2026 年 05 月 13 日

华鑫证券 CHINA FORTUNE SECURITIES

OpenAI 发布 GPT-5.5 Instant，火山引擎推出 Doubao-Seed-2.0-lite

—计算机行业周报

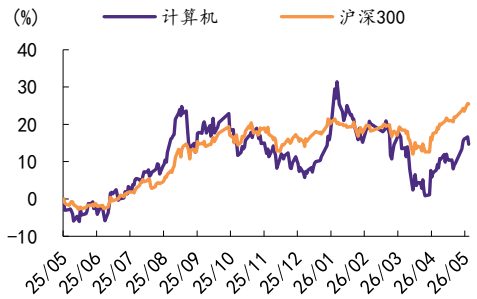
推荐(维持)

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)	6.4	-4.6	15.6
沪深 300	6.5	6.2	27.0

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：OpenAI 发布 GPT-5.5 旗舰大模型，DeepSeekV4 发布》2026-04-28

2、《计算机行业周报：ClaudeOpus4.7 深夜上线，Terafab 项目加速落地》2026-04-21

3、《计算机行业周报：ClaudeMythosPreview 正式发布，HermesAgent 引爆开源社区》2026-04-15

投资要点

算力：算力租赁价格平稳，火山引擎推出首款全模态理解模型 Doubao-Seed-2.0-lite

2026 年 5 月 6 日，字节跳动旗下火山引擎发文官宣豆包家族新版本大模型 Doubao-Seed-2.0-lite，该模型在全模态理解、Agent、Coding 与 GUI 能力等方面进行了同步升级。在同等算力成本下，为企业大规模、批量化部署全模态推理任务提供了更具性价比的选择，并在电竞游戏、在线教育、外电商运营等商业化场景下，表现出积极潜力。

AI 应用：文心一言周访问量环比+15.22%，OpenAI 发布 GPT-5.5 Instant

2026 年 5 月 6 日，OpenAI 推出了 GPT-5.5 Instant 模型。这款模型已经正式上线，并且面向所有 ChatGPT 用户免费开放，成为 ChatGPT 的默认模型。这次更新的核心亮点集中在三个方面：回答更加简洁、具备更强的记忆能力，以及更加个性化的交互体验。

AI 融资动向：RadixArk 完成 1 亿美元种子轮融资

2026 年 5 月，AI 基础设施初创公司 RadixArk 宣布成立，并且完成 1 亿美元种子轮融资，投后估值 4 亿美元。公司由开源推理引擎 SGLang 核心团队成员创立，首创 Day-0 兼容机制，让开源模型的发布即支持成为现实，致力于在开源内核之上，建立一套不锁定模型、不绑架客户、却提供顶级基础设施能力的托管平台。

投资建议

2026 年 5 月 6 日，英伟达与康宁达成多年期商业合作，聚焦 AI 基础设施高端光连接解决方案的美国本土产能扩建。为匹配 AI 数据中心激增的光连接需求，康宁计划将美国本土光连接产品制造产能提升至现有 10 倍，光纤产能扩充超 50%，同时在北卡罗来纳州、得克萨斯州新建三座先进制造工厂，新增 3000 余个高薪就业岗位，扩产产能将专供超大规模数据中心，全面支撑英伟达计算平台规模化部署。当前行业供需偏紧的态势逐步显现，下游厂商纷纷提前锁定稀缺光纤产能，市场供需错配带动光纤价格高增。其中通用普通光纤（G.652D）现货价自阶段低点 20 元/芯公里涨至 80 元-105 元/芯公里；适配高端精密传输场景的特种光纤（G.657.A2）价

格涨至 240 元-260 元/芯公里，体现出高端 AI 光连接资源的稀缺现状。本次合作以共封装光学（Cpo）为核心技术路径。伴随 AI 数据中心 GPU 集群规模化扩容、机间传输带宽与距离同步提升，黄仁勋也直言，下一代人工智能基础设施将需要大量的光学连接，因为计算需求正在迅速增长，以至于铜线已经无法满足需求。

产业链方面，英伟达在 2026 年 3 月斥资 40 亿美元投资激光器及光电转换元器件企业 Coherent 与 Lumentum，完成光电信号转换核心环节布局，搭配本次康宁光纤传输产能落地，持续补齐 AI 光连接的产业链。此次巨头持续加码光通信核心环节，映射 AI 产业底层发展逻辑迎来实质性迭代。过往行业算力竞争主要围绕 GPU 单硬件性能升级展开，当前算力扩容的核心瓶颈已逐步转移至芯片内部、服务器及机柜间的高速数据传输领域；高端场景下光互联替代传统铜缆，逐步变为规模化落地的刚需求。产业龙头的集中布局进一步夯实了高端 AI 光连接、高速光纤的长期增量逻辑，推动光通信行业传统周期属性显著弱化，迈入由 AI 算力建设驱动的高确定性成长周期，赛道长期配置价值值得关注。

中长期，建议关注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	2026-05-13	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	510.09	-0.30	0.30	0.60	-1700.30	1700.30	850.15	买入
301196.SZ	唯科科技	152.69	2.53	3.34	3.98	60.35	45.72	38.36	买入
603859.SH	能科科技	40.76	0.92	1.21	1.50	44.30	33.69	27.17	买入
688615.SH	合合信息	'4	3.24	4.22	5.25	44.06	33.82	27.19	买入

资料来源：Wind，华鑫证券研究

正文目录

图表目录

1、算力动态：算力租赁价格平稳，火山引擎推出首款全模态理解模型 Doubao-Seed-2.0-lite

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 5 月 4 日至 5 月 10 日，周度 Token 消耗量有所上升，调用量为 25.7T，环比上周增加 7.53%。在 Tokens 规模 Leaderboard 前五名中，Tencent 的 Hy3 preview(free)以 2.68T tokens 位居榜首，Moonshotai 的 Kimi K2.6 以 1.61T tokens 位居第二，Anthropic 的 Claude Sonnet 4.6 以 1.45T tokens 位居第三；Anthropic 的 Claude Opus 4.7 以 1.24T tokens 位列第四；DeepSeek 旗下的 DeepSeek V4 Flash 以 1.11T tokens 位居第五；

