

2026 年 10 月 19 日



华鑫证券
CHINA FORTUNE SECURITIES

Gemini 4 Argon 发布，DeepSeek 更新 V4.1-Flash-1017

—计算机行业周报

推荐(维持)

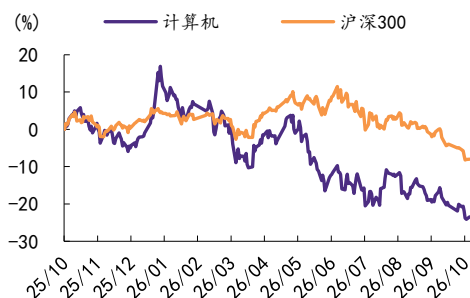
分析师：任春阳 S1050521110006

rency@cfsc.com.cn

行业相对表现

表现	1M	3M	12M
计算机(申万)	-6.9	-3.7	-23.0
沪深 300	-7.4	-7.8	-7.5

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-6.1Sol 上线，Anthropic 发布 ClaudeHaiku5.5》2026-10-14
- 2、《计算机行业周报：GPT-6Sol 与 Luna 推出，Anthropic 发布 ClaudeOpus5.5》2026-09-30
- 3、《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

投资要点

■ 算力：阿里 Qoder 上线 Fast 模式提速降本，DeepSeek 更新 V4.1-Flash-1017 版本开放搜索 API

阿里智能体平台 Qoder 正式上线全新 Fast 模式通过模型智能路由策略、Agent Harness 层系统性重构及网络延迟压降，实现全链路生成速度提升超 150%、Token 消耗降低约 27%，端侧响应时间由 8 秒大幅压缩至 3 秒。同期，DeepSeek 正式发布 V4.1-Flash-1017 版本，新模型在软件工程 DeepSWE 等四项核心基准测试中刷新性能-费用帕累托最优；同时通过自反馈强化学习增强了“许愿式”自然语言编程能力，支持从零构建完整项目，并向公众正式开放定价均为 5 元/千次的 Search 与 Fetch 自研搜索 API 接口。

■ AI 应用：Github 周访问量环比+6.26%，Gemini 4 Argon 开放全球使用

2026 年 10 月 14 日，谷歌正式发布新一代模型 Gemini 4 Argon，标志着 Gemini 4 时代的开启。该模型即日起向全球开放：Pro 用户可在 Gemini 应用与搜索 AI 模式中体验，开发者可通过付费 API 在 AI Studio、Android Studio 中调用，Ultra 用户还可额外在 AI Studio 和 Antigravity 中开展工作。

■ AI 融资动向：无问芯穹完成 23 亿元 C 轮融资，异构算力与 Token 工厂卡位基础设施核心枢纽

国内 AI Infra 企业无问芯穹完成 23 亿元 C 轮融资，累计融资达 43 亿元创行业新高并启动 IPO 筹备，投资方涵盖社保中关村基金、腾讯、百度、红杉中国等国资、产业与财务资本。公司脱胎于清华电子系，深耕“M 模型×N 芯片”异构调度，全网部署触达超 37000P 算力并覆盖 16 种主流芯片；旗下“前店后厂一中心”布局跑通超 4000 公里跨域混训，日均 Token 调用量增长 40 倍且推理成本压降 10 倍，以中立轻资产模式卡位 Token 经济核心生产力枢纽。

■ 投资建议

2026 年 10 月 12 至 15 日，2026 年 OCP 全球峰会在圣何塞举行，CPO 的产业化进展成为本届峰会焦点。英伟达确认，下一代 RubinUltraNVL576 的 scale-up 互联将采用 CPO 方案。博通现场展示第三代 TH6-DavissonCPO 方案，Celestica、台达、

Micas 等合作伙伴演示了基于该方案的 102.4TCPO 整机系统，从交换芯片到整机集成的产业链环节基本打通。可插拔路线同步推进，英伟达与思科联合提出面向单通道 400G 的 USFP 光模块新形态。机架与设施层面，英伟达向 OCP 贡献 VeraRubin 机架与计算托架设计，MGX1.2 规范支持最高 250kW 全液冷机架，并联合 AMD、谷歌、Meta 发布开放数据中心规范 v0.7；AMD 发布智能体机架 Helion，并披露 Helios 首批系统已在 OpenAI、Meta、微软与甲骨文安装运行；Meta 将采用 Spectrum-6 交换机建设 VeraRubinAI 工厂。

本届峰会的核心信号在于，CPO 的应用边界正从交换网络扩展到算力机架内部。此前 CPO 的落地集中在交换机侧的 scale-out 场景，RubinUltraNVL576 的 scale-up 互联采用 CPO，意味着光引擎将进一步贴近计算芯片，scale-up 这个时延与功耗最敏感、此前由铜互联主导的场景开始向光互联打开。供给端，博通第三代方案与多家整机伙伴的 102.4T 系统演示表明，当前 CPO 已从单点技术验证进入芯片、封装、整机协同推进的生态化阶段。需求端的验证窗口也将于下周开启，各大 CSP 将陆续发布新一季财报，Meta、微软、Alphabet 与亚马逊分别定于 10 月 21 日、28 日与 29 日披露业绩，其资本开支指引与资金来源将成为检验 AI 算力需求持续性的直接依据。

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	2026-10-19 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	440.38	-0.40	2.13	4.49	-1100.95	206.75	98.08	买入
301196.SZ	唯科科技	81.85	2.05	1.92	2.39	39.28	42.60	34.20	买入
301369.SZ	联动科技	95.01	0.48	1.08	2.62	197.94	87.97	36.26	买入
603859.SH	能科科技	38.82	0.92	1.05	1.32	42.20	36.97	29.41	买入
688615.SH	合合信息	84.15	3.24	4.22	5.25	25.97	19.94	16.03	买入

