仅供内部参考，请勿外传！

证券研究报告

2026 年 08 月 10 日

华鑫证券 CHINA FORTUNE SECURITIES

千问发布 Qwen3.8-Max，JoyAI-Video-Edit 开源

—计算机行业周报

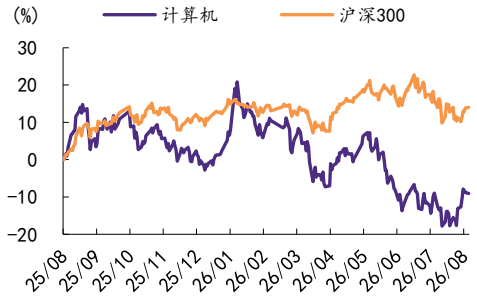
推荐(维持)

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)	-0.1	-14.7	-7.3
沪深 300	-1.6	-3.5	14.5

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：MiniMax 发布 MiniMaxH3，KimiK3 正式开源》2026-08-04

2、《计算机行业周报：Anthropic 发布 ClaudeOpus5，谷歌发布三款 Flash 模型》2026-07-28

3、《计算机行业周报：月之暗面正式发布 KimiK3，StepAOS 与 STEPXNeo 发布》2026-07-22

投资要点

算力：阿里云上线 One Key MCP 服务，千问发布新一代旗舰大模型 Qwen3.8-Max

2026 年 8 月 3 日，阿里巴巴旗下千问大模型正式发布新一代旗舰模型 Qwen3.8-Max。该模型基于 Qwen 3.5 的架构基础构建，参数规模扩展至 2.4 万亿，激活参数 950 亿，支持 100 万上下文 Tokens。新模型在编程、办公、科研以及长周期任务等方面实现了全面提升，为千问首个计划开放权重的 Max 级模型，官方称其为 Qwen 家族迄今最强大的模型。

AI 应用：Canva 周访问量环比+2.27%，JoyAI-Video-Edit 开源

2026 年 8 月 5 日，京东正式开源了 JoyAI-Video-Edit。该模型是首个在流式架构、实时速度与可用质量三个维度上同时达到实用水平的视频编辑工具。其技术核心在于实现了 720P 分辨率下每秒 30 帧的模型推理速度与端到端 24FPS 的稳定输出，并支持无上限时长的流式编辑。

AI 融资动向：OLIX 完成 3.12 亿美元 B 轮融资，投后估值达 33 亿美元

2026 年 8 月 3 日，AI 推理芯片初创公司 OLIX 宣布完成 3.12 亿美元 B 轮融资，投后估值达到 33 亿美元。本轮融资由 Fundomo 领投，Hummingbird Ventures、Crane、Plural、Creandum、Arm、Hudson River Trading、Sovereign AI Fund 等多家机构参投，Netflix 联合创始人 Reed Hastings 以个人身份参投。此前，该公司已于 2026 年 2 月完成 2.2 亿美元 A 轮融资，估值达 10 亿美元，半年时间估值增长至 3 倍，当前累计融资额已超过 5.42 亿美元。

投资建议

2026 年 8 月 3 日，Palantir 发布 2026Q2 财报：当季实现总营收 19.35 亿美元，同比增长 93%、环比增长 19%；毛利率 85%，同比提升 4 个百分点；GAAP 净利润 10.62 亿美元（净利率 55%），摊薄后每股收益 0.41 美元。分业务看，商业业务收入 9.45 亿美元，同比增长 110%，其中美国商业收入 7.64 亿美元，同比增长 149%；政府业务收入 9.90 亿美元，同比增长 79%，其中美国政府收入同比增长 90%；美国区收入占总营收比重由上年同期的 73%升至 81%。订单层面，当季新签 1000 万

行业研究

仅供内部参考，请勿外传！

美元以上合同 73 个；美国商业新签总合同价值（TCV）21.32 亿美元创历史新高，同比增长 153%；期末剩余履约义务（RPO）49 亿美元，约 43%预计于未来 12 个月内确认为收入。Non-GAAP 口径下，调整后营业利润 11.94 亿美元（利润率 62%），Ruleof40 得分达 155%。指引方面，公司将 2026 全年营收指引上调至 81.50-81.58 亿美元（对应约 82%增长），美国商业收入指引上调至 34.24 亿美元以上（对应至少约 134%增长）。

本次财报印证了企级 AI 应用正从试点验证期迈入规模化部署期，AI 软件侧的商业化兑现显著加速。从财报的结构层面看，商业收入的增速显著超越政府业务的增速，这表明公司增长引擎已由传统政府市场切换至企业级商业市场，需求结构更具延展性。从商业模式看，公司在收入接近翻倍的同时实现 85%的毛利率和 62%的调整后营业利润率，Ruleof40 得分达 155%，"高增长+高盈利"的组合打破了软件行业增长与利润率此消彼长的传统约束，印证 AI 应用具备极强的经营杠杆。指引层面，公司的全年营收指引上调幅度远超常规季度修正，印证需求仍在加速而非收敛；CEOKarp 提出的"AI 主权"需求与"将 Token 转化为实际经济价值"的表述，指向企业级 AI 竞争的焦点正从模型能力转向数据治理、决策闭环与行业 know-how 的工程化落地。我们认为，Palantir 作为企业级 AI 应用的风向标，其订单与收入的同步加速表明本轮 AI 产业景气已下游的应用传导，AI 应用侧逐步进入商业化兑现的验证窗口；对国内市场而言，具备行业数据积累、工程化交付能力与政企客户资源的 AI 应用及数据要素相关公司有望承接相似的需求逻辑，配置价值逐步显现。

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	2026-08-10 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	488.18	-0.30	0.30	0.60	-1627.27	1627.27	813.63	买入

仅供内部参考，请勿外传！

301196.SZ	唯科科技	82.67	2.53	3.34	3.98	32.68	24.75	20.77	买入
301369.SZ	联动科技	137.20	0.48	1.08	2.62	285.83	127.04	52.37	买入
603859.SH	能科科技	43.03	0.92	1.05	1.32	46.77	40.98	32.60	买入
688615.SH	合合信息	121.67	3.24	4.22	5.25	37.55	28.83	23.18	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：阿里云上线 ONE KEY MCP 服务，千问发布新一代旗舰大模型 QWEN3.8-MAX 5
1.1、 Tokens 跟踪 5
1.2、 数据跟踪：阿里云上线 One Key MCP 服务，Quick BI 完成 AI-native 架构升级 7
1.3、 产业动态：千问发布新一代旗舰大模型 Qwen3.8-Max，API 服务与 Token Plan 同步上线 8
2、 AI 应用动态：CANVA 周访问量环比+2.27%，JOYAI-VIDEO-EDIT 开源 13
2.1、 周流量跟踪：Canva 周访问量环比+2.27% 13
2.2、 产业动态：JoyAI-Video-Edit 开源，实时视频编辑重塑创作工作流 13
3、 AI 融资动向：OLIX 完成 3.12 亿美元 B 轮融资，投后估值达 33 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） 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：QWEN3.8-MAX 模型性能提升概览 9
图表 4：TEXT ARENA、CODEARENA、VISIONARENA 榜单排名情况 9
图表 5：QWEN3.8-MAX 实现最新研究论文的复现与超越 10
图表 6：QWEN3.8-MAX 随 RL 训练强化性能表现稳步提升 11
图表 7：QWEN3.8-MAX 上线千问 AI 平台 12
图表 8：2026.7.31-2026.8.6 AI 相关网站流量 13
图表 9：JOYAI-VIDEO-EDIT 发布 14
图表 10：JOYAI-VIDEO-EDIT 的架构 14
图表 11：JOYAI-VIDEO-EDIT 在 NVIDIA B200 GPU 上运行时的分析 15
图表 12：人们对 JOYAI-VIDEO-EDIT 与其他模型的偏好评估 15
图表 13：上周 AI 初创公司融资动态 16
图表 14：上周（2026.8.3-2026.8.7 日）指数日涨跌幅 18
图表 15：上周（2026.8.3-2026.8.7 日）AI 算力指数内部涨跌幅度排名 18
图表 16：上周（2026.8.3-2026.8.7 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
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：阿里云上线 One Key MCP 服务，千问发布新一代旗舰大模型 Qwen3.8-Max