从市场份额维度来看，Anthropic 以 3.74T tokens 占据 14.5%的份额，稳居首位；Google 以 3.61T tokens 占据 14.0%，位列第二；Tencent、DeepSeek、OpenAI 则分别以 3.53T、2.99T、2.63T tokens，对应占据 13.7%、11.6%、10.2%的市场份额。

Google 宣布推出多 Token 预测（MTP）起草器，针对计算资源利用率不足的痛点，为 Gemma 4 系列大模型引入推测解码技术，配对重型目标模型（如 Gemma 4 31B）与轻量级起草器（MTP 模型）。数据显示，该架构下，起草器利用闲置算力预测未来 Token，并由目标模型并行验证，其 Token 推理速度最高可提升 3 倍，且不会对输出质量及推理逻辑造成影响，因而开发者将能够在资源受限的环境中部署先进的语言模型，且无需牺牲响应速度或计算精度，同等任务下算力成本有望降低。

5 月 8 日，中国国家数据局副局长余英在 2026 移动云大会上披露，截至 3 月底，我国 Token 调用量日均值已突破 140 万亿，和 2024 年底相比增长超千倍，呈现指数级增长。与此同时，余英指出，Token 调用量的增长正推动算力供给从“卖裸算力”向“卖服务、卖能力”转换，并从互联网、金融等先导行业向传统工业、交通等行业深度渗透，推动形成“人人可及、处处可用、按需服务”的算力发展新生态。

商汤科技正式推出新一代轻量化多模态智能体模型——日日新 SenseNova 6.7 Flash-Lite，该模型采用原生多模态架构，取消了视觉转文本的中间层后，避免了传统“转译”过程中的信息受损，及高昂的 Token 消耗，可大幅提升数据分析、深度调研、PPT 生成等长链路复杂任务的成功率。数据显示，在信息搜索等场景中，推理 Token 消耗对比纯文本智能体直降 60%，可实现毫秒级反馈，更符合高频互动的生产环境需求。

中国移动发布移动模型服务平台 MoMA，平台接入 MiniMax、豆包、GLM 等超 300 款业界主流 AI 模型，模型丰富度行业领先，并首创 Token 集约化运营模式，统一 API 调用，支持大小模型协同与多维策略调度。在成本控制方面，MoMA 基于国产算力部署自研推理引擎，结合智能路由对长尾模型资源调度，实现单位 Token 成本压降约 30%，降低资源占用率 50%以上。

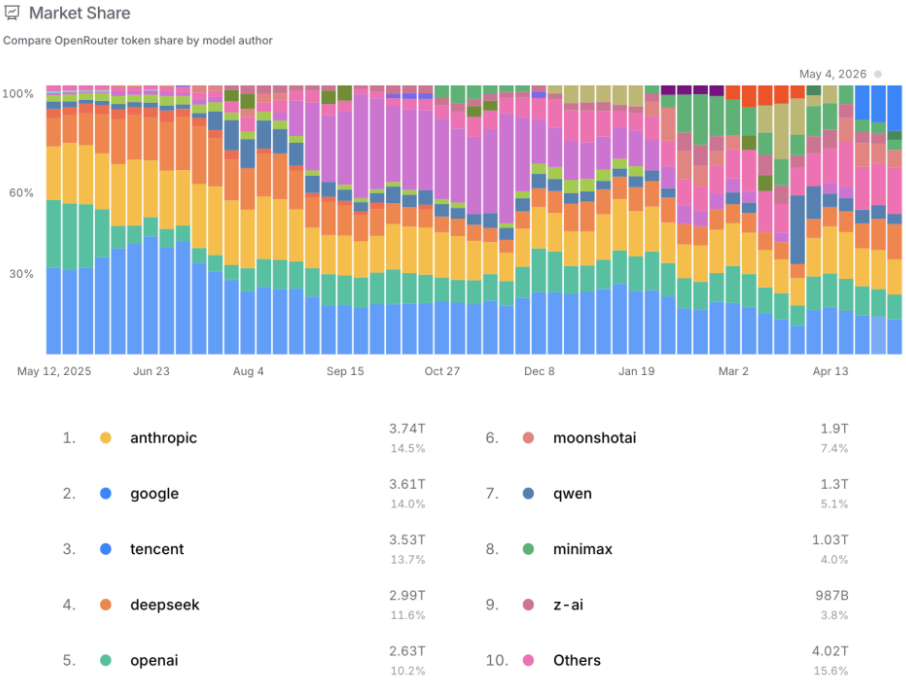
OpenAI 推出首个具备 GPT-5 级别推理能力的语音模型 GPT-Realtime-2，该模型较上一代 GPT-Realtime-1.5 在语音智能评估中的准确率及指令遵循评估中的平均通过率方面分别提升 15.2 和 13.8 个百分点。此外，该模型的上下文窗口从 32K 扩展至 128K，为复杂事务的处理及智能体工作流完整性的加强提供了技术支撑。根据 OpenAI 公布的定价标准，此次公开的

GPT-Realtime 系列 API 服务，GPT-Realtime-2 按音频 token 计费，每百万输入 token 价格为 32 美元，每百万输出 token 价格为 64 美元；GPT-Realtime-Translate（实时翻译服务）与 GPT-Realtime-Whisper（语音转写服务）则按使用时长计费，定价分别为每分钟 0.034 美元与每分钟 0.017 美元。

图表 1：Tokens 规模 Leaderboard

资料来源：OpenRouter，华鑫证券研究

图表 2：市场份额占据示意



资料来源：OpenRouter，华鑫证券研究

1.2、数据跟踪：腾讯云与阿里云持续涨价

近期，国内云服务市场涨价潮持续蔓延，腾讯云、阿里云相继跟进，行业近 20 年只降不升惯例已被打破，我国云计算市场的竞争逻辑正在从低价扩张转向能力定价。

2026 年 5 月 9 日，腾讯云和阿里云相关产品价格先后迎来持续性上涨。公告显示，继 3 月 11 日对部分模型启动正式收费及涨价后，腾讯云针对其 GPU 云服务器以及底层为 GPU 的容器服务（TKE）、大数据服务（EMR）相关产品的价格调整于 5 月 9 日正式生效，包含 AI 算力、容器服务 TKE 及弹性 MapReduce（EMR）在内的相关产品服务刊例价统一上调 5%。据相关人士披露，腾讯云此次调价主要受市场供需与成本压力以及全球行业整体趋势两方面影响，本次调价主要聚焦 AI 算力相关产品，其他云服务价格暂未调整。

