仅供内部参考，请勿外传

证券研究报告

2026 年 08 月 24 日

华鑫证券 CHINA FORTUNE SECURITIES

DeepSeek-V4-Flash-Vision-Exp 上线，Codex Harness 正式开源

—计算机行业周报

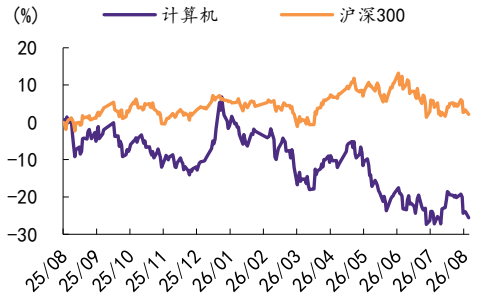
推荐(维持)

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)	2.3	-16.9	-24.5
沪深 300	-1.9	-5.8	4.2

市场表现



资料来源：Wind，华鑫证券研究

相关研究

1、《计算机行业点评报告：Shopify（SHOP）：Q2 营收与利润高增，AI 商业引擎加速激活长尾商业》2026-08-24

2、《计算机行业点评报告：Spotify（SPOT）：营收利润双增长，付费用户突破 3 亿里程碑》2026-08-20

3、《计算机行业点评报告：Airbnb（ABNB.O）：Q2 业绩保持增长，AI 提效与新兴市场双引擎共振》2026-08-19

投资要点

算力：全新多模态视觉理解模型 DeepSeek-V4-Flash-Vision-Exp 上线，DeepSeek Harness 同步更新

2026 年 8 月 21 日，DeepSeek 宣布全新多模态视觉理解模型 DeepSeek-V4-Flash-Vision-Exp 正式上线 DeepSeek API 平台。同日，DeepSeek Harness 进行版本更新，发布新版本 DeepSeek Harness v0.1.1-rc.1，为多模态更新提供官方支持。

AI 应用：QuillBot 周访问量环比+5.36%，Codex Harness 正式开源

2026 年 8 月 20 日，OpenAI 选择将驱动代码模型 Codex 的底层核心框架——Harness，作为开源平台全面向公众开放。开发者可借助 Harness 直接构建个性化产品，实现对界面、数据及权限审批的掌控，而 AI 则在底层承担起高效执行的支撑角色。

AI 融资动向：Etched 完成 7 亿美元 D 轮融资，投后估值达 210 亿美元

当地时间 2026 年 8 月 18 日，AI 推理芯片公司 Etched 宣布完成新一轮 7 亿美元融资，投后估值达到 210 亿美元。本轮融资由量化交易巨头 Jane Street 领投，Kleiner Perkins、Sequoia、Andreessen Horowitz、Tiger Global、Bain Capital Ventures、Neo、Primary、Stripes、Positive Sum 及 Blackstone 等多家知名机构共同参投。此前，该公司已于 2025 年底创下 50 亿美元估值，并于 2026 年 7 月以 103 亿美元估值完成 3 亿美元 C 轮融资，不到一年的时间里，数次完成估值跃迁。截至目前，公司累计融资额已达 19 亿美元。

投资建议

近日，阿里巴巴发布 2027 财年第一季度（截至 2026 年 6 月 30 日）财报。集团当季收入 2,689.53 亿元，同比增长 9%；新口径"AI 云与算力服务"（云智能与平头哥整合）收入 484.37 亿元，总收入及外部客户收入同比均增长 45%，其中 AI 相关产品收入 123.76 亿元，连续 12 个季度实现三位数同比增长，分部经调整 EBITA 利润率 12%；资本开支 676.78 亿元，同比增长 75%（上年同期 386.76 亿元），公司归因于 AI 基础设施

行业研究

仅供内部参考，请勿外传

投入，当季自由现金流净流出 446.70 亿元，期末现金及其他流动投资 4,745.05 亿元。财报发布后的 8 月 23 日，公司宣布拟向美国境外非美国人士配售新股，募资总额 800 亿港元，为 2019 年港股上市以来首次配售，配售净额 100%投向全栈 AI 能力与 AI 基础设施建设。

本次财报的核心信号在于投入与产出同时加速：收入端，云业务 45%的增速较此前几个季度显著抬升，AI 产品收入连续 12 个季度三位数增长，"云收入加速—资本开支加码—AI 收入放量"的正循环首次清晰闭合；投入端，单季 capex 同比+75%，财报后公司随即启动 800 亿港元配售并全额投向 AI，彰显全力投入 AI 的战略决心。对产业链而言，国内头部云厂商资本开支的加速与海外云厂商形成互证，国内算力需求的能见度与持续性进一步增强。我们认为，阿里巴巴财报与配售动作相互印证，国产算力基础设施投入正进入新一轮加速期，建议关注国产算力产业链中 AI 芯片、AI 服务器等环节的投资机会。

中长期，建议关注专注于功率半导体及模拟芯片测试系统研发的联动科技（301369.SZ）、专注于半导体等高端制造业的罗博特科（300757.SZ）、新能源业务高增并供货科尔摩根等全球电机巨头的唯科科技（301196.SZ）、AI 智能文字识别与商业大数据领域巨头的合合信息（688615.SH）、深耕工业 AI 与软件并长期服务高端装备等领域头部客户的能科科技（603859.SH）。