1.1、 Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 8 月 3 日至 8 月 9 日，周度 Token 消耗量大幅上涨，消耗量为 69T，环比上周增加 21.48%。在 Tokens 规模 Leaderboard 前五名中，DeepSeek 的 DeepSeek V4 Flash 0731 以 8.83T tokens 位列榜首；Tencent 的 Hy3 以 8.05T tokens 位居第二；DeepSeek 的 DeepSeek V4 Flash 0423 以 5.88T tokens 位居第三；Xiaomi 的 MiMo-V2.5 以 5.39T tokens 位居第四；OpenAI 的 GPT-5.6 Luna 以 4.43T tokens 位居第五。

文本类模型的市场份额，DeepSeek 以 1.01B requests 占据 27.6%，位居第一；Google 以 925M requests 占据 25.2%的份额，位列第二；OpenAI、Qwen、Anthropic 则分别以 602M、190M、138M requests，对应占据 16.4%、5.2%、3.8%的市场份额。

图像生成类模型调用量大幅回落，调用量为 5.88M requests，环比上周减少 13.27%。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 1.69M requests 位居第一；Nano Banana 2 (Gemini 3.1 Flash Image) 、Nano Banana 2 (Gemini 3.1 Flash Image Preview)、Banana 2 Lite (Gemini 3.1 Flash Lite Image)以 903K、584K、479K requests 位居第二、第三和第四；Nano Banana Pro (Gemini 3 Pro Image)以 440K requests 位列第五。从市场份额来看，Google 以 4.51M requests 占据 76.7%的份额，位居第一；OpenAI 以 632K requests 占据 10.7%，位列第二；xAI、Black Forest Labs、ByteDance Seed 则分别以 323K、163K、148K requests，对应占据 5.5%、2.8%、2.5%的市场份额。

嵌入模型调用量有所下降，调用量为 268M requests，环比上周减少 6.62%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 95.1M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 52.4M requests 位列第二；BAAI 的 bge-m3 以 19.7M requests 位列第三；Perplexity 的 Embed V1 0.6B 以 18.9M requests 位列第四；Google 的 Gemini Embedding 001 以 18.2M requests 位列第五。从市场份额来看，OpenAI 以 113M requests 占据 42.2%的份额，位居第一；Qwen 以 57M requests 占据 21.3%，位列第二；Google、Perplexity、BAAI 则分别以 36.3M、22.2M、22M requests，对应占据 13.5%、8.3%、8.2%的市场份额。

重排序模型调用量显著上升，调用量为 4.79M requests，环比上周增加 21.27%。在 Rerank Leaderboard 中，VoyageAI 的 rerank-2.5 以 1.87M requests 位列榜首；NVIDIA 的 Llama Nemotron Rerank VL 1B V2 以 1.45M requests 位居第二；Cohere 的 Rerank v3.5、Rerank 4 Fast、Rerank 4 Pro 以 672K、349K、328K requests 位列第三、第四和第五。从市场份额来看，VoyageAI 以 1.99M requests 占据 41.5%的份额，位居第一；NVIDIA 以 1.45M requests 占据 30.3%的份额，位居第二；Cohere 以 1.35M requests 占据 28.2%的份额，位列第三。

视频生成类模型调用量有所上升，调用量为 175K requests，环比上周增加 10.06%。在 Video Leaderboard 前五名中，Google 的 Veo 3.1 Fast、Veo 3.1 Lite 分别以 42K、33K requests 位列第一、第二；ByteDance 的 Seedance 2.0、Seedance 2.0 Fast 分别以 20K、

仅供内部参考，请勿外传！

17K requests 位居第三、四；Google 的 Veo 3.1 以 12K requests 位居第五。从市场份额来看，Google 以 87K requests 占据 49.7%的份额，位居第一；ByteDance 以 46K requests 占据 26.3%，位列第二；xAI、MiniMax、KwaiVGI 则分别以 12K、8K、8K requests，对应占据 7.1%、4.7%、4.5%的市场份额。

语音生成类模型调用量有所回落，调用量为 1.54M requests，环比上周减少 6.10%。在 Speech Leaderboard 前五名中，Google 的 Gemini 3.1 Flash TTS Preview 以 547K requests 位居第一；Hexgrad 的 Kokoro 82M 以 405K requests 位列第二；xAI 的 Grok Voice TTS 1.0 以 204K requests 位列第三；Fish Audio 的 S2.1 Pro Free (free)以 128K requests 位列第四；Microsoft 的 MAI-Voice-2 以 69K requests 位列第五。从市场份额来看，Google 以 547K requests 占据 35.6%的份额，位居第一；Hexgrad 以 405K requests 占据 26.4%，位列第二；xAI、Fish Audio、Microsoft 则分别以 204K、193K、81K requests，对应占据 13.3%、12.6%、5.3%的市场份额。

语音转写类模型调用量大幅上涨，调用量为 10.6M requests，环比上周增加 17.00%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3、Whisper Large V3 Turbo、GPT-4o Mini Transcribe 分别以 5.84M、2.3M、804K requests 位列第一、第二和第三；Qwen 的 Qwen3 ASR Flash 以 289K requests 位列第四；OpenAI 的 GPT-4o Transcribe 以 283K requests 位列第五。从市场份额来看，OpenAI 以 9.56M requests 占据 90.4%的份额，位居第一；Qwen 以 289K requests 占据 2.7%，位列第二；Mistral AI、xAI、Microsoft 则分别以161K、128K requests，对应占据 1.8%、1.5%、1.2%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 3 日，高盛亚太互联网团队分析师 Ronald Keung 将 2026 年末中国大模型厂商合计 ARR 预期上调至 130 亿美元，并预计 2030 年该数字将攀升至 1250 亿美元，五年实现约 25 倍增长。与此同时，高盛预测中国大模型的 API 及订阅收入将从 2026 年的 350 亿元增长至 2030 年的 8790 亿元，对应每日 token 消耗量从 350 万亿增至 4600 万亿，收入与 token 消耗量同步实现约 25 倍扩张。