与此同时，阿里云也将从 5 月 15 日起，调整部分 MU（Model Unit）模型单元的服务价格，涨幅从 2%至 7%不等，并将在 7 月份，对其 DDoS 原生防护 2.0（包年包月）、DDoS 高防（中国内地）以及 DDoS 高防（非中国内地）商品的弹性 95 功能进行价格调整。其中，DDoS 高防（中国内地）的弹性 95 由 100 元/Mbps/月调整为 150 元/Mbps/月，涨幅高达 50%。

1.3、产业动态：Doubao-Seed-2.0-lite 升级新版本，豆包官宣新增付费订阅

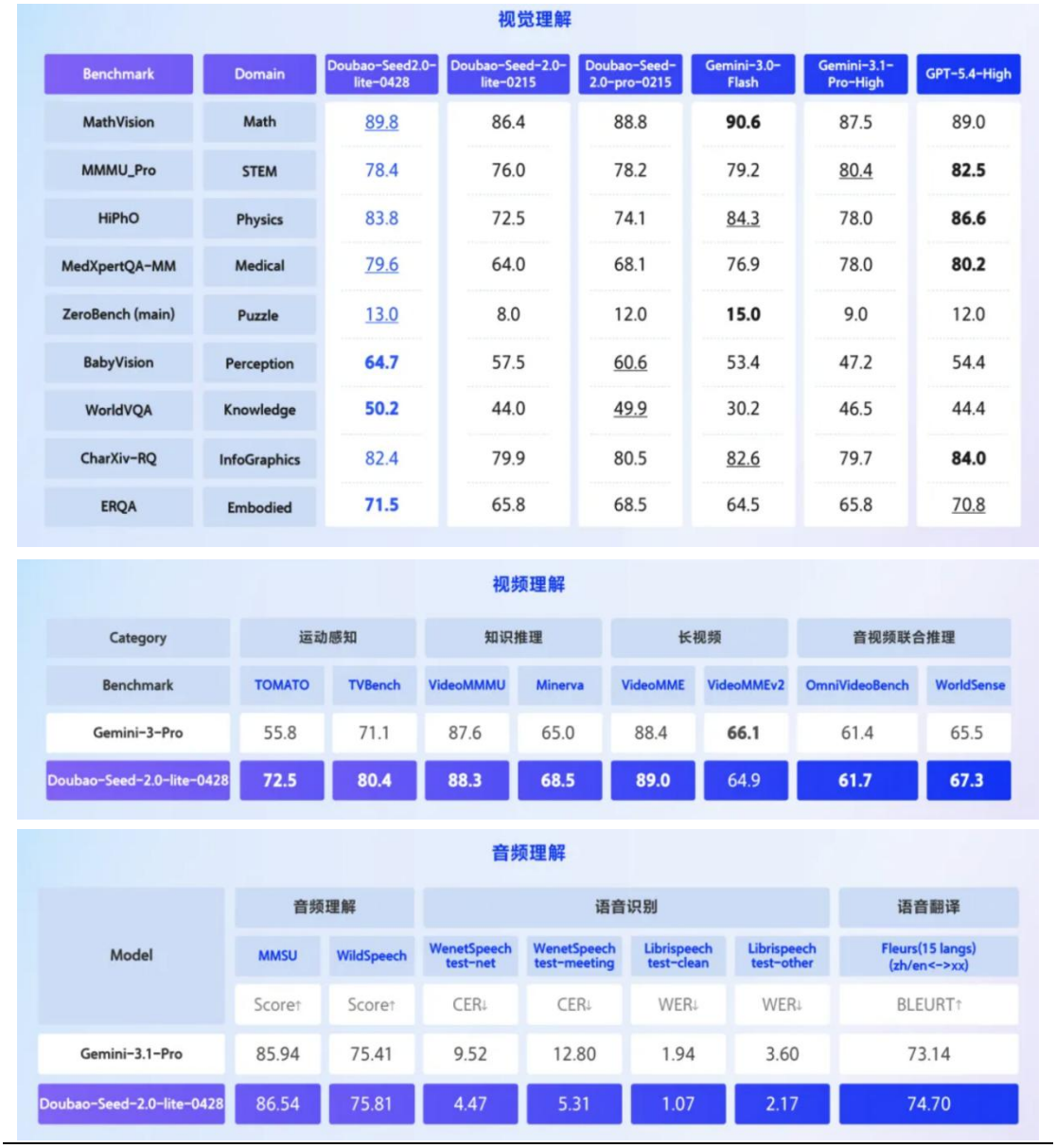
2026 年 5 月 6 日，字节跳动旗下火山引擎推出豆包大模型家族首款全模态理解模型 Doubao-Seed-2.0-lite，该新版本模型面向企业大规模、批量化部署，支持视频、图像、音频、文本原生统一理解，Agent、Coding、GUI 能力同步升级，同等算力成本下性价比更优。目前，该版本已在火山方舟正式上线。

相较 2 月发布的 Seed-2.0-pro，新版本在物理（HiPhO）、医疗（MedXpertQA）等高阶学科推理上的视觉理解能力大幅提升，并在细粒度感知（BabyVision、WorldVQA）与具身理解（ERQA）等关键领域达到 SOTA 水平，更加适应高价值场景规模化部署。

与此同时，新版本模型融入语音理解，可实现多种输入模态同时理解及完成跨模态联合推理。在视频理解场景下，模型能够联合分析视频中的画面与音频信息，精准辨析视频中的视听一致性，同时支持根据自然语言指令精准定位事件发生时间点，及跨越多个时间段提取关键线索，持续追踪人物与事件发展，通过多步逻辑推理还原事件关系与行为脉络等，对复杂业务需求的判断能力进一步提高。