资料来源：Wind，华鑫证券研究

正文目录

1、 算力动态：阿里 QODER 上线 FAST 模式提速降本，DEEPSEEK 更新 V4.1-FLASH-1017 版本开放搜索 API .	5
1.1、 Tokens 跟踪.....	5
1.2、 算力租赁动态：新一代 GPU 租金呈持续走高态势	7
1.3、 云厂动态：阿里智能体平台 Qoder 上线 Fast 模式，全链路优化提速超 150%.....	9
1.4、 产业动态：DeepSeek 发布 V4.1-Flash-1017 版本，代码工程刷新帕累托最优且开放自研搜索 API	9
2、 AI 应用动态：GITHUB 周访问量环比+6.26%，GEMINI 4 ARGON 开放全球使用	12
2.1、 周流量跟踪：Github 周访问量环比+6.26%.....	12
2.2、 产业动态：Gemini 4 Argon 开放全球使用，百万 token 窗口贯通推理与执行	12
3、 AI 融资动向：无问芯穹完成 23 亿元 C 轮融资，异构算力与 TOKEN 工厂卡位基础设施核心枢纽.....	15
4、 行情复盘	17
5、 投资建议	19
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）	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 3：2026.10.12 -2026.10.18 日 H100 价格（单位：美元/小时）	7
图表 4：2026.10.12 -2026.10.18 日 H200 价格（单位：美元/小时）	8
图表 5：2026.10.12 -2026.10.18 日 A100 价格（单位：美元/小时）	8
图表 6：2026.10.12 -2026.10.18 日日 B200 价格（单位：美元/小时）	8
图表 7：2026.10.12 -2026.10.18 日日 B300 价格（单位：美元/小时）	9
图表 8：DEEPSEEK-V4.1-FLASH-1017 与主流前沿模型多维度 BENCHMARK 评测表现对比	10
图表 9：2026.10.10-2026.10.16 AI 相关网站流量	12
图表 10：GEMINI 4 ARGON 在多个测试中的表现.....	13
图表 11：GEMINI 4 ARGON 在 ARTIFICIAL ANALYSIS 智力指数中的表现.....	13
图表 12：GEMINI 4 ARGON 优化开源 AGENT 框架 ADK	14
图表 13：上周 AI 初创公司融资动态	16
图表 14：上周（2026.10.12-2026.10.16 日）指数日涨跌幅	17
图表 15：上周（2026.10.12-2026.10.16 日）AI 算力指数内部涨跌幅度排名	17
图表 16：上周（2026.10.12-2026.10.16 日）AI 应用指数内部涨跌幅度排名	18

图表 17: FICONTEC2025 年年中至今公告订单..... 19

图表 18: 重点关注公司及盈利预测 20

1、算力动态：阿里 Qoder 上线 Fast 模式提速降本，DeepSeek 更新 V4.1-Flash-1017 版本开放搜索 API

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 10 月 12 日至 10 月 18 日，周度 Token 消耗量小幅回落，消耗量达 164T，环比上周下降 4.64%。在 Tokens 规模 Leaderboard 前五名中，DeepSeek 的 DeepSeek V4.1 Flash 以 32.83T tokens 稳居榜首；Z-ai 的 GLM 5.3 Flash 以 12.23T tokens 位列第二；StepFun 的 Step 5 Preview 以 11.47T tokens 位列第三；Xiaomi 的 MiMo-V2.6-Flash 以 9.94T tokens 位列第四；Tencent 的 Hy4 preview 以 7.55T tokens 位列第五。

文本类模型的市场份额方面，deepseek 占据 24.4% 的份额，位列第一；google 占据 19.7% 的份额，位列第二；openai、z-ai、qwen 分别占据 18.5%、12.2%、5.3% 的市场份额。

图像生成类模型调用量稳步增长，调用量为 8.23M requests，环比上周增加 2.49%。在 Image Leaderboard 前五名中，google 的 Nano Banana 2 (Gemini 3.1 Flash Image) 以 1.37M requests 位居第一；OpenAI 的 GPT Image 2.5 Sunburst 以 1.13M requests 位列第二；Google 的 Nano Banana 2 Lite (Gemini 3.1 Flash Lite Image) 以 658K requests 位列第三；Meta 的 Muse Image 1.0 以 646K requests 位列第四；OpenAI 的 GPT Image 2.5 Flare 以 632K requests 位列第五。从市场份额来看，google 以 3.73M requests 占据 45.3% 的份额，位列第一；openai 以 2.72M requests 占据 33.1%，位列第二；meta、bytedance-seed、black-forest-labs 分别以 646K、328K、233K requests，对应占据 7.8%、4.0%、2.8% 的市场份额。

嵌入模型调用量稳步增长，调用量为 404M requests，环比上周增加 9.19%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 157M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 61.9M requests 位列第二；Google 的 Gemini Embedding 2 以 43.3M requests 位列第三；Google 的 Gemini Embedding 001 以 27M requests 位列第四；BAAI 的 bge-m3 以 25.1M requests 位列第五。从市场份额来看，openai 以 182M requests 占据 44.9% 的份额，位居第一；google 以 76.6M requests 占据 18.7% 的份额，位列第二；qwen、sentence-transformers、baai 分别以 69.6M、26M、26M requests，对应占据 17.2%、6.4%、6.4% 的市场份额。

重排序模型调用量小幅增长，调用量为 12.7M requests，环比上周增加 3.25%。在 Rerank Leaderboard 前五名中，VoyageAI 的 rerank-2.5 以 4.68M requests 位居榜首；VoyageAI 的 rerank-2.5-lite 以 3.26M requests 位列第二；Cohere 的 Rerank v3.5 以 1.72M requests 位列第三；Qwen 的 Qwen3 Reranker 8B 以 970K requests 位列第四；VoyageAI 的 rerank-3 以 540K requests 位列第五。从市场份额来看，voyageai 以 8.94M requests 占据 70.7% 的份额，位居第一；cohere 以 2.66M requests 占据 21.0% 的份额，位列第二；qwen、nvidia 分别以 970K、83K requests，对应占据 7.7%、0.7% 的市场份额。

视频生成类模型调用量小幅增长，调用量为 368K requests，环比上周增加 9.52%。在

Video Leaderboard 前五名中，HeyGen 的 HeyGen Video 以 68K requests 位居榜首；Google 的 Veo 3.1 Lite 以 52K requests 位列第二；ByteDance 的 Seedance 2.5 以 36K requests 位列第三；Wan 的 Wan 3.0 以 23K requests 位列第四；xAI 的 Grok Imagine Video 1.5 Lite 以 20K requests 位列第五。从市场份额来看，bytedance 以 95K requests 占据 25.7% 的份额，位列第一；heygen 以 81K requests 占据 22.1% 的份额，位居第二；google、space-x-ai、alibaba 分别以 77K、36K、32K requests，对应占据 20.8%、9.8%、8.8% 的市场份额。

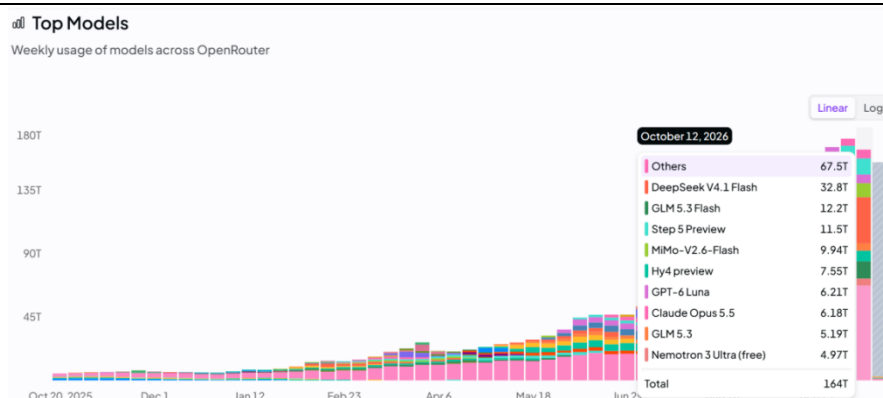
语音生成类模型调用量大幅增长，调用量为 8.39M requests，环比上周增加 70.53%。在 Speech Leaderboard 前五名中，Google 的 Gemini 3.8 Flash TTS 以 1.61M requests 位居榜首；Elevenlabs 的 Eleven v4 Turbo 以 1.27M requests 位列第二；Hexgrad 的 Kokoro 82M 以 828K requests 位列第三；Elevenlabs 的 Eleven v4 以 767K requests 位列第四；Google 的 Gemini 3.8 Flash Lite TTS 以 620K requests 位列第五。从市场份额来看，google 以 2.74M requests 占据 32.6% 的份额，位列第一；elevenlabs 以 2.61M requests 占据 31.1% 的份额，位居第二；fish-audio、hexgrad、space-x-ai 分别以 909K、828K、525K requests，对应占据 10.8%、9.9%、6.3% 的市场份额。