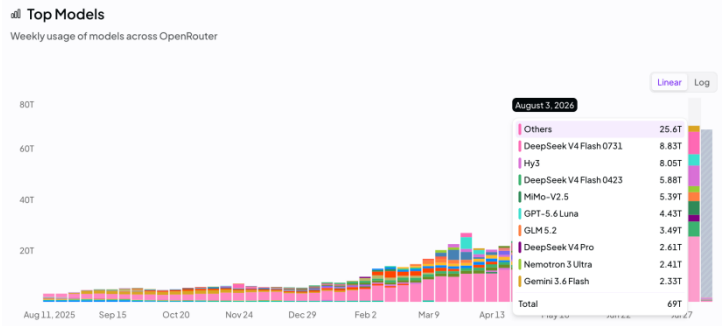
近日，根据美国研究机构 Artificial Analysis 最新测试数据，DeepSeek-V4-Flash 已成为全球主流大模型中运行成本最低的产品（每百万输入 Token 0.14 美元、每百万输出 Token 0.28 美元），运行成本不到 Anthropic 旗舰模型 Claude Fable 5 的百分之一。数据显示，在对每项智能指数任务的加权平均成本测算中，DeepSeek-V4-Flash 的平均测试成本仅为 3 美分，相较之下，月之暗面旗下 Kimi K3 为 86 美分，OpenAI 旗下 GPT-5.6 Sol 为 1.86 美元，Claude Fable 5 高达 3.15 美元。

近日，全国首单 Token 算力工厂科创债券成功落地无锡。本期债券由无锡高新区科创产业发展集团有限公司发行，发行规模 6 亿元，期限 5 年，票面利率 2.03%，认购倍数达 3.1 倍。募集资金将重点投向 AI 算力集群建设，专项用于全国首个“东数西算”、算电融合特大型 Token 工厂——无锡燧弘华创 Token 工厂项目，构建全栈 AI 产业体系。

8 月 6 日，DeepSeek 发布公告称计划于近期对 API 服务价格进行整体上调，预计涨幅较为显著，建议用户根据自身使用需求提前做好规划安排。具体调整方案将以正式通知为准。

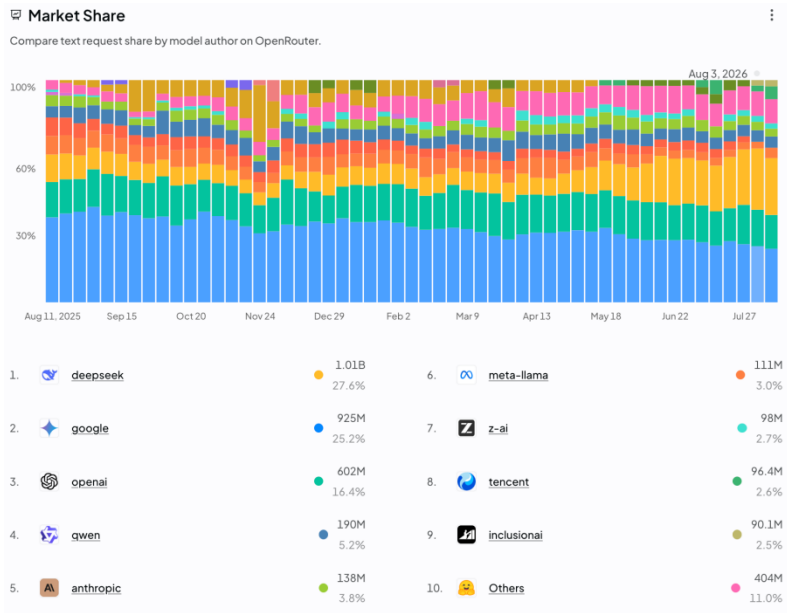
仅供内部参考，请勿外传！

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：阿里云上线 One Key MCP 服务，Quick BI 完成 AI-native 架构升级

2026 年 8 月 5 日，阿里云正式上线 One Key MCP 服务。该服务兼容 Qoder、Codex、Claude Code、Cursor 等主流 Coding Agent，允许开发者通过统一的阿里云百炼 API Key，一键调用所有生态伙伴的 MCP 服务。

当前，开发者可登录阿里云百炼，在 MCP 广场使用现有 API Key 开通并调用相关服务，快速接入生态体系。在此之前，开发者接入不同 MCP 服务时，需要分别申请凭证、独立配置鉴权、逐一完成联调，随着服务数量的增加，接入和维护成本也进一步攀升。新服务将这一流程集中至阿里云百炼平台，为开发者提供五项核心能力，其中包括：统一密钥管理，以减少不同服务商的凭证申请与配置；平台内无感开通，以简化服务商控制台跳转步骤；统一计量与账单，以便于企业进行用量查看与对账；多框架适配，以支持 CLI 和公网服务地址调用；以及安全管控，以支持全链路加密及 API Key 黑白名单 IP 访问限制。

仅供内部参考，请勿外传！

截至目前，该服务首批合作商已覆盖电商供应链、时空与环境数据、金融投资、法律合规、产业与企业数据、物流与车辆六大领域，共吸引 14 家合作伙伴入驻，其中包括 1688、高德云图、青悦信息、东方财富、恒生聚源、今日投资、盈米基金-且慢 AI、龙柏数据、得理科技、e 签宝、北京上奇数字科技、秀派科技、安那其科技、网尚科技。同时，阿里云表示将持续扩展服务覆盖范围，引入更多垂直领域伙伴，共建开放的 Agent 服务生态。

与此同时，阿里云瓴羊企业级智能分析产品 Quick BI 于 8 月 4 日完成 AI-native 架构升级，正式发布 AIPro 版。该版本从数据接入到分析交互各层均以 AI 为底座重新设计，用户通过对话即可直接获取数据结果、洞察和行动建议。

新版本在分析力、理解力、行动力、可信度、成长性五个方面进行了升级：在分析能力方面，AIPro 可自动拆解字段含义、智能匹配数据表，支持多轮追问和层层下钻，最终交付包含归因逻辑和决策建议的分析报告，仅需一个入口，即可完成深度分析；在语义理解层面，该版本增强了企业级业务场景的适配度，能够理解行业术语及企业业务逻辑，支持定制分析模板，并通过 Agent 记忆系统沉淀个人分析行为偏好；在行动力方面，分析结论可与业务系统联动，具体包括自动生成可二次编辑的分析作品并支持分享嵌入、定时自动发现经营数据异常并推送通知，以及通过 MCP 连接器将分析结论融入 OA 系统工作流、更新数据或生成待办；最后，在可信度和成长性方面，每个结论均可追溯数据来源和计算逻辑，行列级权限确保数据安全可控，私有化版本支持配置自有大模型、数据不出域，同时，分析思路可上架至企业技能广场供团队复用，支持跨部门共享，实现从个人提效到组织协同进化。

当前，AIPro 已在制造业产线质检、金融业投资分析、连锁零售渠道巡检及互联网业务协作等场景落地，异常发现到归因缩短至分钟级，并可完成多源数据深度归因分析。目前该版本已正式发布，接入数据后打开对话框即可直接提问分析，首月赠送 12.5 万 Credits 免费试用。