音频理解能力部分，根据公开评测集，Seed-2.0-lite 在语音识别、翻译等多项音频理解基准上优于 Gemini-3.1-Pro。当前，新版本模型支持 19 个语种的精准语音转写、中英文与其他 14 个语种互译，能够深度捕捉情绪变化、环境背景与音乐细节，并输出更加贴近人类自然语言的语义信息。

图表 3：Seed-2.0-lite 视觉、视频、音频理解能力比较数据概览



视觉理解

Benchmark	Domain	Doubao-Seed2.0-lite-0428	Doubao-Seed-2.0-lite-0215	Doubao-Seed-2.0-pro-0215	Gemini-3.0-Flash	Gemini-3.1-Pro-High	GPT-5.4-High
MathVision	Math	89.8	86.4	88.8	**90.6**	87.5	89.0
MMMU_Pro	STEM	78.4	76.0	78.2	79.2	80.4	**82.5**
HiPhO	Physics	83.8	72.5	74.1	84.3	78.0	**86.6**
MedXpertQA-MM	Medical	79.6	64.0	68.1	76.9	78.0	**80.2**
ZeroBench (main)	Puzzle	13.0	8.0	12.0	**15.0**	9.0	12.0
BabyVision	Perception	**64.7**	57.5	60.6	53.4	47.2	54.4
WorldVQA	Knowledge	**50.2**	44.0	49.9	30.2	46.5	44.4
CharXiv-RQ	InfoGraphics	82.4	79.9	80.5	82.6	79.7	**84.0**
ERQA	Embodied	**71.5**	65.8	68.5	64.5	65.8	70.8

视频理解

Category	运动感知		知识推理		长视频		音视频联合推理	
Benchmark	TOMATO	TVBench	VideoMMMU	Minerva	VideoMME	VideoMMEv2	OmniVideoBench	WorldSense
Gemini-3-Pro	55.8	71.1	87.6	65.0	88.4	**66.1**	61.4	65.5
Doubao-Seed-2.0-lite-0428	**72.5**	**80.4**	**88.3**	**68.5**	**89.0**	64.9	**61.7**	**67.3**

音频理解

Model	音频理解		语音识别				语音翻译
	MMSU	WildSpeech	WenetSpeech test-net	WenetSpeech test-meeting	Librispeech test-clean	Librispeech test-other	Fleurs(15 langs) (zh/en<->xx)
	Score↑	Score↑	CER↓	CER↓	WER↓	WER↓	BLEURT↑
Gemini-3.1-Pro	85.94	75.41	9.52	12.80	1.94	3.60	73.14
Doubao-Seed-2.0-lite-0428	86.54	75.81	4.47	5.31	1.07	2.17	74.70

资料来源：火山引擎，华鑫证券研究

Seed-2.0-lite 在 Agent 和 Coding 方面也进行了能力升级。Agent 方面，新模型对多轮、多步、多约束的用户指令遵循度显示出明显提升，反思推理与多 Agent 协同调度能力继续增强，Agent 在长程任务中表现更加稳定。不仅如此，Seed-2.0-lite 还能与 OpenClaw、Hermes Agent 等框架达成深度适配，深度搜索与 Skill 动态调用得以强化。Coding 方面，新模型能力已全面覆盖前端页面、3D 场景与游戏开发，交付产物在视觉美观度与工程完整度都得到了进一步提升，能够胜任从原型到上线的前后端深度开发。

图表 4：LLM、Agent、Coding 方面大模型表现对比

Category	Benchmark	Doubao-Seed-2.0-lite-0428	Doubao-Seed-2.0-lite-0215	Doubao-Seed-2.0-pro-0215	GPT-5.4-Mini	Gemini-3-Flash
Knowledge	GPQA Diamond	88.4%	85.1%	88.9%	88.0%	**90.7%**
	SuperGPQA	69.6%	67.5%	68.7%	63.9%	**72.7%**
	HLE (no tool, text only)	25.7%	28.2%	**32.4%**	28.2%	31.7%
Reasoning	BeyondAIME	79.0%	76.0%	**86.5%**	80.0%	82.0%
	FrontierSci-olympiad	72.0%	70.0%	**74.0%**	70.0%	73.0%
	Superchem (text-only)	**55.0%**	48.0%	51.6%	29.1%	54.4%
	BABE	**57.9%**	50.2%	53.5%	49.0%	55.2%
Instruction Following	CL-Bench	20.1%	20.0%	**20.8%**	14.9%	16.1%
	MultiChallenge	**69.9%**	63.2%	68.3%	62.5%	69.3%
SearchAgent	WideSearch	70.3%	74.5%	**74.7%**	73.0%	64.0%
	BrowseComp	64.0%	72.1%	**77.3%**	61.3%	41.5%
	ResearchRubrics	**59.2%**	50.8%	50.7%	47.1%	36.9%
	XPert Bench	56.8%	63.3%	**64.5%**	41.8%	50.1%
Real World	SkillsBench	43.7%	42.1%	42.3%	**45.4%**	26.4%
	GDPval	53.1%	47.3%	**54.4%**	50.6%	13.7%
	FinSearchComp	63.8%	65.1%	**70.2%**	61.8%	43.7%
	Tob-Agent	51.4%	45.2%	**52.6%**	43.0%	37.4%
CodingAgent	SWE-Bench Pro	46.6%	46.0%	46.9%	**54.4%**	46.7%
	NL2Repo-Bench	28.7%	24.6%	27.9%	**37.3%**	27.6%
	SWE Multilingual	66.6%	64.4%	71.7%	**73.6%**	71.1%
	PaperBench	52.5%	**54.6%**	53.8%	49.1%	33.9%
	Terminal Bench 2.0	43.3%	45.0%	55.8%	60.0%	**60.0%**
	Vibe Coding 人工评估	49.4%	48.7%	48.4%	**57.4%**	56.9%

资料来源：火山引擎，华鑫证券研究

新版本在 GUI 能力上方面也进行了升级，打通了从界面理解到操作执行一体化闭环，既能精准识别按钮、菜单、表单、弹窗等界面元素及其状态，也能稳定完成点击、输入、右键、滚动、拖拽等 Browser Use 与 Computer Use 操作，让 Agent 具备了完整的交付能力。