语音转写类模型调用量小幅增长，调用量为 29M requests，环比上周增加 3.20%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3 以 9.53M requests 位居榜首；OpenAI 的 Whisper Large V3 Turbo 以 7.05M requests 位列第二；Microsoft 的 MAI-Transcribe 2 以 3.93M requests 位列第三；OpenAI 的 GPT-4o Mini Transcribe 以 854K requests 位列第四；OpenAI 的 GPT Transcribe 以 674K requests 位列第五。从市场份额来看，openai 以 18.9M requests 占据 65.2% 的份额，位居第一；microsoft 以 3.97M requests 占据 13.7% 的份额，位列第二；elevenlabs、qwen、mistralai 分别以 2.12M、911K、654K requests，对应占据 7.3%、3.1%、2.3% 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。

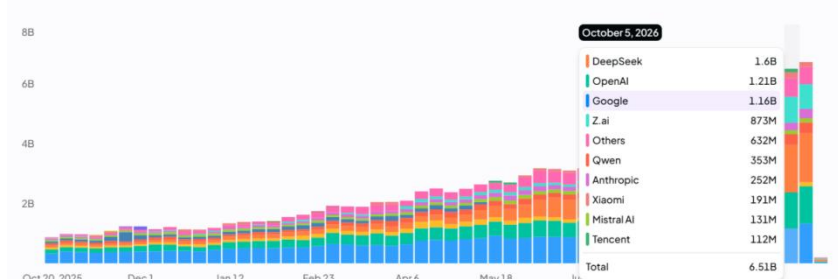
10 月 11 日，润建股份牵头组建的“广西词元经济及智能应用概念验证中心”成功入选首批广西科技成果转化基地。该中心联合广西大学、东盟技术转移中心等多家单位共建，立足区位优势，依托五象云谷词元工厂智算底座与 Token 服务，聚焦解决 AI 应用验证成本高、转化链条长与跨境落地难等痛点，打造面向东盟的词元与智能应用跨境流通枢纽，开辟中国—东盟数字经济合作新通道。

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

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

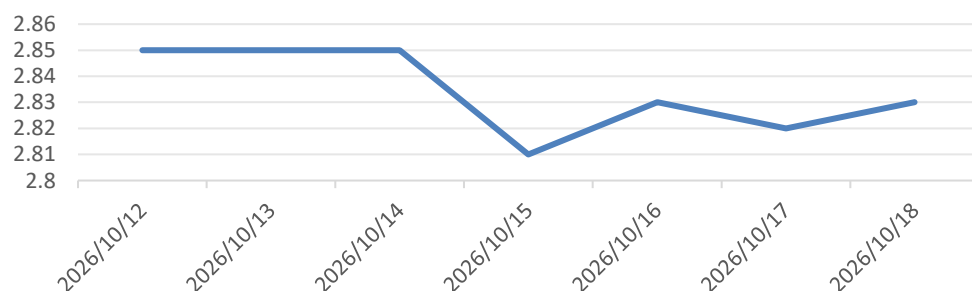
1.2、算力租赁动态：新一代 GPU 租金呈持续走高态势

本周（2026.10.12-2026.10.18 日），算力租赁市场整体成熟卡型相对稳定，新一代高算力 GPU 租金的持续走高。H100 指数/H200 指数/A100 指数/B200 指数/B300 指数日租金（美元/小时）最大值分别为 2.85/3.51/1.53/6.26/7.01，H100 指数/H200 指数/A100 指数/B200 指数/B300 指数日租金（美元/小时）最小值分别为 2.81/3.35/1.51/5.94/6.65。

对比上周（10.5-10.11 日）数据观察，不同代际卡型呈现分化走势：成熟主力卡型中，H100 与 A100 日租金极值与上周基本持平，整体运行平稳；H200 租金出现小幅上扬（最大值由上周的 3.44 美元/小时上升至 3.51 美元/小时，最小值由 3.29 美元/小时提升至 3.35 美元/小时）；而 Blackwell 架构新卡型租金上涨趋势显著，B200 租金由周初的 5.94 美元/小时逐日上行至周末的 6.26 美元/小时，突破上周 5.93 美元/小时的上限水平；本周新纳入追踪的 B300 租金亦呈现阶梯式走高，由周初的 6.65 美元/小时逐步攀升至周末的 7.01 美元/小时。

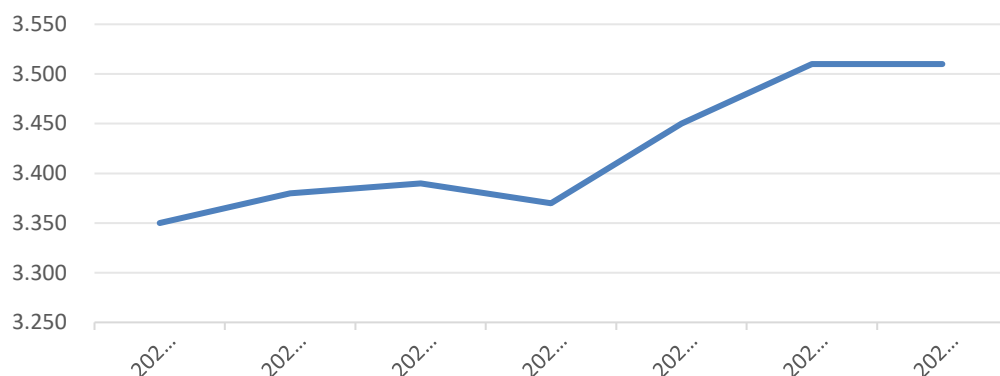
* 计算方法说明：以上数据均来源于 silicon data。每条 GPU 租赁记录首先会被转换为标准化的每 GPU 每小时费率——即单价等于总租赁价格除以（GPU 数量乘以租赁时长）。预处理后的数据会经过多阶段过滤，以剔除上报错误，同时不会忽略真实的结构性价变动。由于单纯汇总价格会随每日配置组合的变化而波动，因此我们采用基差调整（basis adjustment）将每条记录标准化为基准等价合约，在此过程中会综合考虑租赁类型、地理位置、CPU 平台以及 GPU 型号等因素。经基差调整后的价格会在提供商层面进行汇总，最终的指数值是各提供商数据的加权平均值。