风险提示

1）AI 底层技术迭代速度不及预期。2）政策监管及版权风险。3）AI 应用落地效果不及预期。4）推荐公司业绩不及预期风险。

重点关注公司及盈利预测

公司代码	名称	2026-08-24 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	496.40	-0.30	0.30	0.60	-1654.67	1654.67	827.33	买入
301196.SZ	唯科科技	88.98	2.53	3.34	3.98	35.17	26.64	22.36	买入
301369.SZ	联动科技	131.09	0.48	1.08	2.62	273.10	121.38	50.03	买入
603859.SH	能科科技	37.20	0.92	1.05	1.32	40.43	35.43	28.18	买入
688615.SH	合合信息	89.73	3.24	4.22	5.25	27.69	21.26	17.09	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：DEEPSEEK-V4-FLASH-VISION-EXP 上线，DEEPSEEK HARNESS 同步更新 4

1.1、 Tokens 跟踪 4

1.2、 云厂动态：阿里云韩国第三座数据中心正式上线 6

1.3、 产业动态：DeepSeek 上线全新多模态视觉理解模型 DeepSeek-V4-Flash-Vision-Exp，DeepSeek Harness 同步更新 7

2、 AI 应用动态：QUILLBOT 周访问量环比+5.36%，CODEX HARNESS 正式开源 9

2.1、 周流量跟踪：QuillBot 周访问量环比+5.36% 9

2.2、 产业动态：Codex Harness 正式开源，AI 嵌入业务场景迈向新高度 9

3、 AI 融资动向：ETCHED 完成 7 亿美元 D 轮融资，投后估值达 210 亿美元 12

4、 行情复盘 14

5、 投资建议 16

6、 风险提示 17

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） 6

图表 3：DEEPSEEK-V4-FLASH-VISION-EXP 基准测试结果横向对比图 7

图表 4：DEEPSEEK-V4-FLASH-VISION-EXP 定价对比图 8

图表 5：2026.8.15-2026.8.21 AI 相关网站流量 9

图表 6：OPENAI 总裁 GREG BROCKMAN 对 CODEX HARNESS 的评价 10

图表 7：CODEX APP-SERVER 11

图表 8：TAX AI 的完成度在报税季持续加速提升 11

图表 9：上周 AI 初创公司融资动态 13

图表 10：上周（2026.8.17-2026.8.21）AI 指数日涨跌幅 14

图表 11：上周（2026.8.17-2026.8.21 日）AI 算力指数内部涨跌幅度排名 14

图表 12：上周（2026.8.17-2026.8.21 日）AI 应用指数内部涨跌幅度排名 15

图表 13：FICONTEC2025 年年中至今公告订单 16

图表 14：重点关注公司及盈利预测 17



仅供内部参考，请勿外传

1、算力动态：DeepSeek-V4-Flash-Vision-Exp 上线，DeepSeek Harness 同步更新

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 8 月 17 日至 8 月 23 日，周度 Token 消耗量显著上涨，消耗量为 93.4T，环比上周增加 24.04%。在 Tokens 规模 Leaderboard 前五名中，DeepSeek 的 DeepSeek V4 Flash 0731 以 11.6T tokens 位列榜首；Stealth 的 Ox Alpha 以 11.6T tokens 位居第二；Xiaomi 的 MiMo-V2.5 以 9.94T tokens 位居第三；Tencent 的 Hy3 以 8.21T tokens 位居第四；DeepSeek 的 DeepSeek V4 Flash 0423 以 5.46T tokens 位居第五。

文本类模型的市场份额，DeepSeek 以 1.13B requests 占据 26.3%，位居第一；Google 以 932M requests 占据 21.8%的份额，位列第二；OpenAI、Qwen、Stealth 则分别以 814M、223M、146M requests，对应占据 19.0%、5.2%、3.4%的市场份额。

图像生成类模型调用量略微下降，调用量为 5.73M requests，环比上周减少 0.87%。在 Image Leaderboard 中，Google 的 Nano Banana 系列包揽榜单前四名席位，其中，Nano Banana 2 (Gemini 3.1 Flash Image)以 1.38M requests 位居第一；Nano Banana (Gemini 2.5 Flash Image) 、Banana 2 Lite (Gemini 3.1 Flash Lite Image)、Nano Banana 2 (Gemini 3.1 Flash Image Preview)分别以 830K、714K、557K requests 位居第二、第三和第四；OpenAI 的 GPT Image 2 以 460K requests 位列第五。从市场份额来看，Google 以 4.2M requests 占据 73.3%的份额，位居第一；OpenAI 以 705K requests 占据 12.3%，位列第二；Black Forest Labs、xAI、ByteDance Seed 则分别以 214K、196K、188K requests，对应占据 3.7%、3.4%、3.3%的市场份额。

嵌入模型调用量有所上涨，调用量为 301M requests，环比上周增加 5.99%。在 Embedding Leaderboard 前五名中，OpenAI 的 Text Embedding 3 Small 以 114M requests 位居榜首；Qwen 的 Qwen3 Embedding 8B 以 82.3M requests 位列第二；Google 的 Gemini Embedding 2 以 18.3M requests 位列第三；OpenAI 的 Text Embedding 3 Large 以 18M requests 位列第四；BAAI 的 bge-m3 以 17.2M requests 位列第五。从市场份额来看，OpenAI 以 136M requests 占据 45.3%的份额，位居第一；Qwen 以 85.6M requests 占据 28.5%，位列第二；Google、BAAI、Perplexity 则分别以 34M、17.9M、8.57M requests，对应占据 11.3%、5.9%、2.8%的市场份额。

重排序模型调用量小幅回落，调用量为 6.17M requests，环比上周减少 1.60%。在 Rerank Leaderboard 中，VoyageAI 的 rerank-2.5 以 2.63M requests 位列榜首；NVIDIA 的 Llama Nemotron Rerank VL 1B V2 以 1.56M requests 位居第二；Cohere 的 Rerank v3.5、Rerank 4 Pro、Rerank 4 Fast 分别以 861K、321K、283K requests 位列第三、第四和第五。从市场份额来看，VoyageAI 以 2.89M requests 占据 46.8%的份额，位居第一；NVIDIA 以 1.56M requests 占据 25.2%的份额，位居第二；Cohere、Qwen 分别以 1.46M、262K requests 占据 23.7%、4.3%的份额。