图表 5：大模型 GUI 能力对比

Benchmark	Doubao-Seed-2.0-lite-0428	Doubao-Seed-1.8	Claude-Opus-4.7	Claude-Sonnet-4.5	Gemini-3.1-Pro
OSWorld-Verfied	64.4%	61.9%	**78.0%**	62.9%	64.0%
MobileWorld	**64.6%**	52.1%	56.4%	47.8%	57.3%

资料来源：火山引擎，华鑫证券研究

全新升级的 Seed-2.0-lite 模型将全模态理解与持续增强的 Agent、Coding 和 GUI 能力进行结合，在电竞游戏、在线教育、海外电商运营等领域的商业化场景下，都呈现出了积极的潜力：电竞场景下，新模型具备实现从多维度切片点评到跨回合走位追踪，到生成包含战况图谱及时间轴复盘在内的高质量复盘界面，到执行 25 小时长程任务并提供精准提升建议的能力；在线教育场景中，新模型能够针对课堂录像进行定时分析，自动生成包含课堂亮点与学生高光时刻在内的可视化报告，并进行分发，实现高效协作与教学反馈闭环；海外电商运营场景中，新模型可利用自身 GUI 能力完成海外电商多语言爆款视频的自动搜索、下载、要素拆解与 Skill 回写，生成多语言推广视频，并完成自动登录与平台发布。

关于付费方面，继国产大模型 Kimi 推出付费订阅服务后，近日豆包 AppStore 页面也出现付费版本服务声明，三档订阅价格分别为：标准版连续包月每月 68 元（连续包年 688 元）、加强版连续包月每月 200 元（连续包年 2048 元）、专业版连续包月每月 500 元（连续包年 5088 元）。本次付费更新主要聚焦于复杂任务和生产力场景（如 PPT 生成、数据分析、影视制作等），作为针对不同用户差异化需求推出的增值服务。目前相关方案细节还在测试阶段，相关付费选项和功能尚未开通。

2、AI 应用动态：文心一言周访问量环比+15.22%，OpenAI 发布 GPT-5.5 Instant

2.1、周流量跟踪：文心一言周访问量环比+15.22%

本期（2026.5.2-2026.5.8）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1289.0M）、Bing（794.2M）和 Gemini（653.4M），访问量环比增速第一为文心一言（15.22%）；平均停留时长前三位分别为 Character.AI（00:15:01）、Discord（00:11:02）和 Kimi（00:08:26）;平均停留时长环比增速第一为 ChatGPT（0.57%）。

图表 6：2026.5.2-2026.5.8AI 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	1289.0	-0.08%	5:53	0.57%
Bing	搜索	微软	794.2	-1.97%	7:30	-1.32%
Gemini	聊天机器人	谷歌	653.4	-0.88%	7:08	-0.93%
Canva	在线设计	Canva	230.2	-1.88%	5:50	-0.85%
Github	代码托管	微软	144.4	-0.82%	6:28	-1.27%
Discord	游戏社区	微软	141.8	2.46%	11:02	0.00%
Character.AI	聊天机器人	Character.AI	38.65	-7.34%	15:01	-6.92%
NotionAI	文本/笔记	Notion	38.34	-5.45%	7:54	-1.86%
Perplexity	AI 搜索	Perplexity	34.05	-5.76%	4:44	-1.39%
DeepI	翻译工具	DeepL	26.83	-3.35%	2:24	-2.70%
QuillBot	释义工具	QuillBot	10.88	-5.14%	2:48	-1.18%
Kimi	聊天机器人	Moonshot AI	9.88	-8.11%	8:26	0.00%
文心一言	聊天机器人	百度	0.64	15.22%	2:29	-1.97%

资料来源：similarweb,华鑫证券研究

2.2、产业动态：OpenAI 发布 GPT-5.5 Instant，智能与个性化全面升级

2026 年 5 月 6 日，OpenAI 推出了 GPT-5.5 Instant 模型。这款模型已经正式上线，并且面向所有 ChatGPT 用户免费开放，成为 ChatGPT 的默认模型。这次更新的核心亮点集中在三个方面：回答更加简洁、具备更强的记忆能力，以及更加个性化的交互体验。

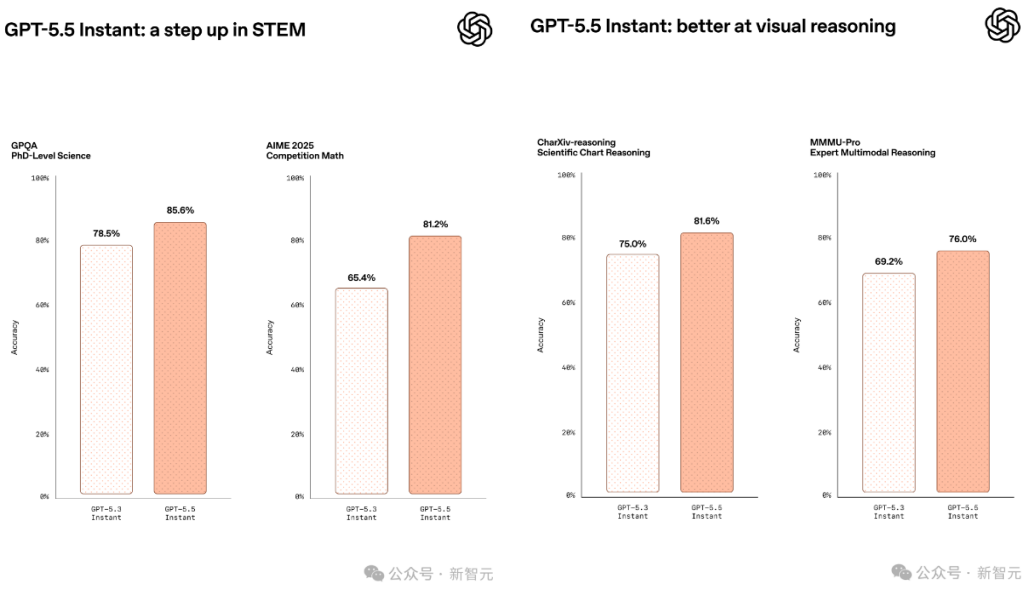
图表 7：GPT-5.5 Instant 取代 GPT-5.3 Instant，成为 ChatGPT「默认模型」