图表 3：2026.10.12 -2026.10.18 日 H100 价格（单位：美元/小时）



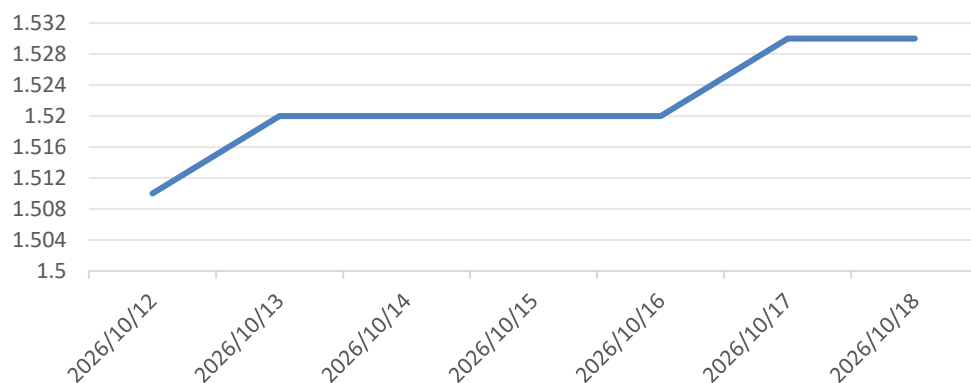
资料来源：silicon data，华鑫证券研究

图表 4：2026.10.12 -2026.10.18 日 H200 价格（单位：美元/小时）



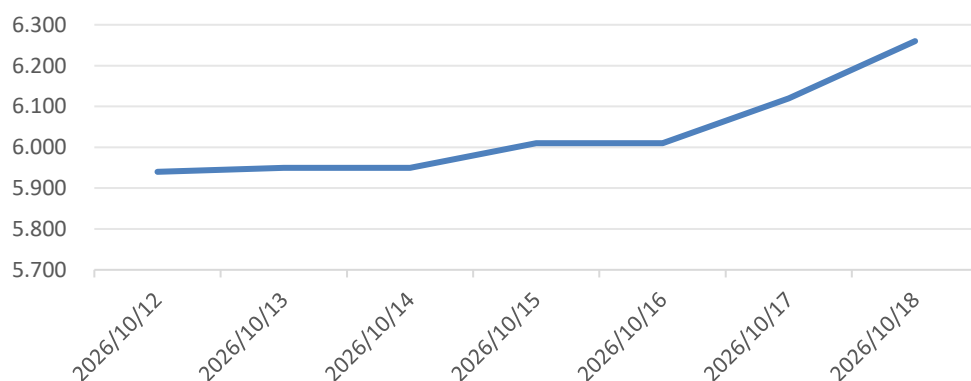
资料来源：silicon data，华鑫证券研究

图表 5：2026.10.12 -2026.10.18 日 A100 价格（单位：美元/小时）



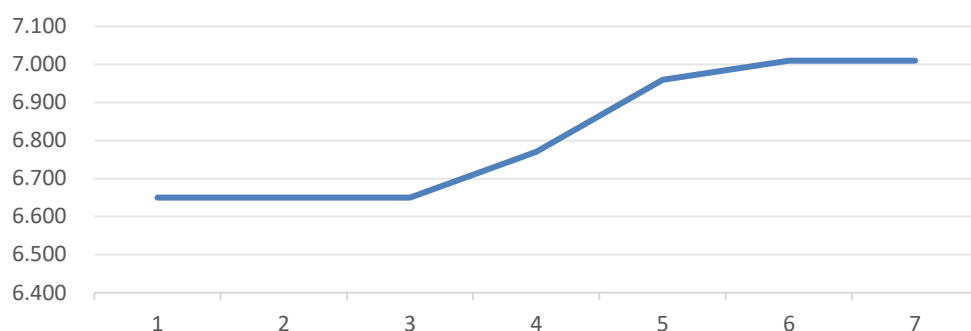
资料来源：silicon data，华鑫证券研究

图表 6：2026.10.12 -2026.10.18 日日 B200 价格（单位：美元/小时）



资料来源：silicon data，华鑫证券研究

图表 7：2026.10.12 -2026.10.18 日日 B300 价格（单位：美元/小时）



资料来源：silicon data，华鑫证券研究

1.3、云厂动态：阿里智能体平台 Qoder 上线 Fast 模式，全链路优化提速超 150%

2026 年 10 月 12 日，阿里智能体平台 Qoder 正式上线全新 Fast 模式，所有用户即日起可在 Qoder 中国版和国际版同步体验。该模式相比国际版原有性能模式，生成速度提升超 150%，Token 消耗降低约 27%，在显著提升生成速度的同时大幅压降用户使用成本。在应用定位上，针对写日报、修 BUG、代码补充等高频日常任务无需长时间深度思考而更依赖即时响应与稳定输出的特点，Qoder 团队确立了分档处理策略，日常高频任务交由 Fast 模式承接以保障生产力流转，少量复杂深度问题则切换至更强档位处理。自 2025 年 8 月面向全球发布以来，Qoder 产品家族已延伸覆盖 Qoder、Qoder IDE、Qoder CLI、Qoder JetBrains 插件以及数字员工产品 QoderWake 等矩阵，累计服务全球超 800 万用户及超 10 万家企业客户。

技术架构与底层机制方面，Fast 模式从模型、Agent Harness 与网络三个层面实现了全链路系统级优化：模型层面，平台接入新一代更高速模型并引入智能路由策略，能够根据任务类型与复杂度自动匹配最优模型，带动推理速度提升 12%；Agent Harness 层面，针对上下文获取、工具调用及任务规划等关键环节实施系统性重构，整体运行效率提升 28%；网络层面，全链路压缩网络等待时间，端到端延迟降低 15%。

实测表现与交互体验方面，测试数据显示 Fast 模式的端侧响应时间由原有性能模式的 8 秒大幅压降至 3 秒；在 10 个常见任务的横向实测中，Fast 模式完成回答的整体耗时仅为同类产品的三分之一。在具体高频场景测试中，Fast 模式生成一份旅游攻略仅用时 1 秒，基于当日工作日报输出次日规划用时 4 秒，完成单次头脑风暴快问快答仅耗时 1 秒，整体交互体感接近即时响应，进一步强化了端侧轻量任务下的工程交付效率。

1.4、产业动态：DeepSeek 发布 V4.1-Flash-1017 版本，代码工程刷新帕累托最优且开放自研搜索 API

2026 年 10 月 17 日，DeepSeek 正式将 V4.1 Flash 模型更新至 DeepSeek-V4.1-Flash-1017 版本，官方网页端、App 与 API 已同步更新，模型权重亦在 HuggingFace 与 ModelScope 开源开放下载。此次迭代距离 9 月 10 日发布全面超越 DeepSeek-V4-Pro 的 V4.1-Flash 仅过

