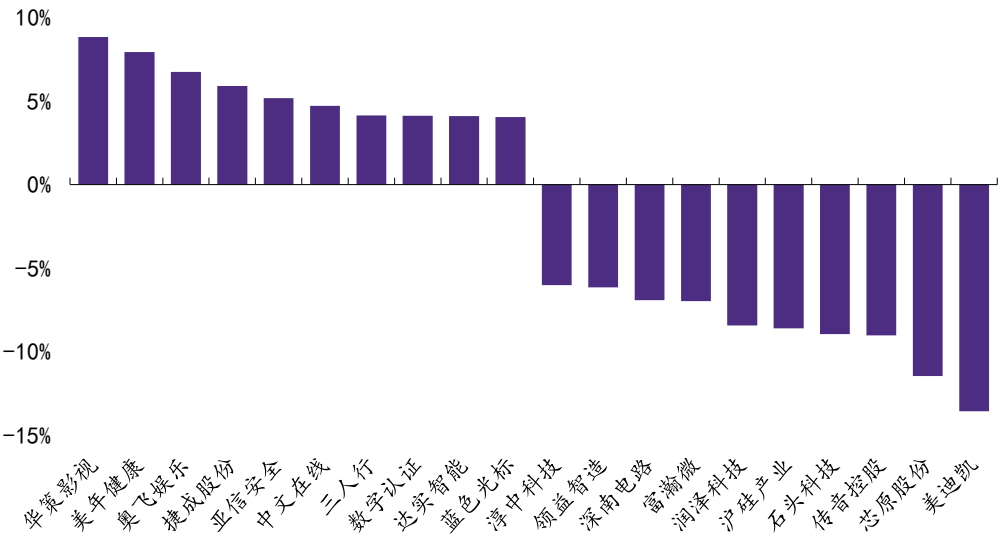
图表 15：上周（2026.10.5-2026.10.9 日）AI 算力指数内部涨跌幅度排名



资料来源：wind，华鑫证券研究

仅供内部参考，请勿外传

图表 16：上周（2026.10.5-2026.10.9 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

仅供内部参考，请勿外传

5、 投资建议

10 月 6 日，Sierra 与 Meta 联合发布个人智能体协议（PersonalAgentProtocol，代号 Poppy），首批合作伙伴包括 Genesys、Instinct、Rocket、Shopify、Stripe 与 Walmart。该协议为开放标准，定义个人智能体代表用户与企业交互时的身份认证、权限与会话管理方式，基于 OAuth 等现有标准构建；企业可据此识别来访的智能体及其代表的用户，并选择通过网页、API 或自有智能体承接任务。10 月 9 日，Sierra 公布协议草案并新增 35 家设计伙伴，包括 OpenAI、Visa、Mastercard、PayPal、美国银行、Cloudflare、Target、美联航等，覆盖模型厂商、支付网络、银行、零售与航空多个环节；支付、消息推送等能力暂列为后续扩展方向。

本次协议发布的核心信号在于，智能体竞争已经延伸至交互标准的制定。先前公司 Muse 智能体的爆发式增长验证了消费级智能体的需求，而当智能体开始代表用户完成购物、改签、售后等真实交易，智能体与企业系统之间缺乏统一的交互规则就成为新的瓶颈，规则的制定权将成为入口之争以外的第二层竞争。从参与方结构看，发起方与新增伙伴覆盖模型厂商、支付网络、银行与零售龙头，头部模型 OpenAI 亦以设计伙伴身份加入，交易链条上的关键角色已齐聚同一协议框架。我们认为，个人智能体协议的落地标志着智能体经济开始完善规则层基础设施，智能体与支付、电商、客服系统的标准化对接将成为 AI 应用落地的下一条主线。

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

仅供内部参考，请勿外传

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 18：重点关注公司及盈利预测

公司代码	名称	2026-10-13	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	481.60	-0.40	2.13	4.49	-1204.00	226.10	107.26	买入
301196.SZ	唯科科技	75.38	2.05	1.92	2.39	39.28	39.23	31.50	买入
301369.SZ	联动科技	111.17	0.48	1.08	2.62	231.60	102.94	42.43	买入
603859.SH	能科科技	36.57	0.92	1.05	1.32	39.75	34.83	27.70	买入
688615.SH	合合信息	84.17	3.24	4.22	5.25	25.98	19.95	16.03	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。

仅供内部参考，请勿外传

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。

仅供内部参考，请勿外传

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。