资料来源：新智元，华鑫证券研究

在具体的技术表现上，GPT-5.5 Instant 在处理日常任务时比上一代模型表现得更加出色。无论是分析用户上传的图片、解答 STEM 难题，还是精准判断何时需要自动调用网页搜索功能，它都表现得游刃有余。更为重要的是，它在减少幻觉方面取得了显著进步。GPT-5.5 Instant 相比前代产品，幻觉率降低了，尤其是在医疗、法律和金融这些对事实准确性要求极高的专业领域，这一提升显得尤为关键。同时，在用户主动标记为事实错误的高难度对话中，不准确陈述的数量也减少了 37.3%。通过实际的测试也能看到，当面对一个一开始包含计算错误的数学问题时，GPT-5.5 Instant 能够及时自我纠正，在最初附和了错误答案之后，自己发现将某个数值代回原方程并不成立，进而找出真正的计算错误，并用正确的公式解出答案。相比之下，上一代模型虽然也发现了错误，却没有进一步核算，直接得出了无实数解的错误结论。

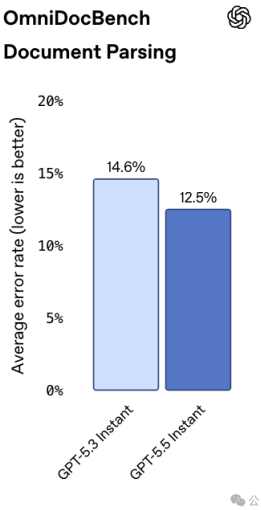
从各项基准测试的成绩来看，GPT-5.5 Instant 的表现相当亮眼。在数学 AIME 2025 测试中，其成绩从 65.4%跃升到了 81.2%。在博士级科学题 GPQA 测试中，得分从 78.5%提高到了 85.6%。在多模态推理 MMMU-Pro 测试中，也从 69.2%提升到了 76.0%。此外，在专门针对医学文档的基准测试 OmniDocBench 中，幻觉率也下降了 2.1%。

图表 8：GPT-5.5 Instant 在多个测试中的表现



资料来源：新智元，华鑫证券研究

图表 9：GPT-5.5 Instant 在 OmniDocBench 测试中的表现



资料来源：新智元，华鑫证券研究

除了准确性层面的提升，用户能够立刻感受到的另一个明显变化是，ChatGPT 的回答变得更简洁了。官方数据显示，相比前代产品，GPT-5.5 Instant 的回复字数减少了 30.2%，行数减少了 29.2%。例如，当被问到如何委婉地告诉同事不要总是唠叨时，以前的模型会列出多种策略、若干注意事项，还会贴心地询问同事的性格类型。而新模型直接砍掉了一半篇幅，先是一句提醒，然后按照用户可能希望采用的强硬程度分级提供话术，最后给出总结。整体的语气拿捏得更加精准，显得随性、实用且不过度。与此同时，OpenAI 还专门针对之前版本中无意义的表情符号泛滥、废话过多、格式过度等问题进行了纠正。GPT-5.5 Instant 学会了判断什么时候该给出详细的回答，什么时候用户仅仅只需要一句干脆利落的回复。

这次升级中最关键的部分在于个性化的提升。GPT-5.5 Instant 现在可以主动调用过往的聊天记录、用户上传过的文件，甚至是已经连接的电子邮箱，来为回复提供个性化的上下文。这意味着它开始真正记住用户了。同样是推荐一家新开的茶饮店，上一代模型给出的推荐看起来和推荐给任何普通用户的没有区别，只是泛泛地罗列了一些热门选择。而 GPT-5.5 Instant 则完全不同，它知道用户平时常去哪些店铺，了解用户偏爱什么样的茶饮风格，然后据此推荐了更符合个人口味的新店，甚至帮助用户做了决策，比如哪家适合成为新的日常打卡地，哪家适合寻求特别的体验。

与之同步上线的功能叫做记忆来源，面向所有个人计划用户开放。这个功能让用户能够看到 ChatGPT 具体引用了哪些过去的聊天记录或保存的记忆来生成当前的回复。如果某些信息已经过时或者不正确，用户可以直接删除或修改。当然，OpenAI 也承认，记忆来源功能可能无法穷尽所有影响最终回答的因素，它会展示几条最相关的历史聊天记录，但不一定是全部。

在实际操作的层面，GPT-5.5 Instant 即日起面向所有 ChatGPT 用户逐步推出。付费用户在接下来的三个月内仍然可以手动切换到 GPT-5.3 Instant，但三个月后，上一个模型将正式退役。基于聊天记录、文件和电子邮件的增强个性化功能，目前仅面向网页端的 Plus 和 Pro 用户开放，移动端即将上线。Free、Go、Business 和 Enterprise 用户将在未来几周内陆续获得权限。在 API 端，对应的模型 ID 也已经公布为 chat-latest

3、AI 融资动向：RadixArk 完成 1 亿美元种子轮融资，估值 4 亿美元

2026 年 5 月 5 日，AI 基础设施初创公司 RadixArk 宣布成立，并且完成 1 亿美元种子轮融资，投后估值 4 亿美元。该公司由来自 xAI 和 NVIDIA 的 AI 基础设施和模型专家 Ying Sheng 和 Banghua Zhu 共同创立，正基于两个开源项目 SGLang 和 Miles 开发商业化 AI 开发工具。

2023 年，Sheng 等人创建了开源推理引擎 SGLang，用于大规模服务模型部署。往后的两年间，SGLang 以惊人的迭代速度迅速成为开源大模型标杆，并在 GitHub 上积累了 27K+ 星值，在 400K+ GPU 上现实部署，每日支撑数万亿 token 的生产流量，用户包括 Google、Microsoft、NVIDIA、Oracle、AMD、LinkedIn、xAI、Thinking Machines Lab 等企业。