去一个多月；在版本发布前的 10 月 15 日，DeepSeek 曾发生近 5 小时 API 服务中断，官方随后公布补偿方案，全额以赠金形式返还用户 10 月 8 日至 14 日的消耗金额。架构与推理性能方面，新版本聚焦任务执行效率与完整项目开发能力，在提升测试成绩的同时显著缩短了任务运行时间与输出长度；海外开发者实测显示其运算速度为 244 个 Token，运算速率超越 GPT-6.1 Sol 的超快模式，同时具备与 Sol 相当的智能水平，部分开发者将其定位为“V4.2-Pro”。平台适配与生态支持方面，DeepSeek Harness 同步升级至 0.2.1 版本（官方建议升级至 0.2.1-rc.1 及以上），并对腾讯 CodeBuddy 与 WorkBuddy 进行了框架优化；符合条件的用户首次登录 DeepSeek Harness 可获赠 6 元限量体验金（7 天内有效）。

基准测试与“许愿式”项目构建方面，新模型在软件工程 DeepSWE、业务自动化 Automation-Bench、复杂任务执行 Agents' Last Exam 以及专业知识工作 Apex Bench 四项评测中展现出明显优势，跑分领先 Kimi K3、GLM 5.3、GPT-5.6 Sol 与 Claude Opus 5 等前沿模型。在考察软件工程能力的 DeepSWE v1.1 测试中，新模型 max 档位单次任务通过率达 77.0%，超越 GPT-6 Astra、GPT-6.1 Sol、Gemini 3.8 Flash 以及 Claude Opus 5，较上一版本提升 2.8 个百分点；单项任务平均费用则由约 3.3 元压降约六成至约 1.3 元，在多个 Harness 下达成了全新的性能-费用帕累托最优（以更低成本取得更高任务表现）。在编程开发能力上，新模型通过基于自反馈的强化学习（RL）训练，强化了从零构建完整项目的的能力，支持以自然语言描述需求的“许愿式”编程，仅需较少提示词与参考素材即可以极少交互轮次完成游戏开发与 3D 建模等复杂任务。

图表 8：DeepSeek-V4.1-Flash-1017 与主流前沿模型多维度 Benchmark 评测表现对比

Benchmarks	DeepSeek V4.1-Flash-1017	DeepSeek V4.1-Flash	Kimi K3	GLM 5.3	GPT 5.6-Sol	Claude Opus 5
HLE (w/o tools)	42.6/44.0*	36.8/39.1*	43.5	42.0*	44.5	56.3
Terminal-Bench 4.0	42.3	31.2	12.6	37.9	39.9	51.8
DeepSWE v1.1	77.0	74.2	67.5	66.9	73.0	74.0
ProgramBench	29.1	20.3	17.5	19.0	23.0	37.0
NL2Repo-Bench	68.2	65.4	58.0	58.0	56.8	75.3
Automation-Bench V1.0.6	60.5	54.8	46.7	48.8	45.8	50.3
Agents' Last Exam	35.4	31.8	27.6	28.5	26.7	28.6
Apex Bench v1.1	72.1	64.6	47.8	56.6	54.3	65.8
Chartography (w/tools)	83.0	78.9	68.1	-	79.9	84.0
BabyVision (w/tools)	92.5	89.6	85.7	-	88.9	94.1
ZeroBench-main (w/tools)	51.0	49.0	41.0	-	53.0	52.0

*代表仅测试 HLE 的文本子集；我们在对 Terminal-Bench 4.0 进行的测试中使用 ClaudeCode 框架，DeepSWE 使用 Mini-SWE-V2 框架；ZeroBench-main 使用 Pass@5 评测。

资料来源：智东西，华鑫证券研究

搜索商业化与接口开放方面，DeepSeek 正式向公众开放自研搜索引擎的 Search（搜索）与 Fetch（抓取）两项 API 接口。该搜索引擎已索引全球千亿级网页，日处理数亿次搜索请求，支持多语言搜索并针对 SEO 作弊、GEO 操纵及低质 AI 生成内容进行了深度过滤。商业化资费上，Search 与 Fetch 接口定价均为 5 元/千次；为降低生态接入门槛，目前用户在 DeepSeek Harness、Claude Code 以及 Codex 中调用 DeepSeek API 时，通过服务端工具访问的 Search 与 Fetch 接口暂未计费，免费调用期将持续至北京时间 11 月 1 日 00:00，届时将正式转为按次收费。

2、AI 应用动态：Github 周访问量环比 +6.26%，Gemini 4 Argon 开放全球使用

2.1、周流量跟踪：Github 周访问量环比+6.26%

本期（2026.10.10-2026.10.16）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1384.0M）、Bing（857.5M）和 Gemini（607.0M），访问量环比增速第一为 Github（6.26%）；平均停留时长前三位分别为 Character.AI（00:14:39）、Discord（00:10:35）和 NotionAI（00:07:43）；平均停留时长环比增速第一为 QuillBot（2.26%）。

图表 9：2026.10.10-2026.10.16 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	1384.0	0.36%	6:36	-0.75%
Bing	搜索	微软	857.5	1.52%	7:04	-0.24%
Gemini	聊天机器人	谷歌	607.0	0.58%	6:21	-0.78%
Canva	在线设计	Canva	217.1	-0.60%	5:35	-0.59%
Github	代码托管	微软	169.7	6.26%	6:26	-0.26%
Discord	游戏社区	微软	147.3	2.43%	10:35	-0.16%
Character.AI	聊天机器人	Character.AI	35.02	3.33%	14:39	-2.55%
DeepL	翻译工具	DeepL	22.07	0.96%	2:16	-2.16%
Perplexity	AI 搜索	Perplexity	20.21	-0.64%	4:20	-0.38%
QuillBot	释义工具	QuillBot	8.23	-2.70%	3:01	2.26%
Kimi	聊天机器人	Moonshot AI	4.63	3.42%	5:35	0.90%
NotionAI	文本/笔记	Notion	1.685	-37.62%	7:43	-1.91%
文心一言	聊天机器人	百度	0.003166	-37.97%	0:00	

资料来源：similarweb, 华鑫证券研究

2.2、产业动态：Gemini 4 Argon 开放全球使用，百万 token 窗口贯通推理与执行

2026 年 10 月 14 日，谷歌正式发布新一代模型 Gemini 4 Argon，标志着 Gemini 4 时代的开启。该模型即日起向全球开放：Pro 用户可在 Gemini 应用与搜索中体验，开发者可通过付费 API 在 AI Studio、Android Studio 中调用，Ultra 用户还可额外在 AI Studio 和 Antigravity 中开展工作。