视频生成类模型调用量大幅下降，调用量为 156K requests，环比上周减少 9.83%。在 Video Leaderboard 前五名中，ByteDance 的 Seedance 2.0 Mini 以 26K requests 位列第一；

仅供内部参考，请勿外传

Google 的 Veo 3.1 Lite 以 23K requests 位列第二；ByteDance 的 Seedance 2.0 Fast、Seedance 2.0、Seedance 2.5 分别以 17K、13K、12K requests 位居第三、第四和第五。从市场份额来看，ByteDance 以 79K requests 占据 50.3%的份额，位列第一；Google 以 37K requests 占据 24.0%的份额，位居第二；xAI、KwaiVGI、Alibaba 则分别以 13K、10K、7K requests，对应占据 8.5%、6.2%、4.6%的市场份额。

语音生成类模型调用量大幅上涨，调用量为 4.67M requests，环比上周增加 74.25%。在 Speech Leaderboard 前五名中，Fish Audio 的 S2.1 Pro Free (free)以 3.24M requests 位居第一；Google 的 Gemini 3.1 Flash TTS Preview 以 614K requests 位列第二；Hexgrad 的 Kokoro 82M 以 368K requests 位列第三；xAI 的 Grok Voice TTS 1.0 以 196K requests 位列第四；Fish Audio 的 S2.1 Pro 以 74K requests 位列第五。从市场份额来看，Fish Audio 以 3.31M requests 占据 70.9%的份额，位居第一；Google 以 614K requests 占据 13.1%，位列第二；Hexgrad、xAI、Microsoft 则分别以 368K、196K、73K requests，对应占据 7.9%、4.2%、1.6%的市场份额。

语音转写类模型调用量有所上涨，调用量为 12.2M requests，环比上周增加 2.52%。在 Transcription Leaderboard 前五名中，OpenAI 的 Whisper Large V3、Whisper Large V3 Turbo、GPT-4o Mini Transcribe 分别以 6.78M、3.06M、560K requests 位列第一、第二和第三；MistralAI 的 Voxtral Mini Transcribe 以 327K requests 位列第四；OpenAI 的 Whisper 1 以 264K requests 位列第五。从市场份额来看，OpenAI 以 10.9M requests 占据 89.2%的份额，位居第一；MistralAI 以 341K requests 占据 2.8%，位列第二；Qwen、Deepgram、Microsoft 则分别以 293K、220K、149K requests，对应占据 2.4%、1.8%、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。

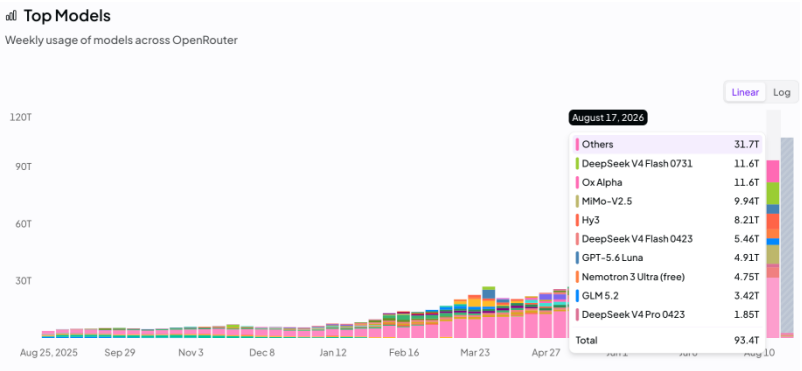
8 月 18 日，百度公布 2026 年第二季度财报。财报显示，百度核心 AI 驱动业务营收达 125 亿元，且 AI 业务收入已连续两个季度在一般性业务收入中占比达到甚至超过 50%。细分来看，在百度智能云板块内部，GPU 云业务营收同比增长 283%，已连续四个季度实现三位数增长；此外，百度千帆大模型服务及 Agent 平台来自外部客户的 Token 使用收入同比增长超 9 倍。

8 月 20 日，中国平安发布 2026 年中期业绩，首次将 Token 消耗量写入财报。中报显示，公司日均 Token 消耗量从 2025 年 12 月的 300 亿跃升至 2026 年 6 月的超 1200 亿，半年增长至 4 倍。指标背后透露出关键信号：AI 正从概念展示走向规模化应用，从少数场景试用拓展至越来越多的真实业务流程。

8 月 21 日，青岛自贸片区与浪潮云签署战略合作协议，双方将围绕数据跨境流动、算力出海及大模型服务出海等领域开展深度协同，共同搭建北方首个跨境 Token 工厂。该项目旨在打通“算力调用—词元计费—跨境结算—多方生态分润”的完整商业闭环，降低海外客户、出海企业及 OPC 创作者的接入门槛，提供标准化、低成本、高合规的算力词元服务。