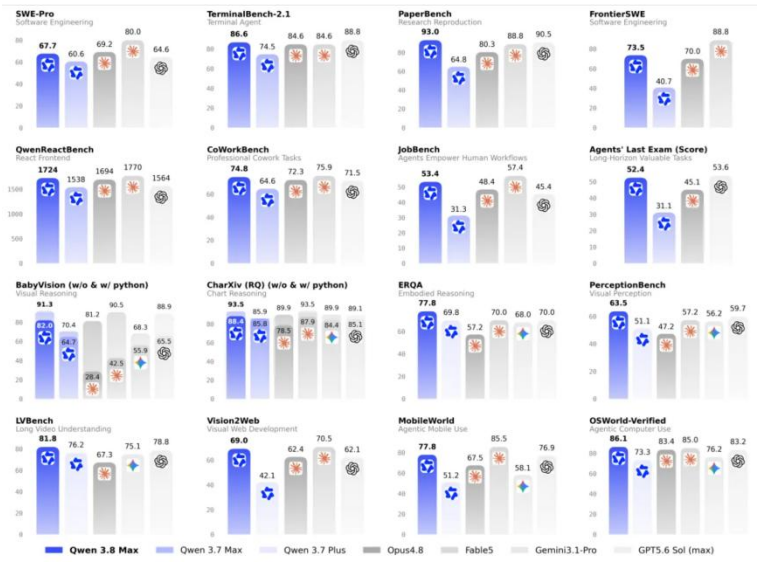
1.3、产业动态：千问发布新一代旗舰大模型 Qwen3.8-Max，API 服务与 Token Plan 同步上线

2026 年 8 月 3 日，阿里巴巴旗下千问大模型正式发布新一代旗舰模型 Qwen3.8-Max。该模型基于 Qwen 3.5 的架构基础构建，参数规模扩展至 2.4 万亿，激活参数 950 亿，支持 100 万上下文 Tokens，为千问首个计划开放权重的 Max 级模型，官方称其为 Qwen 家族迄今最强大的模型。

新模型在编程、办公、科研以及长周期任务等方面实现了全面提升，性能表现正逐步接近 Claude Fable 5、GPT-5.6 Sol 等前沿大模型，在部分基准上已实现超越，在另一些基准上仍有追赶空间。具体来看，多项基准测试结果显示，该模型在论文复现基准 PaperBench 中得分同时超过 GPT-5.6 Sol、Claude Fable 5 和 Claude Opus 4.8，取得 93.0 的成绩，位列榜首；在电脑操作能力基准 OSWorld-Verified 中，该模型以 86.1 分位居参评模型首位。与此同时，在 TerminalBench 2.1 中，该模型获得 86.6 的评分，略低于 GPT-5.6 Sol 的 88.8 分；在 SWE-bench Pro 中，该模型以 67.7 的得分落后于 Fable 5 的 80.0 和 Claude Opus 4.8 的 69.2；在 VideoMME v2 中，该模型得分 68.3，低于 GPT-5.6 Sol 的 71.1 分。

仅供内部参考，请勿外传！

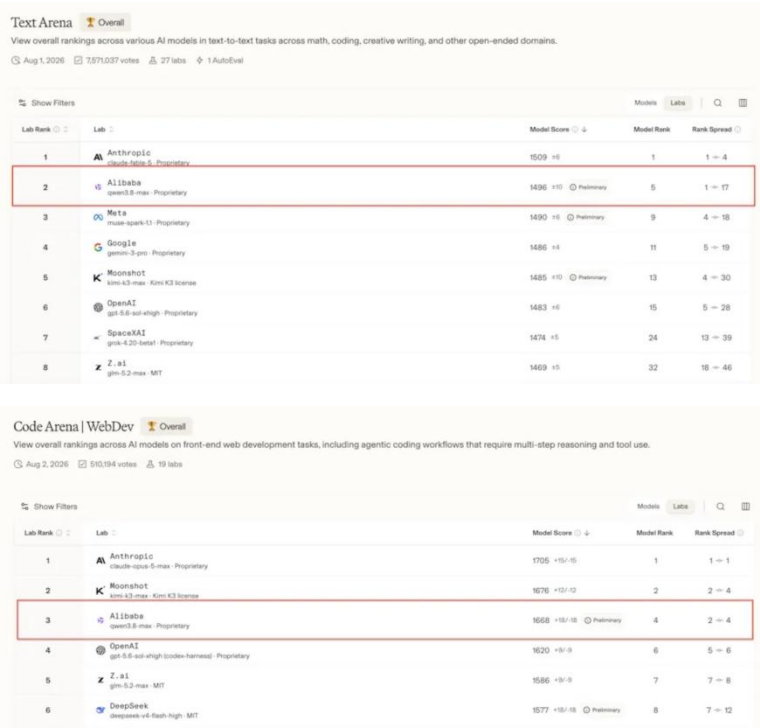
图表 3：Qwen3.8-Max 模型性能提升概览



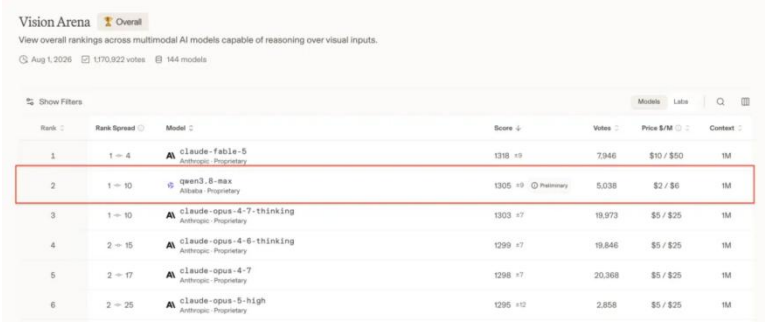
资料来源：千问大模型，华鑫证券研究

此外，新模型于第三方评测中表现亮眼，在国际权威评测榜单 Arena 中，该模型以 1496 的得分位列文本任务 Text Arena 实验室榜第二，仅次于 Claude Fable 5 的 1509；在前端编程能力榜单 CodeArena 中，该模型以 1668 的得分位列实验室榜第三，Claude Opus 5 Max 和 Kimi K3 Max 分别以 1705、1676 的得分位列第一、第二；在视觉能力榜单 VisionArena 中，该模型以 1305 的得分，仅次于 Claude Fable 5 的 1318，略高于 Claude Opus 4.7 Thinking 的 1303，排名第二。

图表 4：Text Arena、CodeArena、VisionArena 榜单排名情况



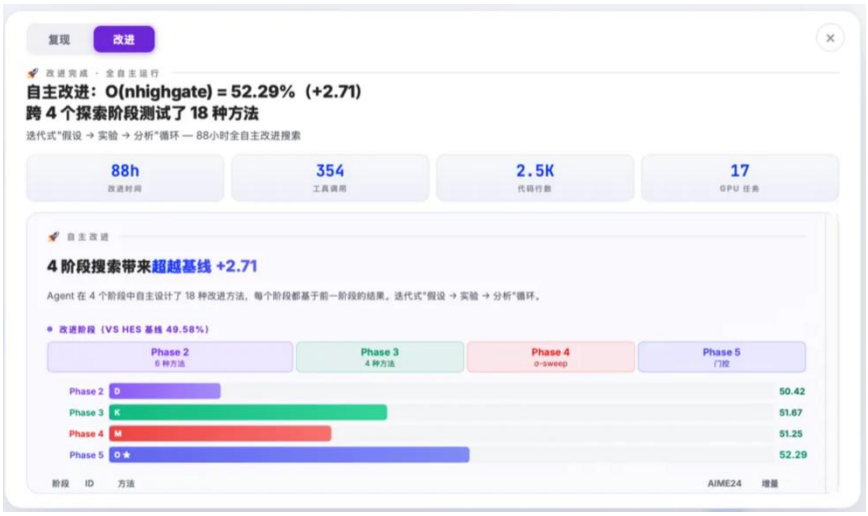
仅供内部参考，请勿外传！



资料来源：千问大模型，华鑫证券研究

编程方面，该模型已具备从零到一的独立项目交付能力，能够从一个空文件夹出发，独立完成跨越数天的真实项目，自行测试迭代并做到交付。在自进化Harness项目的测试案例中，模型被要求从零创建 oh-my-cli 项目，并在十天级长程自动编程中构建一套自进化harness，结果显示，模型完全自主运行约 16 天后，仓库累计留下 265 次 commits、127 个 PR 和 151 个 issues，展现出持续演进的自动编程能力。在论文复现案例中，模型被要求复现一篇最新研究论文《Unified Data Selection for LLM Reasoning》，结果显示，在连续独立工作的约 125 小时里，模型用约 37 小时从零搭建起论文的整套训练与评测流水线，完整复现了论文的六项主要结论，并在随后的约 88 小时里，执行了从提出假设、编写代码，到 GPU 训练、分析结果，再到重复尝试的科研循环，四轮下来独立提出并测试了 18 种改进方向，最终找到超越论文原方法的新做法，同时在竞赛级数学基准 AIME24 上再提升了 2.7 分，实现了论文的复现与超越。