Gemini 4 Argon 将图像、视频与语言理解能力同推理、代码执行能力相结合，能够在理解任务后调用工具、执行代码并持续推进任务。在官方公开的 19 项基准对比中，最高思考档的 Argon 拿下 13 项单独最高、1 项并列最高，实力对标 GPT-6 Astra 与 Table 5.1；第三方 Artificial Analysis 智力指数中，Argon (high) 得 53 分，与 GPT-6 Astra (max) 同分，并高出 GPT-6.1 Sol (max) 1 分。谷歌还展示了其在真实工程、科学与机器人难题上的成果：Argon 可连续工作 11 小时，跑完 16892 项单元测试，将 Chrome 一项架构重构从数月压缩至数周；使 Gemma4 E2B 在指定推理配置下提速 2.85 倍且输出 token 保持一致；机器人连续投中 7 次沙包；AI 辅助设计的 5 种新酶全部通过实验室催化活性验证。该模型提供百万 token 统一工作空间，输入、思考与输出共用窗口，截至 11 月 30 日每百万输入/输出 token 分别为 2 美元和 10 美元，之后恢复为 4 美元和 20 美元。

图表 10: Gemini 4 Argon 在多个测试中的表现

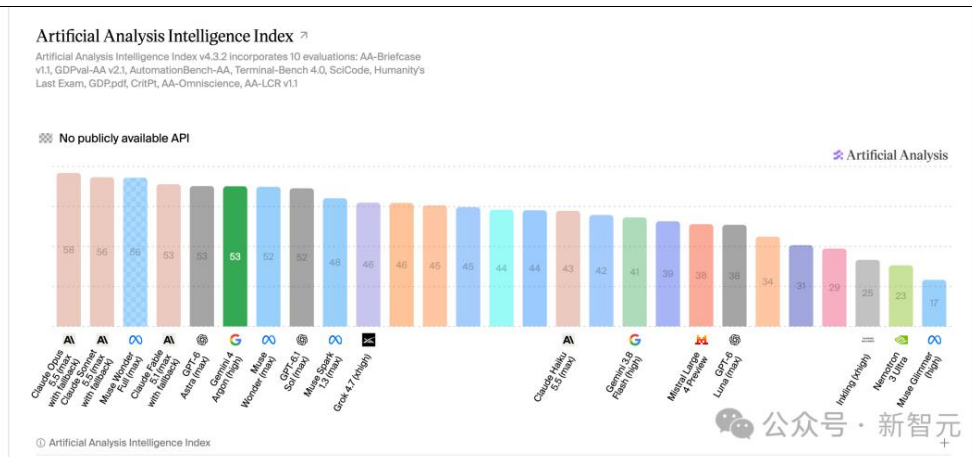
Benchmark		Gemini 4 Argon	GPT-6 Astra	Claude Fable 5.1	Claude Opus 5.5	
Knowledge work	Vals Index	68.9%	63.1%	65.8%	67.0%	
	AutomationBench	Score 51.3%	41.4%	31.4%	42.5%	
	Vals Finance Agent v2	65.4%	53.5%	58.9%	58.6%	
	Harvey's Legal Agent Benchmark	19.6%	5.4%	6.7%	3.8%	
Agentic coding	DeepSWE v1	77.9%	74.1%	67.4%	74.2%	
	FrontierSWE v2	55.0%	65.5%	56.3%	62.3%	
	Vibe Code Bench	91.9%	89.6%	90.3%	90.3%	
	Terminal-bench 4.0	57.4%	58.2%	57.9%	66.4%	
ML engineering	PostTrainBench	45.3%	44.3%	40.2%	49.3%	
Science and math	Terminal-Bench Science 0.1	57.6%	68.1%	52.6%	63.3%	
	LABBench 2	88.8%	85.4%	68.6%	73.1%	
	RiemannBench	76.0%	72.0%	65.6%	69.6%	
Long context	GraphWalks	Up to 128k, BFS (F1)	99.7%	98.7%	91.4%	90.6%
		256k to 1M, BFS (F1)	84.2%	71.8%	65.0%	66.8%
Computer use	Agent's Last Exam	Pass rate 39.5%	34.2%	—	38.2%	
	OSWorld-2.0	Offline subset Partial score 69.2%	72.6%	—	—	
Multimodal understanding	Chartography	71.6%	71.0%	46.2%	66.3%	
	LVNBench	91.7%	87.5%	79.7%	83.7%	
Cybersecurity	CWE-bench v1	68.0%	68.0%	58.0%	67.0%	

Methodology: deepmind.google/models/evals-methodology/gemini-4-argon

 公众号·新智元

资料来源：新智元，华鑫证券研究

图表 11: Gemini 4 Argon 在 Artificial Analysis 智力指数中的表现



资料来源：新智元，华鑫证券研究

在代码生成与 3D 世界构建方面，Argon 展现出突出能力：在 Antigravity 中搭建中国风寺庙建筑群，补齐飞檐、斗拱、植被与光影细节；依据办公桌照片搭建可交互微型赛车场，借助 Blender 构建场景与碰撞检测、NanoBanana 补充纹理、Three.js 接入物理模拟与操控，并以 Gemini Omni Flash 生成引擎声；不依赖现成 3D 引擎，用 Python 和 NumPy 编写光线追踪程序，从零渲染出阳光图书馆并配套视频工具与混音器；制作巫师国际象棋，完成棋子建模、骨骼绑定、动作贴图、网页交互与单人对战 AI。机器人任务中，Argon 观看人类投沙包视频后在仿真环境练习，将动作迁移至 Apollo 真机，并在真实场地根据落点微调释放位置，最终连续投中 7 次，原本需数周的人工仿真建模被压缩至数小时。

在大型工程维护方面，谷歌将约 4000 万行代码的 Chromium 代码库交由 Argon 处理，其中一个旧浏览器功能目录约 600 万行，需拆出核心网页处理逻辑、改动横跨数百文件；Argon 并行推进多组修改，将数月重构压缩至数周，使网页在标签页、侧边栏与浮动窗口间移动时无需重新加载，并消除一类浏览器崩溃问题。在开源 Agent 框架 ADK 上，Argon 面对约 20.5 万行 Python 代码，按维护规则与目标模式下无人值守运行 11 小时，完成瓶颈排查、代码修改与自我审查，16,892 项单元测试，提交涉及 32 个文件、6 项优化的合并请求，维护者保留 2 项并结合首轮成果合入主分支，最终长会话内存会话准备耗时减少约 40%，冷导入耗时减少 21%。Argon 还优化了 Gemma4 E2B 推理程序，在 A100、vLLM 配置下将吞吐从每秒约 3.79 万 token 提升至 10.8 万以上，端到端提速 2.85 倍。同时其在 Antigravity 中的使用使 Agent 代码提交量增加超过 35%，代码回滚减少近三分之一。

图表 12：Gemini 4 Argon 优化开源 Agent 框架 ADK

Same goal. Round 2 added the repo rulebook and /goal mode.

How Gemini 4 Argon's overnight behavior changed once maintainer rules and /goal mode were active.