过去的两年里，该模型架构经历了 MoE、长上下文、Reasoning 模型、多模态融合等一系列剧变，团队首创性的 Day-0 兼容机制，让开源模型的发布即支持成为现实，形成了得天独厚竞争优势，也因此收到了投资人极高的评价。目前，SGLang 对几乎所有开源模型家族（Llama、Qwen、DeepSeek、Kimi、GLM、GPT、Gemma、Mistral 等）和硬件提供商（NVIDIA GPU、AMD GPU、Intel CPU、Google TPU 等）都提供 Day-0 支持，这些框架共同构成了托管基础设施和工具套件的起点，支持从个人开发者到初创公司、企业和研究实验室的所有 AI 系统构建者，实现云端托管 AI 模型。

RadixArk 认为，下一代 AI 不应受限于对私有基础设施的访问权，更多团队应当能够拥有自己的模型、系统和未来。经历种子轮融资后，RadixArk 下一步将致力于让 SGLang 成为任何新模型的 Day-0 生产标准；把 Miles 做成大规模训练与 RL 的基础设施级框架；在开源内核之上，建立一套不锁定模型、不绑架客户、却提供顶级基础设施能力的托管平台。

图表 10：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
阶跃星辰	端侧 AI 基础设施	产业链资本（华勤技术、龙旗科技、豪威集团、中兴通讯等）+ 香港投资管理有限公司	Pre-IPO 轮	近 25 亿美元	超 32 亿美元	100 亿美元
月之暗面	AI 原生应用平台	美团龙珠	D 轮	约 20 亿美元	超 47 亿美元	200 亿美元

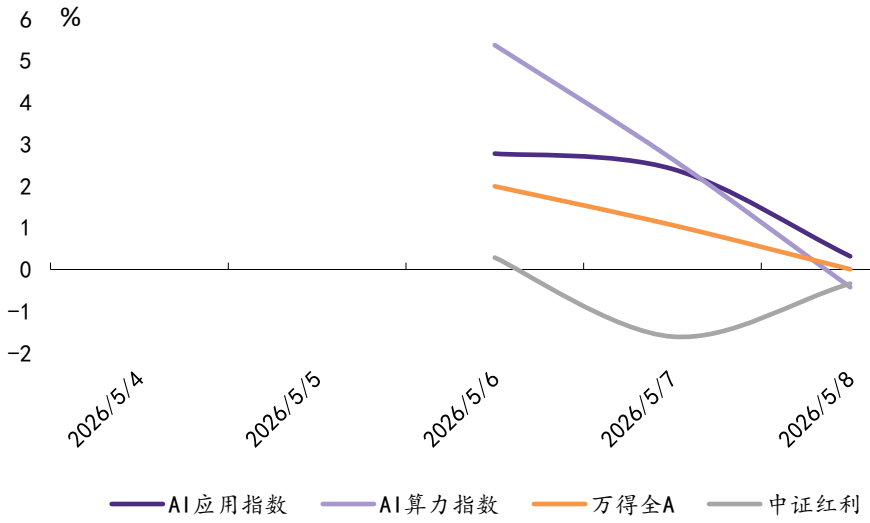
RadixArk	AI 基础设施	Accel、Spark Capital（联合领投）	种子轮	1 亿美元	1 亿美元	4 亿美元

资料来源：wind，Saasverse，华鑫证券研究

4、 行情复盘

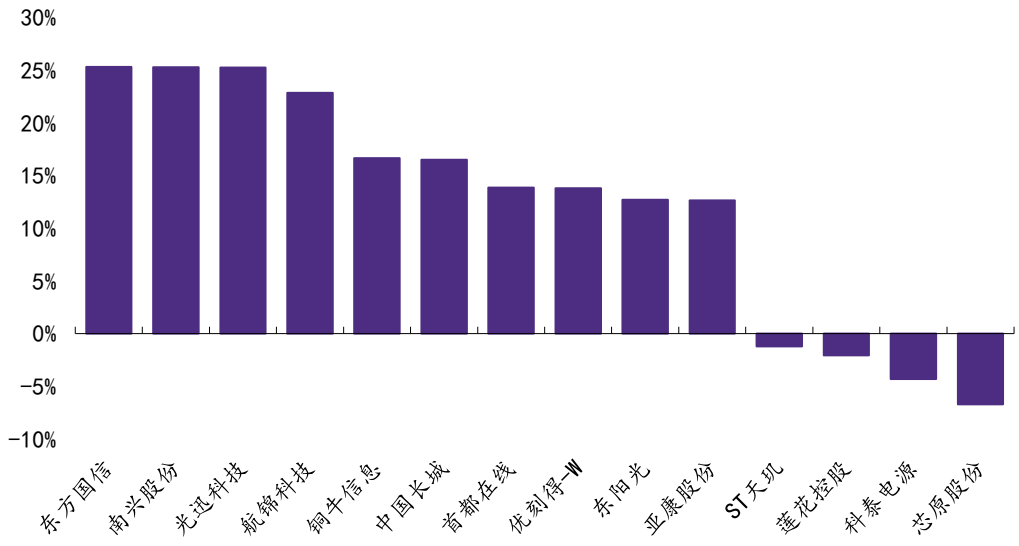
上周（2026.5.4-2026.5.8 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 2.78%/5.38%/1.99%/0.29%，AI 算力指数/中证红利日跌幅最大值分别为-0.42%/-1.61%。AI 算力指数内部，东方国信以 25.32%录得上周最大涨幅，芯原股份以-6.72%录得上周最大跌幅。AI 应用指数内部，易点天下以 28.17%录得上周最大涨幅，芯原股份以-6.72%录得上周最大跌幅。