图表 5：Qwen3.8-Max 实现最新研究论文的复现与超越



资料来源：千问大模型，华鑫证券研究

通用办公与专业任务方面，为提升模型在真实工作任务中的表现，团队对真实世界的强化学习系统进行了拓展。与此同时，随着训练的持续强化，该模型在数十个 in-house / public 工作基准上均取得了稳步的性能提升，并在 QwenWork、Claude Code、 Codex、OpenClaw、Hermes 等众多 harness 上取得了相当的性能表现。不仅如此，该模型还在数百种职业的高频工作场景中展现出了跨领域的交付能力，覆盖法律、设计、餐饮、建筑等场景，如在一小时内从数百份企业合规文件中标记出 1284 条相关条款；生成包含 8 个页面的可交互数字银行 App 原型；结合上百份食材供应资料设计 26 道菜菜单并将成本率控制在 33.8%；依据图纸还原出 30 层写字楼的抗震结构模型等等。同时在量化研究案例中，模型仅

仅供内部参考，请勿外传！

根据六条因子描述，将动量、价值、质量、投资、低风险和情绪六类因子分别拆解为 50 个研究方向，调度约 330 个子 Agent，完成约 6000 次回测，将原本需要研究员数周甚至数月串行推进的工作，在单次对话内实现规模化交付。

图表 6：Qwen3.8-Max 随 RL 训练强化性能表现稳步提升



资料来源：千问大模型，华鑫证券研究

长程任务方面，面对极其复杂的长周期、多约束任务，该模型展现出了超越以往的系统级自主规划能力。在芯片设计案例中，模型在集成仿真、综合、物理设计工具链的沙箱中独立完成了 GCD/RSA 密码硬件加速器从算法架构、RTL 代码到多轮优化的全过程，单次不间断运行中历经约 500 轮交互、71 次评估、13 个关键里程碑，将设计从 8,298 门一路优化至 678 门，在所有参评模型中表现最优。在电商经营模拟案例中，模型在模拟 365 天长周期电商经营的 E-Commerce Bench 基准中，手握 10 万元启动资金同时运营多家网店，独立完成选品、供应链议价、库存管理、动态定价等全链路决策，最终以高达 416,252 元（4.16 倍回报）总现金胜出，测评成绩相较第二名 GLM 5.2 高出 38%，相较上一代旗舰模型 Qwen3.7-Max 提升 152%。

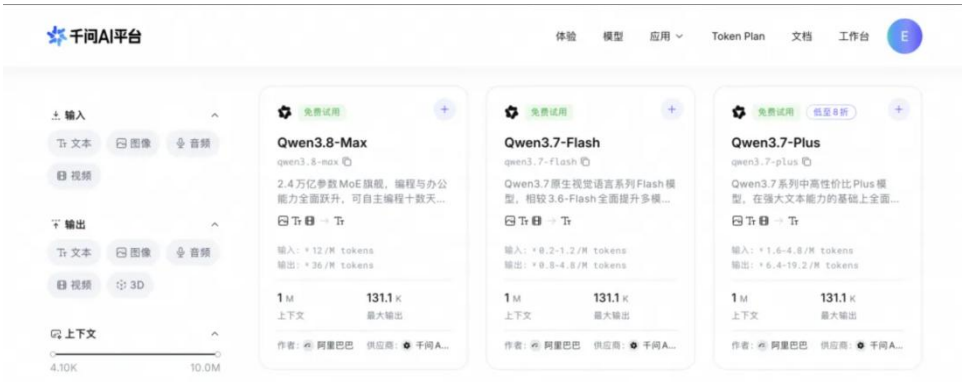
多模态智能体方面，该模型展示出了贯穿任务全流程的视觉生产力。在理解层面，模型可处理超过 200 页的财报和复杂 PDF，具备跨页读懂文字、图表与版面并提炼结论的能力，同时能处理超过 100 小时的长视频素材并构建视频记忆。在执行层面，模型具备将生活素材剪成 Vlog、将界面截图还原为完整前端项目、将户型图建成 3D 装修效果的动手交付能力。在校验层面，模型能够在执行过程中持续观察中间产物，一旦发现偏差即自主修正。同时，该模型具备混合智能体能力，能够通过编程高效完成繁重工作，并结合 GUI 操作触达人类所能操作的一切并带回真实反馈。

当前，Qwen3.8-Max 的 API 服务已在千问 AI 平台首发上线，全球开发者可直接调用。价格方面，国内定价为每百万 Tokens 输入 12 元、输出 36 元，隐式缓存命中 1.5 元；国际定价为每百万 Tokens 输入 2 美元、输出 6 美元，隐式缓存命中 0.25 美元，价格分别为 Opus5 输入、输出价格的 40%与 24%，实现了相近智能水平下的更优性价比。

此外，千问 AI 平台为 Token Plan 用户推出专属限时优惠：夜间时段使用新模型享 5 折 Credits 消耗；8 月 5 日前所有已购买个人版 Token Plan 的用户可获取 1 次 7 天额度重置权益。

仅供内部参考，请勿外传！

图表 7：Qwen3.8-Max 上线千问 AI 平台



资料来源：阿里云，华鑫证券研究

仅供内部参考，请勿外传！

2、 AI 应用动态：Canva 周访问量环比+2.27%，JoyAI-Video-Edit 开源

2.1、 周流量跟踪：Canva 周访问量环比+2.27%

本期（2026.7.31-2026.8.6）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1247.0M）、Bing（864.5M）和 Gemini（588.1M），访问量环比增速第一为 Canva（2.27%）；平均停留时长前三位分别为 Character.AI（00:16:25）、Discord（00:11:11）和 NotionAI（00:08:53）；停留时长环比增速第一为文心一言（4.76%）。

图表 8：2026.7.31-2026.8.6 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	1247.0	-0.72%	6:35	-0.25%
Bing	搜索	微软	864.5	-1.06%	7:21	0.68%
Gemini	聊天机器人	谷歌	588.1	-2.06%	7:01	0.00%
Canva	在线设计	Canva	189.4	2.27%	5:57	-0.28%
Discord	游戏社区	微软	153.4	-0.78%	11:11	-0.30%
Github	代码托管	微软	145.7	-1.09%	6:26	0.52%
Character.AI	聊天机器人	Character.AI	36.43	-1.65%	16:25	-0.81%
DeepI	翻译工具	DeepL	22.76	-3.07%	2:29	0.00%
Perplexity	AI 搜索	Perplexity	21.97	-4.06%	4:19	0.78%
Kimi	聊天机器人	Moonshot AI	15.35	-12.19%	7:07	-1.16%
QuillBot	释义工具	QuillBot	6.746	-4.03%	2:51	-1.16%
NotionAI	文本/笔记	Notion	2.70	-0.66%	8:53	-1.48%
文心一言	聊天机器人	百度	0.04	-68.62%	2:34	4.76%