What we compared	Round 1 (no repo rulebook)	Round 2 (with .agents/skills + /goal mode)
Active runtime	About 2 hours	11 hours overnight (9:11 PM to 8:19 AM)
First pass	Wrote a benchmark at 7 min; first ADK edit at 70 min	Read 12 .agents/skills guideline files first; ran baseline benchmark at 10 min before any edit
Public schemas and visibility	Altered AuthConfig JSON schema	Left AuthConfig untouched; placed new helpers in private _-prefixed modules
Test verification	Ran 216 nearby unit tests	Added regression tests and ran all 16,892 unit tests in the repo, plus type checks and linters
Pull request handoff	No benchmark table in PR description	Opened one unmerged PR in the morning with before-and-after tables across four workloads

公众号 · 新智元

资料来源：新智元，华鑫证券研究

在科学研究领域，Argon 也表现优秀。谷歌与马里兰大学研究者合作，针对 1967 年提出的 Grad 猜想找到了反例。研究团队借助 Argon 去掉原有解中的仿星器对称性限制，几小时内得到候选解并编写验证代码，该成果为仿星器模拟程序提供校验基准。谷歌科学家还与 Argon 设计出 5 种切开特定化学键的新酶，将原本数月的筛选调模环节压缩至数小时，5 种候选酶均能催化目标反应，产物在紫外光下发亮。Argon 还优化了内部 AlphaQubit 纠错解码器，使单位时间处理数据接近翻倍，并写出经 Lean 形式化验证的反正弦量子子程序，时空开销降至已发表量子 CORDIC 方案的 1/100 以下。

3、AI 融资动向：无问芯穹完成 23 亿元 C 轮融资，异构算力与 Token 工厂卡位基础设施核心枢纽

国内 AI 原生基础设施服务商无问芯穹近日宣布完成 23 亿元 C 轮融资，累计融资额达 43 亿元，单轮与累计融资额均创国内 AI Infra（人工智能基础设施）赛道新高，且据外媒报道公司已启动 IPO 筹备。公司自 2023 年成立以来背后已集结多元股东阵容，涵盖国资机构、产业资本与市场化财务投资机构。其中，国资机构包括杭州高新金投、社保中关村基金（君联资本管理）、珠海科技产业集团、厦门国兴投资、徐汇资本、孚腾资本、中保投资、国开科创等；产业投资方包括 Dabanc Holding、联想、九坤创投、同歌创投、腾讯、百度等；财务投资机构包括红杉中国、洪泰基金、惠远资本、启明创投、达晨财智、顺为资本、砺思资本等。学术与创始团队背景方面，公司深度植根于清华大学电子工程系自 2008 年起深耕软硬件协同的 NICS-EFC 实验组；公司发起人汪玉为清华大学电子工程系长聘教授、时任系主任及 IEEE Fellow；联合创始人兼 CEO 夏立雪本硕博均毕业于清华电子系并曾在阿里云负责产品开发与管理；联合创始人戴国浩现任上海交通大学副教授；CTO 李伯勋曾任旷视科技算法总监。

业务布局与技术壁垒方面，无问芯穹围绕“M 种模型×N 种芯片”构建多元异构软硬件协同体系，在 WAIC 上正式提出“前店后厂一中心”战略架构。一中心为“算力集散中心”（Agentic Infra 平台），负责汇聚和调度异构算力，目前在全国部署触达超 37000P 算力、覆盖 16 种主流芯片；其面向异构集群的跨域训练系统已完成生产级检验：联合中国电信完成超 4000 公里（新疆哈密至广东深圳）跨域混训且性能提升 165%、打通 4 个不同代际 GPU 集群使协同性能提升 41%、实现本地私有集群与公有云混训使整体性能提升 68%，此前更完成全球首个“4+2”式 700 亿参数大模型跨国产与进口 6 种芯片混合训练（算力利用率达 97.6%），实现新芯片 2 个工作日内适配、适配成本降低超 70%、万亿参数 MoE 训练吞吐提升 41%至 75%。后厂为“Token 工厂”（Agentic MaaS 大模型服务平台），负责将底层异构算力转化为标准化 Token 服务，平台单日 Token 调用量较去年 12 月爆发增长 40 倍；其首创跨集群异构 PD 分离架构 PDD，实现首 Token 延迟降低 51.5%、单 Token 成本降低 37.5%，过去一年推动推理成本下降 10 倍且未来仍有 10 倍降本空间。前店为“AI 生产力商店”（Agentic Infra 行业解决方案），以智能体自适应技术将 Token 落地为交付业务成果，在 3D 游戏（携手 VAST Tripo 大模型打造 AGENTIC 3D GAME GEN）、法律终端（打造律所专属 AI 合伙人“律盾芯人”）及能源电力（联合南网能源探索智算中心能耗预测与调度）实现商业化落地。商业模式与生态卡位方面，无问芯穹采用不自建实体数据中心的轻资产运营模式，保持不自研芯片、不做基础模型的中立定位，作为独立算网枢纽连接底层硬件与上层模型，平台已上线百余款主流大模型并服务 Kimi、智谱等头部客户与科研机构。

图表 13：上周 AI 初创公司融资动态

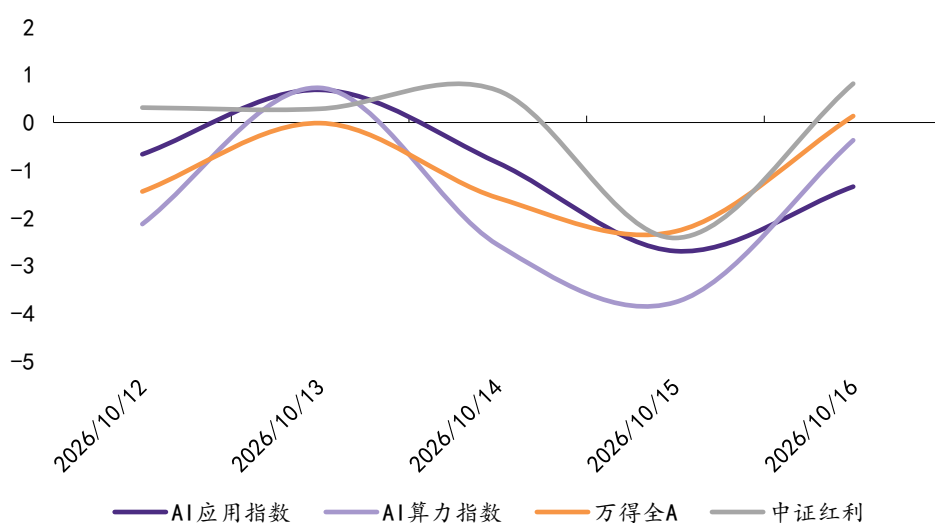
应用	应用类型	领投方	融资轮	融资额	目前累计 融资额	目前估值
Modal Labs	AI 云基础设施平台	Accel	D 轮	7.5 亿美元	暂未披露	157.5 亿美元
无问芯穹	AI 原生基础设施	高新金投、社保中关村基金、腾讯、百度等	C 轮	23 亿元	43 亿元	暂未披露
Oxide	云计算基础设施公司	Eclipse	D 轮	4.45 亿美元	暂未披露	暂未披露