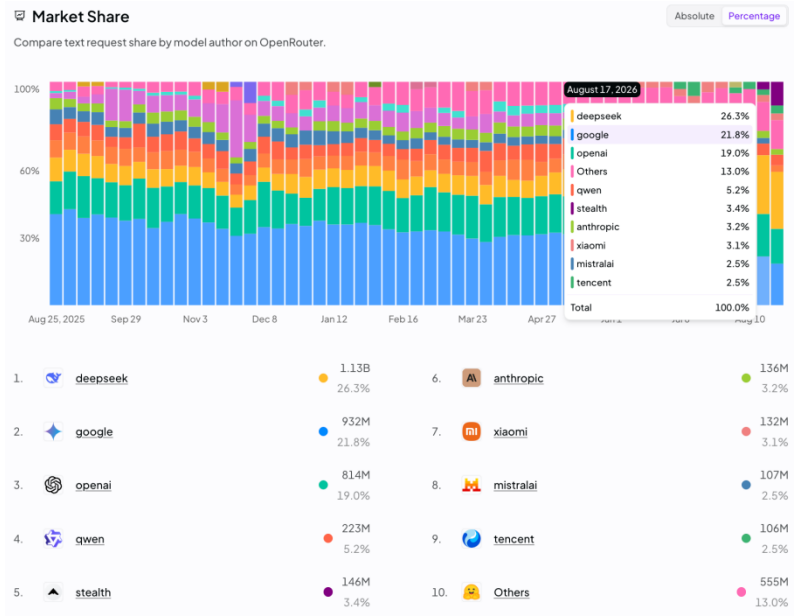
仅供内部参考，请勿外传

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、 云厂动态：阿里云韩国第三座数据中心正式上线

2026 年 8 月 18 日，阿里云宣布其位于韩国的第三座数据中心正式上线运营。此举是阿里云在韩国市场持续加大基础设施投入的重要体现，反映出当地企业数字化转型步伐加快以及重点行业对云服务需求的持续攀升。

当前，AI 已成为韩国云市场的重要增长引擎。随着 AI 应用加速落地，企业对稳定、弹性、安全的云基础设施需求日益增长。此次上线的数据中心旨在覆盖智能体开发、部署、运营及安全管理等多元需求，不仅提供计算、存储、网络、安全、数据库、容器及云原生等企业级服务，还同步布局 AgentRun、STAROps 等六项 Agentic AI 服务，从而满足不断扩大的市场需求。

在生态建设方面，阿里云持续通过开源模型推动 AI 开发者生态的扩展。截至目前，Qwen 模型家族全球累计下载量已突破 20 亿次，开发者基于 Qwen 衍生的模型数量超过 30 万

仅供内部参考，请勿外传

个。相关 AI 能力已在韩国内容、电商、互联网等多个行业的企业中实现落地。

阿里云 AI 能力已在韩国多家企业中实现落地应用。总部位于光州的 AI 公司 Gendive 已将 Wan、HappyHorse 等阿里云视频生成模型集成至其 Agentic WorkOS 产品 DevDive 中，帮助企业在低代码甚至无代码环境下完成文档分析、视觉及多媒体内容生成等自动化工作流。此外，Naver 旗下 Naver Z 运营的 Zepeto 平台，也借助阿里云基础设施支持亚洲市场的实时 3D 虚拟形象生成与沉浸式互动体验。

此次对韩国市场的持续投入，是阿里云全球云与 AI 基础设施战略布局的重要组成部分。近期，阿里云已在欧洲、东亚、东南亚、拉美等区域持续完善云基础设施网络，致力于为全球客户提供更加稳定、可扩展的云与 AI 服务能力。

1.3、产业动态：DeepSeek 上线全新多模态视觉理解模型 DeepSeek-V4-Flash-Vision-Exp，DeepSeek Harness 同步更新

2026 年 8 月 21 日，DeepSeek 宣布全新多模态视觉理解模型 DeepSeek-V4-Flash-Vision-Exp 正式上线 DeepSeek API 平台。同日，DeepSeek Harness 进行版本更新，发布新版本 DeepSeek Harness v0.1.1-rc.1，为多模态更新提供官方支持。

模型方面，DeepSeek-V4-Flash-Vision-Exp 为实验性质模型。基准测试结果显示，该模型在涵盖 Agent、推理、世界知识等方面的纯文本能力上与 DeepSeek-V4-Flash 正式版持平。与此同时，相较于正式版，新模型在需要视觉理解的 Agent Benchmark 上显示出了大幅度的能力跃升。此外，该模型的多模态 Agent 能力也已同 Opus-4.8 相接近。

图表 3：DeepSeek-V4-Flash-Vision-Exp 基准测试结果横向对比图

Benchmark		DeepSeek V4-Flash-Vision-Exp	DeepSeek V4-Flash-0731	Opus-4.8
文本 Agent 能力评测*	Terminal Bench 2.1	83.9	82.7	85.0
	NL2Repo	57.7	54.2	69.7
	Cybergym	75.3	76.7	78.3
	DeepSWE	59.3	54.4	58.0
	Toolathlon-Verified	75.9	70.3	76.2
	DSBench-Hard	63.6	59.6	71.7
	AutomationBench (Public)	25.7	25.1	27.2
多模态 Agent 能力评测	ApexBench (Pass@1)	36.5	26.2**	39.4
	Agents' Last Exam	27.3	25.2**	25.7
	Chartography	64.3	-	65.0
	ZeroBench (Pass@5)	35.0	-	34.0

资料来源：智东西，华鑫证券研究

仅供内部参考，请勿外传

场景应用方面，V4-Flash-Vision-Exp 在各类 Agent 框架内展现出了较强的多模态适配能力。例如，用户可利用该模型制作 PPT 内容，模型能够理解真实摄影图片风格、野性、原始、探索感等较为抽象的风格化要求。该模型还可用于视觉编程任务，如在 DeepSeek Harness 官网中加入黑蓝深海、玻璃 UI 和 ASCII 原子像素等视觉要素，完成对官网的二次创作。