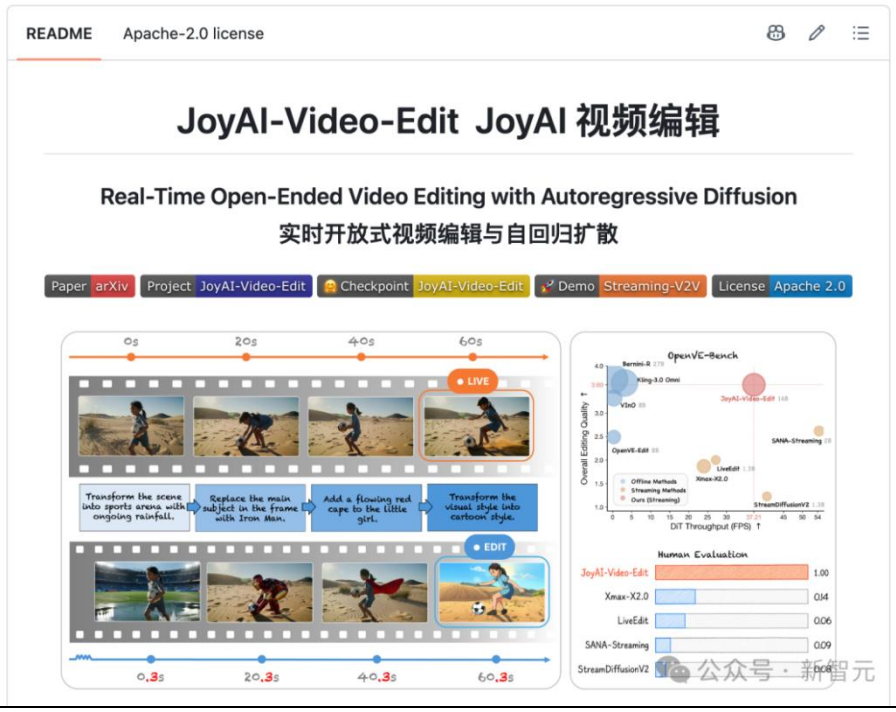
资料来源：similarweb，华鑫证券研究

2.2、 产业动态：JoyAI-Video-Edit 开源，实时视频编辑重塑创作工作流

2026 年 8 月 5 日，京东正式开源了 JoyAI-Video-Edit。该模型是首个在流式架构、实时速度与可用质量三个维度上同时达到实用水平的视频编辑工具。其技术核心在于实现了 720P 分辨率下每秒 30 帧的模型推理速度与端到端 24FPS 的稳定输出，并支持无上限时长的流式编辑。

仅供内部参考，请勿外传！

图表 9：JoyAI-Video-Edit 发布



资料来源：新智元，华鑫证券研究

要实现如此高难度的实时流式编辑，必须攻克两道主要技术难关。首要挑战是处理速度。视频由每秒 24 至 30 帧的画面构成，模型处理速度一旦低于播放速度，便会产生卡顿。JoyAI-Video-Edit 的解决方案是采用逐帧处理、即时输出的流式管线，无需预知完整视频内容。其底层架构集成了 MLLM 条件编码器、因果视频 VAE 及一个拥有 16B 参数的多模态扩散 Transformer。为达到实时性，团队通过名为 SA-DMD 的蒸馏训练方法，将原本需要十几步的迭代生成压缩至两步完成，极大提升了推理效率。在单张 Nvidia B200 GPU 上，端到端延迟仅为 226 毫秒，吞吐量达 30.1 帧/秒。此前同类流式编辑器为追求速度普遍采用小参数量与低分辨率，导致画质不佳，而 JoyAI-Video-Edit 以 160 亿参数和 720P 分辨率实现了速度与质量的兼得。

图表 10：JoyAI-Video-Edit 的架构

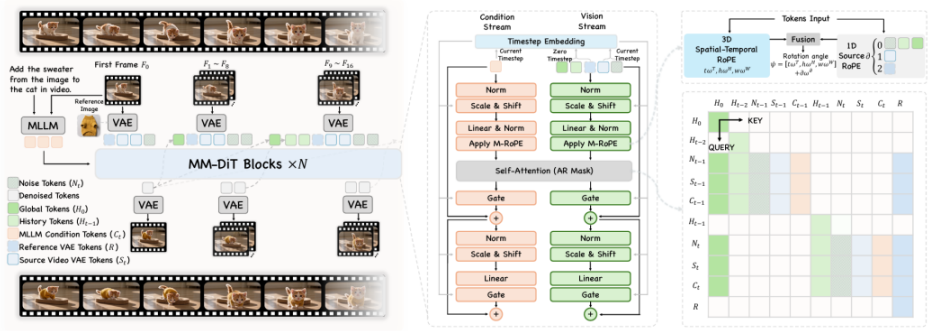


Figure 3 Architecture of JoyAI-Video-Edit. Our autoregressive diffusion architecture consists of an MLLM, a causal video VAE, and an MM-DiT diffusion backbone. The MLLM extracts condition tokens from textual and visual inputs, while the VAE projects videos and optional reference images into a shared latent space. The MM-DiT jointly models the condition tokens and latent visual tokens to generate edited-video latents, which are subsequently decoded by the VAE into the final output video.

公众号 · 新智元

资料来源：新智元，华鑫证券研究

仅供内部参考，请勿外传！

图表 11：JoyAI-Video-Edit 在 Nvidia B200 GPU 上运行时的分析

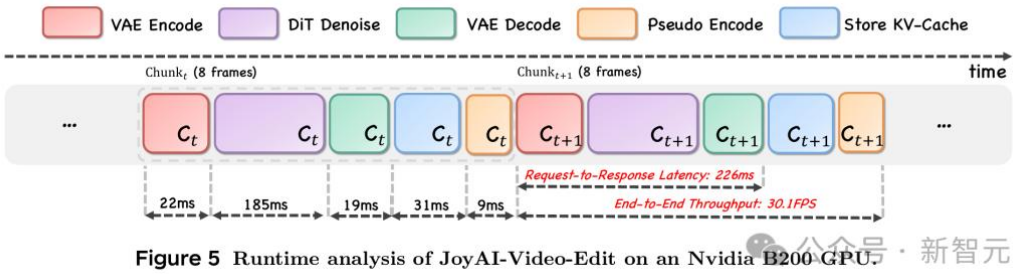


Figure 5 Runtime analysis of JoyAI-Video-Edit on an Nvidia B200 GPU.