图表 11：上周（2026.5.4-2026.5.8 日）指数日涨跌幅



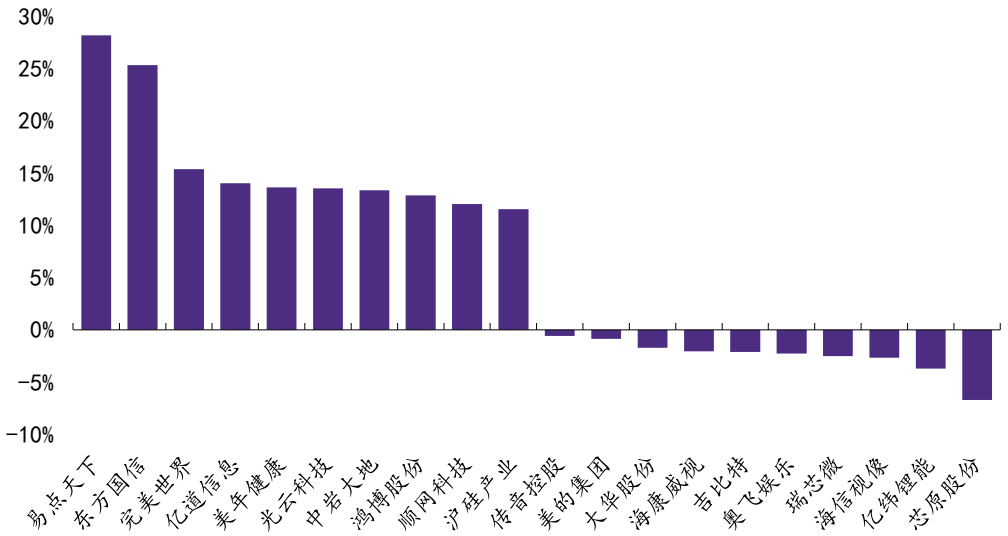
资料来源：wind，华鑫证券研究

图表 12：上周（2026.5.4-2026.5.8 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 13：上周（2026.5.4-2026.5.8 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 5 月 6 日，英伟达与康宁达成多年期商业合作，聚焦 AI 基础设施高端光连接解决方案的美国本土产能扩建。为匹配 AI 数据中心激增的光连接需求，康宁计划将美国本土光连接产品制造产能提升至现有 10 倍，光纤产能扩充超 50%，同时在北卡罗来纳州、得克萨斯州新建三座先进制造工厂，新增 3000 余个高薪就业岗位，扩产产能将专供超大规模数据中心，全面支撑英伟达计算平台规模化部署。当前行业供需偏紧的态势逐步显现，下游厂商纷纷提前锁定稀缺光纤产能，市场供需错配带动光纤价格高增。其中通用普通光纤（G.652D）现货价自阶段低点 20 元/芯公里涨至 80 元-105 元/芯公里；适配高端精密传输场景的特种光纤（G.657.A2）价格涨至 240 元-260 元/芯公里，体现出高端 AI 光连接资源的稀缺现状。本次合作以共封装光学（Cpo）为核心技术路径。伴随 AI 数据中心 GPU 集群规模化扩容、机间传输带宽与距离同步提升，黄仁勋也直言，下一代人工智能基础设施将需要大量的光学连接，因为计算需求正在迅速增长，以至于铜线已经无法满足需求。

产业链方面，英伟达在 2026 年 3 月斥资 40 亿美元投资激光器及光电转换元器件企业 Coherent 与 Lumentum，完成光电信号转换核心环节布局，搭配本次康宁光纤传输产能落地，持续补齐 AI 光连接的产业链。此次巨头持续加码光通信核心环节，映射 AI 产业底层发展逻辑迎来实质性迭代。过往行业算力竞争主要围绕 GPU 单硬件性能升级展开，当前算力扩容的核心瓶颈已逐步转移至芯片内部、服务器及机柜间的高速数据传输领域；高端场景下光互联替代传统铜缆，逐步变为规模化落地的刚需求。产业龙头的集中布局进一步夯实了高端 AI 光连接、高速光纤的长期增量逻辑，推动光通信行业传统周期属性显著弱化，迈入由 AI 算力建设驱动的高确定性成长周期，赛道长期配置价值值得关注。

中长期，建议关注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 14：ficonTEC2025 年年中至今公告订单

签约日期	客户/描述	业务类型	金额	折合人民币
2025/6/20	美国某头部公司 A 及其子公司	光电子封测设备	约 1,710 万欧元	约 1.36 亿元
2025/7/11	美国某头部公司 B 及其子公司	光电子封测设备	约 1,418 万美元	约 0.98 亿元
2025/9/3	瑞士某头部公司 C 的子公司	全自动硅光子封装整线设备或服务	约 946.50 万欧元	约 0.75 亿元
2025/10/21	武汉驿路通科技股份有限公司	光纤预制及组装线相关自动化设备	约 900 万美元	约 0.62 亿元
2026/1/6	瑞士某头部公司 C 的子公司	第二条全自动 OCS（光交换机）封装整线设备及服务	约 770.00 万欧元	约 0.61 亿元
2025/9/24-2026/1/26	以色列的纳斯达克上市的头部公司 E	单面晶圆测试设备及服务	约 921.60 万美元	约 0.64 亿元

2026/3/13	暂未披露	双面晶圆测试设备及服务	约 608.09 万欧元	约 0.48 亿元
2026/3/19-2026/3/25	纳斯达克上市的公司 F 及其子公司	耦合设备及服务（可用于可插拔硅光高速光模块封装制程核心环节的量产）	约 6 亿元人民币	约 6 亿元
2026/4/1	纳斯达克上市的公司 F	耦合设备及服务（可用于可插拔硅光高速光模块封装制程核心环节的量产）	约 3,570 万美元	约 2.46 亿元
2026/4/8-2026/5/1	纽约证券交易所上市的公司 B 的子公司	耦合设备及相关服务	约 2680 万美元	约 1.83 亿元
2026/4/8-2026/5/1	纳斯达克上市的公司 F	视觉检测设备、高精度激光 bar 条封装设备及相关服务	约 3226 万美元	约 2.20 亿
	总金额			约 17.93 亿元

资料来源：Wind，公司公告，华鑫证券研究

图表 15：重点关注公司及盈利预测

公司代码	名称	2026-05-13	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	510.09	-0.30	0.30	0.60	-1700.30	1700.30	850.15	买入
301196.SZ	唯科科技	152.69	2.53	3.34	3.98	60.35	45.72	38.36	买入
603859.SH	能科科技	40.76	0.92	1.21	1.50	44.30	33.69	27.17	买入
688615.SH	合合信息	142.74	3.24	4.22	5.25	44.06	33.82	27.19	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

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。