调用方式上，该多模态 API 同时兼容 Chat Completions、Messages、Responses 三种格式，支持图文混合输入，方便接入各类 Agent 工具。图片传入支持 base64 内联、外部 URL 和 Files API 三种方式。其中，Files API 为 DeepSeek 同步上线的免费服务，同一张图片无需重复上传，可有效节省重复传输的开销。

计费方面，图片会先转换为 Token 再计入用量，单张图片最多按 384 Tokens 计算，价格与 V4-Flash 模型持平。

图表 4：DeepSeek-V4-Flash-Vision-Exp 定价对比图

模型			deepseek-v4-flash	deepseek-v4-pro	deepseek-v4-flash-vision-exp
价格[1][2]	百万tokens输入（缓存命中）	空闲时段	0.05元	0.15元	0.05元
		高峰时段	0.10元	0.30元	0.10元
	百万tokens输入（缓存未命中）	空闲时段	1.5元	4.5元	1.5元
		高峰时段	3.0元	9.0元	3.0元
	百万tokens输出	空闲时段	4.5元	13.5元	4.5元
		高峰时段	9.0元	27.0元	9.0元
并发限制[3]			2500	500	2500

资料来源：智东西，华鑫证券研究

官方支持方面，日前，DeepSeek Harness 已完成公测后的首次重要更新，正式发布 v0.1.0-rc.8 版本。此次更新共带来 14 项调整，覆盖多模态输入、子代理协作、终端体验、工具调用和开发者支持等多个方向，既补齐了 Agent 处理图片等多模态任务的能力，也进一步完善了子代理协作、终端交互和工具调用体验。同日，DeepSeek Harness 再度完成更新，正式发布 v0.1.1-rc.1 版本，在 DeepSeek 模型适配器中新增 DeepSeek-V4-Flash-Vision-Exp 多模态模型选项，并支持配置原生图片请求，多模态能力进一步增强。

仅供内部参考，请勿外传

2、AI 应用动态：QuillBot 周访问量环比+5.36%，Codex Harness 正式开源

2.1、周流量跟踪：QuillBot 周访问量环比+5.36%

本期（2026.8.15-2026.8.21）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1290.0M）、Bing（860.3M）和 Gemini（597.7M），访问量环比增速第一为 QuillBot（5.36%）；平均停留时长前三位分别为 Character.AI（00:16:12）、Discord（00:11:04）和 NotionAI（00:08:30）;平均停留时长环比增速第一为文心一言（41.18%）。

图表 5：2026.8.15-2026.8.21 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	1290.0	2.14%	6:37	0.76%
Bing	搜索	微软	860.3	3.36%	7:22	0.00%
Gemini	聊天机器人	谷歌	597.7	3.23%	6:57	-0.24%
Canva	在线设计	Canva	196.5	2.24%	5:56	0.00%
Discord	游戏社区	微软	149.4	-2.10%	11:04	-0.60%
Github	代码托管	微软	147.7	0.75%	6:26	0.26%
Character.AI	聊天机器人	Character.AI	35.53	-2.42%	16:12	-0.82%
DeepI	翻译工具	DeepL	22.53	0.36%	2:28	0.68%
Perplexity	AI 搜索	Perplexity	21.47	-1.78%	4:25	0.00%
Kimi	聊天机器人	Moonshot AI	10.86	-19.32%	7:04	-0.70%
QuillBot	释义工具	QuillBot	7.27	6%	2:51	0.00%
NotionAI	文本/笔记	Notion	2.589	0.00%	8:30	-3.41%
文心一言	聊天机器人	百度	0.00	-96.67%	3:12	41.18%

资料来源：similarweb,华鑫证券研究

2.2、产业动态：Codex Harness 正式开源，AI 嵌入业务场景迈向新高度

2026 年 8 月 20 日，OpenAI 选择将驱动代码模型 Codex 的底层核心框架——Harness，作为开源平台全面向公众开放。开发者可借助 Harness 直接构建个性化产品，实现对界面、数据及权限审批的掌控，而 AI 则在底层承担起高效执行的支撑角色。

仅供内部参考，请勿外传

图表 6：OpenAI 总裁 Greg Brockman 对 Codex Harness 的评价



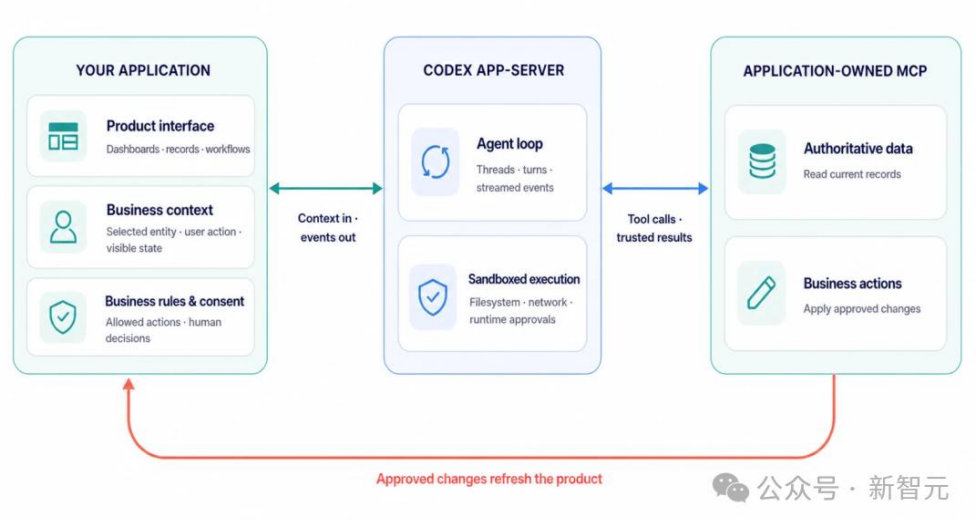
资料来源：新智元，华鑫证券研究

作为此次开源的焦点——Harness 框架，能够显著提升模型的实际表现。以高难度基准测试 ARC-AGI-3 为例，通过仅对 Harness 实施两项关键调整——即保留推理能力与上下文压缩，GPT-5.6 Sol 模型的得分实现了从 13.3%至 38.3%的跃升，同时输出 Token 数量大幅缩减至原来的六分之一。