公众号 · 新智元

资料来源：新智元，华鑫证券研究

第二个关键难题是长视频处理中的稳定性。过往模型在处理超过几分钟的视频时，常因自回归推理中的误差累积导致画面漂移、色彩失真或风格不一致。流式编辑的难点在于它需同时满足三重约束：时序连贯性、与原视频的相似性以及编辑指令的一致性。JoyAI-Video-Edit 的解决方案是“有界 KV 状态推理”，即限制模型的记忆窗口，仅保留最近几帧与视频首帧作为参考，从而确保无论视频多长，计算负载与内存占用均保持恒定。同时，模型在训练阶段便针对有限记忆条件下的长视频编辑进行了专项优化，使其能在极简信息下维持画面的长期稳定，从而实现无时长限制的流畅编辑。

在性能评估上，该模型表现出了显著优势。在通用视频编辑基准 OpenVE-Bench 中，其总分达到 3.60，在局部删除任务上甚至超越了所有离线对比模型。在针对长视频的 LongV2VBench 测试中，该模型在全部五个场景类别中均位列第一，速度也领先于同类竞品。人类偏好评估进一步印证了其优势，在与四个流式模型的两两盲比中均获得明显更高的偏好率，分别为 90%、87%、87%和 81%。

图表 12：人们对 JoyAI-Video-Edit 与其他模型的偏好评估

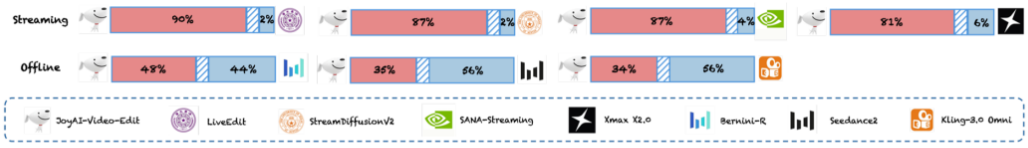


Figure 9 Pairwise human preference between JoyAI-Video-Edit and competing methods. Red and blue denote preferences for JoyAI-Video-Edit and the competing method, respectively; the hatched region denotes ties.

公众号 · 新智元

资料来源：新智元，华鑫证券研究

这种实时流式编辑能力带来的影响是深远的。首先，它消除了创作过程中的“渲染等待”，JoyAI-Video-Edit 的响应延迟只有 226 毫秒，可以将响应延迟压缩至人类感知阈值以下，极大提升了创作手感。其次，它改变了内容生产的逻辑，使得“一次拍摄”能够衍生出“无限多个成品”。例如，在电商场景中，同一直播流可针对不同观众实时呈现不同配色或款式的服装展示，实现个性化内容的按需生成。更深层次地，该技术为摄像头与屏幕之间引入了一个可编程的语义理解层。它不同于简单的滤镜，而是真正理解画面内容并进行连贯的语义级编辑。这将深刻影响预演、远程协作、虚拟试衣、实时风格化等众多领域，其核心价值不仅在于加速最终成品的产出，更在于通过极速反馈帮助团队高效排除错误方案，加速决策与共识的形成。

仅供内部参考，请勿外传！

3、AI 融资动向：OLIX 完成 3.12 亿美元 B 轮融资，投后估值达 33 亿美元

2026 年 8 月 3 日，AI 推理芯片初创公司 OLIX 宣布完成 3.12 亿美元 B 轮融资，投后估值达到 33 亿美元。本轮融资由 Fundomo 领投，Arm、Hudson River Trading、Sovereign AI Fund 等新资方参投，Netflix 联合创始人 Reed Hastings 以个人身份参投，Hummingbird Ventures、Crane、Plural、Creandum 等现有股东持续跟投。此前，该公司已于 2026 年 2 月完成 2.2 亿美元 A 轮融资，估值达 10 亿美元，半年时间估值增长至 3 倍，当前累计融资额已超过 5.62 亿美元。

OLIX 成立于 2024 年 3 月，创始人 James Dacombe 年仅 25 岁，16 岁从高中辍学，18 岁创立脑监测初创公司 CoMind（累计融资约 1 亿美元），23 岁创立 OLIX，专注 AI 推理芯片设计。公司总部位于英国伦敦，同时在布里斯托尔、多伦多、奥斯汀和旧金山均设有办公室，员工总数超过 140 人，其中涵盖来自美国大型科技企业和半导体行业的核心人才。

公司致力于打造专用推理芯片，认为 AI 推理行业正从通用计算迈向专业化分工的新阶段，单一芯片处理所有任务的时代已经结束。基于这一理念，X-1 平台作为该公司的核心产品，将大模型推理任务拆解并分发至海量专用芯片并行运行，每颗芯片专职处理模型计算中的某一环节。在技术路线上，OLIX 做出了三项差异化选择：其一，摒弃当前供应紧张的 HBM 高带宽内存，改用片上 SRAM 保存模型数据，规避传统 GPU 架构在存储带宽和供应链上的双重约束；其二，舍弃传统铜线电气传输，改用光信号在芯片之间直接完成数据交互，以大量低速通道替代少数高速通道，降低纠错电路的复杂度和能耗；其三，采用整机柜交付模式，交付包含芯片、激光器、网络设备和软件在内的完整服务器机架，而非单张加速卡，客户接入电源即可使用。

与此同时，公司推出首款定位为大模型推理加速器的 DX-1 芯片。根据公司自行披露的架构模拟数据，该产品在千亿参数模型场景下单用户每秒可生成超过 1 万个 Token，多机柜扩展架构可适配十万亿参数级别的超大规模模型。然而，目前该性能指标基于内部周期精确和比特精确模拟结果，尚未经过真实硅片或独立第三方测试验证，公司计划于 2027 年下半年向首批客户交付产品。

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
OLIX	AI 推理芯片	Fundomo	B 轮	3.12 亿美元	超 5.62 亿美元	33 亿美元

仅供内部参考，请勿外传！

HappyRobot	AI 智能体	Prysm Capital、Eurazeo	C 轮	1.5 亿美元	超 2 亿美元	12 亿美元
Convex	AI 基础设施	Insight Partners	B 轮	5700 万美元	超 1 亿美元	未披露