资料来源：Saasverse，智东西，新浪财经，华鑫证券研究

4、行情复盘

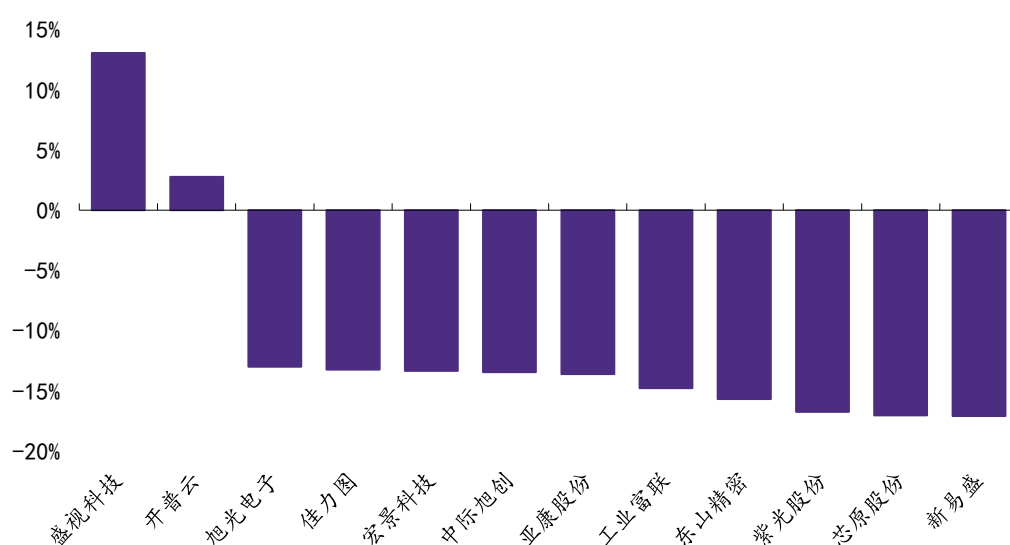
上周（2026.10.12-2026.10.16日），AI应用指数/AI算力指数/万得全A/中证红利日涨幅最大值分别为0.69%/0.73%/0.14%/0.82%，AI应用指数/AI算力指数/万得全A/中证红利日跌幅最大值分别为-2.7%/-3.77%/-2.28%/-2.42%。AI算力指数内部，盛视科技以13.06%录得上周最大涨幅，新易盛以-17.1%录得上周最大跌幅。AI应用指数内部，金自天正以15.1%录得上周最大涨幅，芯原股份以-17.05%录得上周最大跌幅。

图表 14：上周（2026.10.12-2026.10.16日）指数日涨跌幅



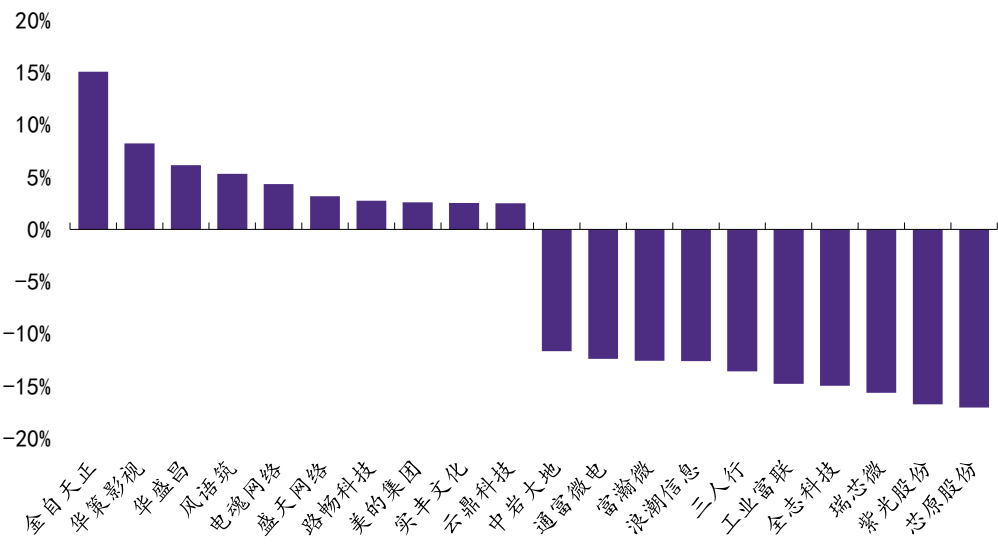
资料来源：wind, 华鑫证券研究

图表 15：上周（2026.10.12-2026.10.16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. 10. 12-2026. 10. 16 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 10 月 12 至 15 日，2026 年 OCP 全球峰会在圣何塞举行，CPO 的产业化进展成为本届峰会焦点。英伟达确认，下一代 RubinUltraNVL576 的 scale-up 互联将采用 CPO 方案。博通现场展示第三代 TH6-DavissonCPO 方案，Celestica、台达、Micas 等合作伙伴演示了基于该方案的 102.4TCPO 整机系统，从交换芯片到整机集成的产业链环节基本打通。可插拔路线同步推进，英伟达与思科联合提出面向单通道 400G 的 USFP 光模块新形态。机架与设施层面，英伟达向 OCP 贡献 VeraRubin 机架与计算托架设计，MGX1.2 规范支持最高 250kW 全液冷机架，并联合 AMD、谷歌、Meta 发布开放数据中心规范 v0.7；AMD 发布智能体机架 Helion，并披露 Helios 首批系统已在 OpenAI、Meta、微软与甲骨文安装运行；Meta 将采用 Spectrum-6 交换机建设 VeraRubinAI 工厂。

本届峰会的核心信号在于，CPO 的应用边界正从交换网络扩展到算力机架内部。此前 CPO 的落地集中在交换机侧的 scale-out 场景，RubinUltraNVL576 的 scale-up 互联采用 CPO，意味着光引擎将进一步贴近计算芯片，scale-up 这个时延与功耗最敏感、此前由铜互联主导的场景开始向光互联打开。供给端，博通第三代方案与多家整机伙伴的系统演示表明，当前 CPO 已从单点技术验证进入芯片、封装、整机协同推进的生态化阶段。需求端的验证窗口也将于下周开启，各大 CSP 将陆续发布新一季财报，Meta、微软、Alphabet 与亚马逊分别定于 10 月 21 日、28 日与 29 日披露业绩，其资本开支指引与资金来源将成为检验 AI 算力需求持续性的直接依据。

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
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	2026-10-19 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	440.38	-0.40	2.13	4.49	-1100.95	206.75	98.08	买入
301196.SZ	唯科科技	81.85	2.05	1.92	2.39	39.28	42.60	34.20	买入
301369.SZ	联动科技	95.01	0.48	1.08	2.62	197.94	87.97	36.26	买入
603859.SH	能科科技	38.82	0.92	1.05	1.32	42.20	36.97	29.41	买入
688615.SH	合合信息	84.15	3.24	4.22	5.25	25.97	19.94	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。

■ 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。