为便于开发者实现快速集成，OpenAI 依据 Apache-2.0 许可协议，一次性发布了三大开源组件。其中，CLI 工具“codex exec”适用于运行自动化流水线，能够执行有边界的 Agent 工作流并返回结构化输出，满足脚本运行、持续集成任务等一次性后台需求。其次，官方 Codex SDK 为应用程序开发者提供了编程式接口，支持 TypeScript 与 Python 语言，允许开发者通过代码精确控制线程与任务的生命周期，实现任务的启动、恢复或流式传输。最为核心的组件当属“Codex app-server”，它是将 Agent 深度融入产品的关键引擎。该服务器通过记录详细的 JSON-RPC 客户端协议，使应用得以连接至本地 Codex 进程，从而支持持久对话状态维护、事件流式传输、任务中断、自定义工具暴露以及人机交互审批等功能。这种将前端审批界面与底层执行循环分离的设计，显著简化了 Agent 嵌入业务仪表盘的流程。

仅供内部参考，请勿外传

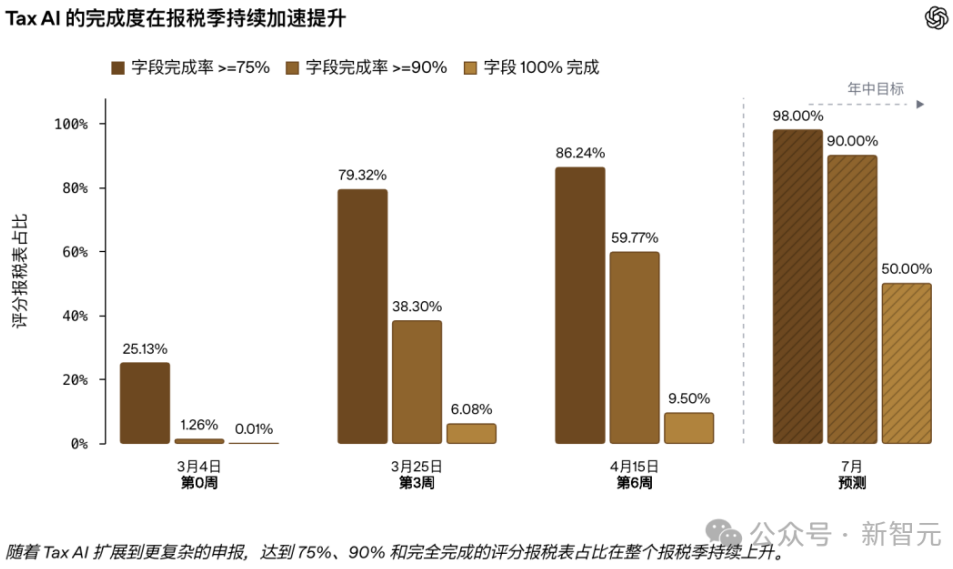
图表 7：Codex app-server



资料来源：新智元，华鑫证券研究

值得注意的是，Codex Harness 的应用场景已突破代码生成领域。在税务行业，合作伙伴 Thrive Holdings 与 Crete 将其嵌入税务准备工作流，系统在处理复杂税务逻辑的同时融合税务师反馈，成功处理了约 7000 份申报表，将税务准备时间缩减约三分之一，展现了 AI 在垂直专业工具中的巨大价值。科技巨头 Cisco 则利用 Codex SDK 在其云控制平台内构建了 App Builder，使客户能够通过自然语言创建自定义应用程序，底层逻辑流转由 Harness 自动处理。此外，OpenAI 官方展示了名为“Relay”的虚拟物流运营看板应用，生动呈现了人机协同的工作模式：用户无需编写提示词，仅需选中延误货单并选择操作，应用即可自动将界面数据与物流信息作为上下文传递至 AI，Codex 随后调用自有 MCP 工具获取实时数据，分析生成最优方案后提交人工审批，待业务员确认后方可执行写入操作，最终实现看板状态的自动同步，构建起真正意义上的人机协同体系。

图表 8：Tax AI 的完成度在报税季持续加速提升



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：Etched 完成 7 亿美元 D 轮融资，投后估值达 210 亿美元

当地时间 2026 年 8 月 18 日，AI 推理芯片公司 Etched 宣布完成新一轮 7 亿美元融资，投后估值达到 210 亿美元。本轮融资由量化交易巨头 Jane Street 领投，Kleiner Perkins、Sequoia、Andreessen Horowitz、Tiger Global、Bain Capital Ventures、Neo、Primary、Stripes、Positive Sum 及 Blackstone 等多家知名机构共同参投。此前，该公司已于 2025 年底创下 50 亿美元估值，并于 2026 年 7 月以 103 亿美元估值完成 3 亿美元 C 轮融资，不到一年的时间里，数次完成估值跃迁。截至目前，公司累计融资额已达 19 亿美元。

该公司致力于打造可支持多种前沿模型架构的推理系统，而非单一模型专用芯片，核心产品为 Frontier Inference Clusters，即前沿推理集群。公司认为，在大模型推理时代，客户真正购买的是系统能力，尤其是在高并发、低延迟、长上下文、多模型混合、MoE 架构和 Agent 工作流等场景下，单颗芯片的理论性能并不足以决定最终效果，因而，Etched 选择以整机机架形式交付产品，其希望交付的不是一颗孤立芯片，而是一套完整系统，包括芯片、互联、内存、散热、供电、软件和机架级部署能力。