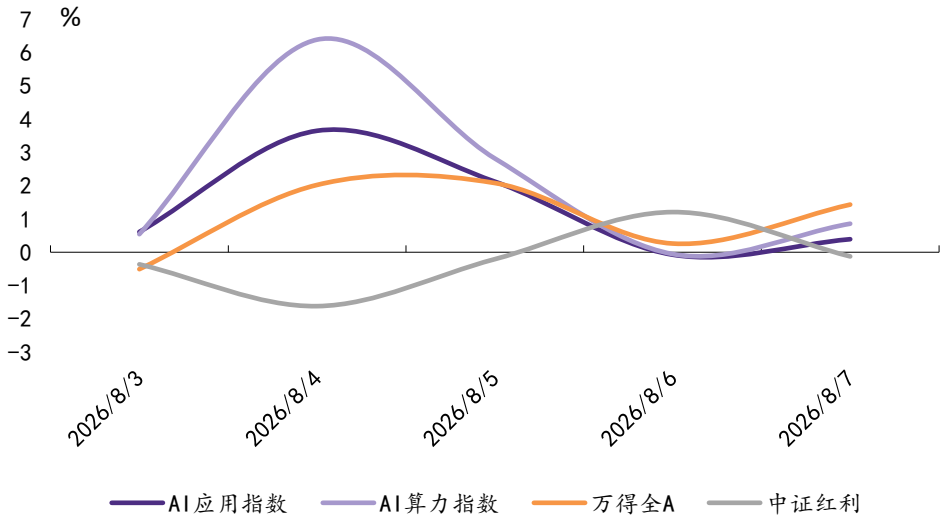
资料来源：wind，Saasverse，华鑫证券研究

仅供内部参考，请勿外传！

4、 行情复盘

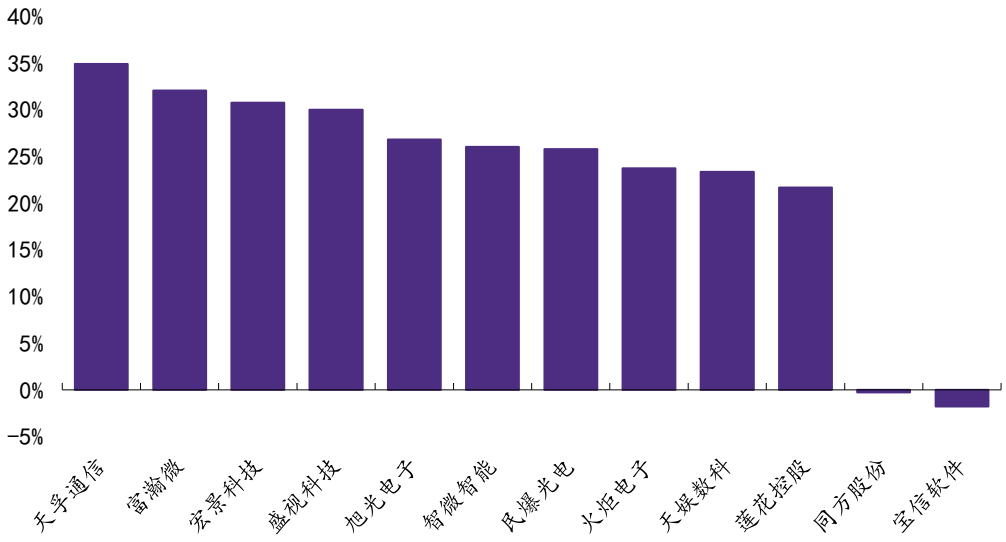
上周（2026.8.3-2026.8.7 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 3.65%/6.38%/2.08%/1.21%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-0.07%/-0.05%/-0.51%/-1.62%。AI 算力指数内部，天孚通信以 34.89%录得上周最大涨幅，宝信软件以-1.78%录得上周最大跌幅。AI 应用指数内部，胜宏科技以 47.16%录得上周最大涨幅，传音控股以-11.54%录得上周最大跌幅。

图表 14：上周（2026.8.3-2026.8.7 日）指数日涨跌幅



资料来源：wind，华鑫证券研究

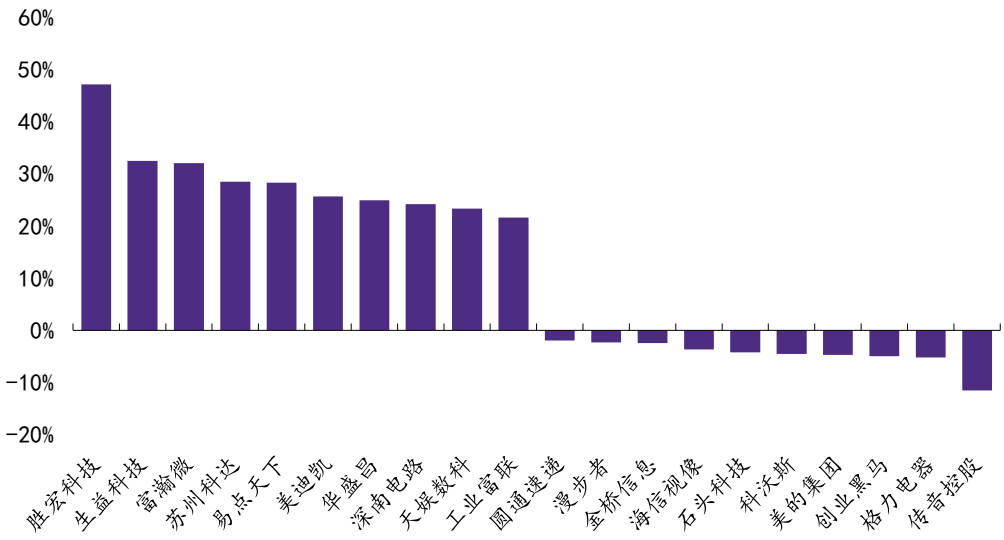
图表 15：上周（2026.8.3-2026.8.7 日）AI 算力指数内部涨跌幅度排名



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（2026.8.3-2026.8.7 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

2026 年 8 月 3 日，Palantir 发布 2026Q2 财报：当季实现总营收 19.35 亿美元，同比增长 93%、环比增长 19%；毛利率 85%，同比提升 4 个百分点；GAAP 净利润 10.62 亿美元（净利率 55%），摊薄后每股收益 0.41 美元。分业务看，商业业务收入 9.45 亿美元，同比增长 110%，其中美国商业收入 7.64 亿美元，同比增长 149%；政府业务收入 9.90 亿美元，同比增长 79%，其中美国政府收入同比增长 90%；美国区收入占总营收比重由上年同期的 73%升至 81%。订单层面，当季新签 1000 万美元以上合同 73 个；美国商业新签总合同价值（TCV）21.32 亿美元创历史新高，同比增长 153%；期末剩余履约义务（RPO）49 亿美元，约 43%预计于未来 12 个月内确认为收入。Non-GAAP 口径下，调整后营业利润 11.94 亿美元（利润率 62%），Ruleof40 得分达 155%。指引方面，公司将 2026 全年营收指引上调至 81.50-81.58 亿美元（对应约 82%增长），美国商业收入指引上调至 34.24 亿美元以上（对应至少约 134%增长）。

本次财报印证了企级 AI 应用正从试点验证期迈入规模化部署期，AI 软件侧的商业化兑现显著加速。从财报的结构层面看，商业收入的增速显著超越政府业务的增速，这表明公司增长引擎已由传统政府市场切换至企业级商业市场，需求结构更具延展性。从商业模式看，公司在收入接近翻倍的同时实现 85%的毛利率和 62%的调整后营业利润率，Ruleof40 得分达 155%，"高增长+高盈利"的组合打破了软件行业增长与利润率此消彼长的传统约束，印证 AI 应用具备极强的经营杠杆。指引层面，公司的全年营收指引上调幅度远超常规季度修正，印证需求仍在加速而非收敛；CEOKarp 提出的"AI 主权"需求与"将 Token 转化为实际经济价值"的表述，指向企业级 AI 竞争的焦点正从模型能力转向数据治理、决策闭环与行业 know-how 的工程化落地。我们认为，Palantir 作为企业级 AI 应用的风向标，其订单与收入的同步加速表明本轮 AI 产业景气正向下游的应用传导，AI 应用侧逐步进入商业化兑现的验证窗口；对国内市场而言，具备行业数据积累、工程化交付能力与政企客户资源的 AI 应用及数据要素相关公司有望承接相似的需求逻辑，配置价值逐步显现。

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

仅供内部参考，请勿外传！

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	2026-08-10	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	488.18	-0.30	0.30	0.60	-1627.27	1627.27	813.63	买入
301196.SZ	唯科科技	82.67	2.53	3.34	3.98	32.68	24.75	20.77	买入
301369.SZ	联动科技	137.20	0.48	1.08	2.62	285.83	127.04	52.37	买入
603859.SH	能科科技	43.03	0.92	1.05	1.32	46.77	40.98	32.60	买入
688615.SH	合合信息	121.67	3.24	4.22	5.25	37.55	28.83	23.18	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。