在技术路线上，该公司的推理性能主要来自低电压推理（Low Voltage Inference，LVI）和集群规模内存（Cluster Scale Memory，CSM）两项关键技术。大模型推理包含预填充和解码两个核心阶段，其中预填充阶段计算密集度较高，对矩阵运算、注意力机制和吞吐能力要求较高，Etched 的 LVI 技术主要服务于这一阶段，其数学运算模块运行电压低于主流 AI 加速器的一半，能够显著降低芯片工作温度，从而支持芯片在高负载条件下以更高密度运行计算单元，并在同等功耗约束下提供更高吞吐。CSM 则主要面向解码阶段，Etched 的 CSM 不是只在单颗芯片或单个节点内部做内存优化，而是试图在整个集群层面构建一个高效率的大规模共享内存池，允许多块芯片互联并共享同一内存池，同时保持极低延迟和高速通信。

除了 LVI 和 CSM，Etched 还在机架层面集成了多项自研组件，包括自研互联技术、自研冷板、专有电压调节模块以及面向整机供电和散热效率的系统级优化，其互联方案在某些通信任务上仅需约 700 毫秒，而竞品方案通常需要约 4000 毫秒。

在商业化进展方面，Etched 自种子轮融资起不到三年便实现了首次流片成功，其首款原型芯片已在 TSMC 生产线上下线，并成功完成早期验证。2026 年 6 月，该公司以超过 400 人团队规模和可运行芯片正式亮相，首个完整推理集群在约 40 天内完成启动部署，成功运行多款前沿 AI 大模型，并于 7 月向第一位正式商业客户 Jane Street 完成首批机架交付。与此同时，公司正在同步推进三代硬件的并行研发，试图通过更快的产品周期追赶甚至改变 AI 推理硬件的竞争节奏。

仅供内部参考，请勿外传

图表 9：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
Etched	AI 推理芯片	Jane Street	D 轮	7 亿美元	19 亿美元	210 亿美元
Higgsfield	AI 视频与图像生成	DST Global	B 轮	4 亿美元	超 4.8 亿美元	54 亿美元
Wispr Flow	AI 语音输入	Menlo Ventures	B 轮	2.8 亿美元	3.61 亿美元	20 亿美元

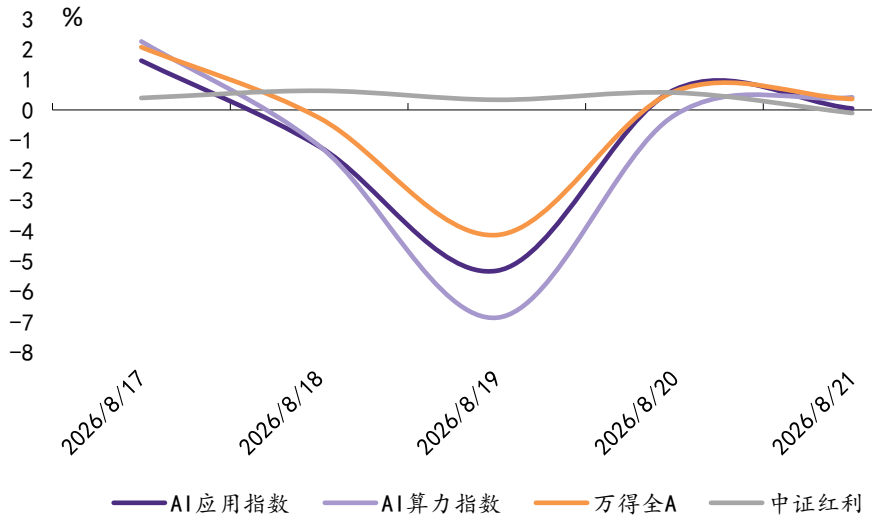
资料来源：wind，Saasverse，华鑫证券研究

仅供内部参考，请勿外传

4、 行情复盘

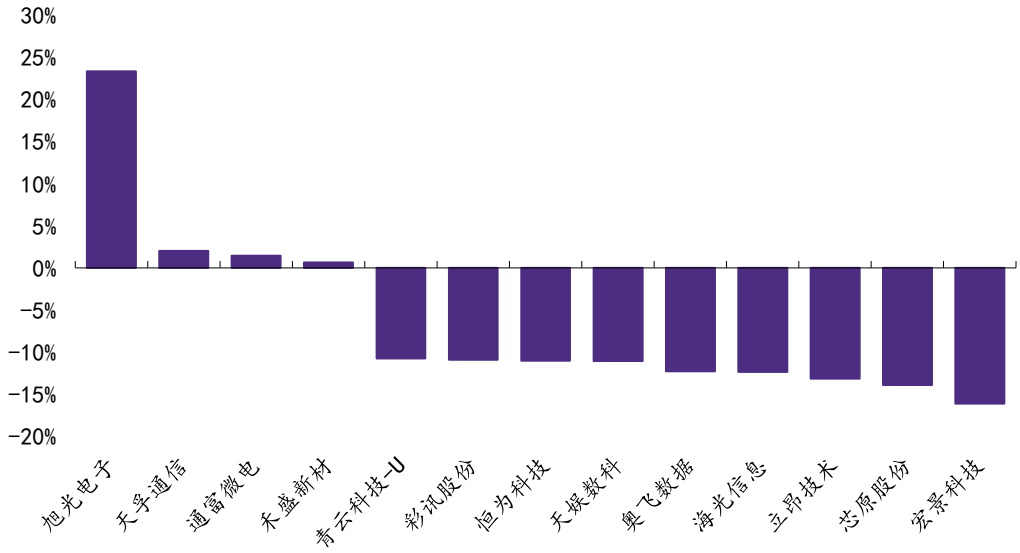
上周（2026.8.17-2026.8.21 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 1.63%/2.26%/2.07%/0.63%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-5.31%/-6.86%/-4.13%/-0.1%。AI 算力指数内部，旭光电子以 23.36%录得上周最大涨幅，宏景科技以-16.14%录得上周最大跌幅。AI 应用指数内部，科沃斯以 9.39%录得上周最大涨幅，恒银科技以-14.62%录得上周最大跌幅。

图表 10：上周（2026.8.17-2026.8.21 日）指数日涨跌幅



资料来源：wind，华鑫证券研究

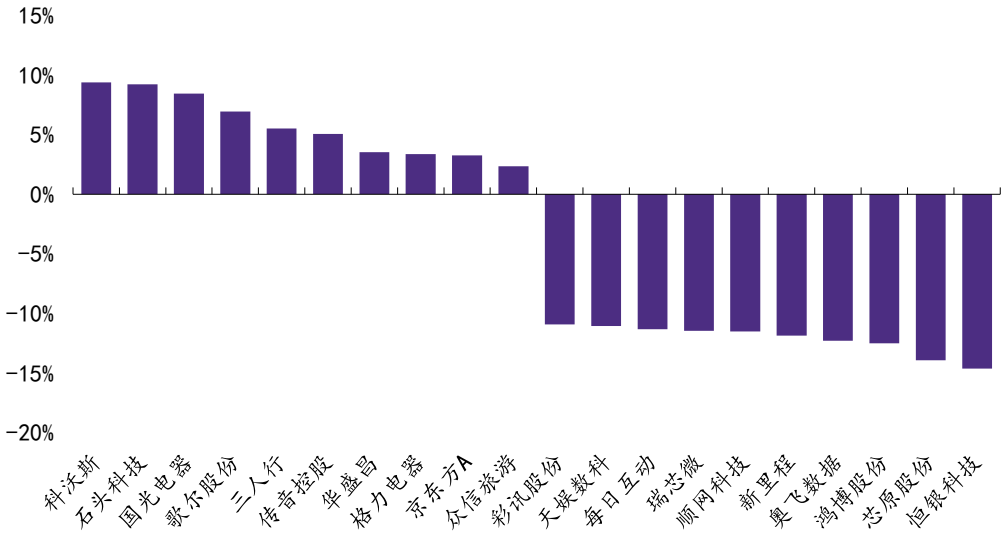
图表 11：上周（2026.8.17-2026.8.21 日）AI 算力指数内部涨跌幅度排名



资料来源：wind，华鑫证券研究

仅供内部参考，请勿外传

图表 12：上周（2026.8.17-2026.8.21 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

近日，阿里巴巴发布 2027 财年第一季度（截至 2026 年 6 月 30 日）财报。集团当季收入 2,689.53 亿元，同比增长 9%；新口径"AI 云与算力服务"（云智能与平头哥整合）收入 484.37 亿元，总收入及外部客户收入同比均增长 45%，其中 AI 相关产品收入 123.76 亿元，连续 12 个季度实现三位数同比增长，分部经调整 EBITA 利润率 12%；资本开支 676.78 亿元，同比增长 75%（上年同期 386.76 亿元），公司归因于 AI 基础设施投入，当季自由现金流净流出 446.70 亿元，期末现金及其他流动投资 4,745.05 亿元。财报发布后的 8 月 23 日，公司宣布拟向美国境外非美国人士配售新股，募资总额 800 亿港元，为 2019 年港股上市以来首次配售，配售净额 100%投向全栈 AI 能力与 AI 基础设施建设。

本次财报的核心信号在于投入与产出同时加速：收入端，云业务 45%的增速较此前几个季度显著抬升，AI 产品收入连续 12 个季度三位数增长，"云收入加速—资本开支加码—AI 收入放量"的正循环首次清晰闭合；投入端，单季 capex 同比+75%，财报后公司随即启动 800 亿港元配售并全额投向 AI，彰显全力投入 AI 的战略决心。对产业链而言，国内头部云厂商资本开支的加速与海外云厂商形成互证，国内算力需求的能见度与持续性进一步增强。我们认为，阿里巴巴财报与配售动作相互印证，国产算力基础设施投入正进入新一轮加速期，建议关注国产算力产业链中 AI 芯片、AI 服务器等环节的投资机会。

中长期，建议关注专注于功率半导体及模拟芯片测试系统研发的联动科技（301369.SZ）、专注于半导体等高端制造业的罗博特科（300757.SZ）、新能源业务高增并供货科尔摩根等全球电机巨头的唯科科技（301196.SZ）、AI 智能文字识别与商业大数据领域巨头的合合信息（688615.SH）、深耕工业 AI 与软件并长期服务高端装备等领域头部客户的能科科技（603859.SH）。

图表 13：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

仅供内部参考，请勿外传

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 14：重点关注公司及盈利预测

公司代码	名称	2026-08-24	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	496.40	-0.30	0.30	0.60	-1654.67	1654.67	827.33	买入
301196.SZ	唯科科技	88.98	2.53	3.34	3.98	35.17	26.64	22.36	买入
301369.SZ	联动科技	131.09	0.48	1.08	2.62	273.10	121.38	50.03	买入
603859.SH	能科科技	37.20	0.92	1.05	1.32	40.43	35.43	28.18	买入
688615.SH	合合信息	89.73	3.24	4.22	5.25	27.69	21.26	17.09	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。