2026 年 06 月 10 日

华鑫证券 CHINA FORTUNE SECURITIES

Gemma 4 12B 开启本地多模态 AI 新时代，MiniMax M3 正式发布

—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)	-15.0	-16.3	1.3
沪深 300	-1.4	2.7	24.2

市场表现

资料来源：Wind，华鑫证券研究

相关研究

1、《计算机行业周报：ClaudeOpus4.8 发布，小米 MiMo 大模型 API 永久降价》2026-06-02

2、《计算机行业点评报告：Symbotic（SYM）：Q2 营收增长势头强劲，调整后 EBITDA 同比翻倍》2026-05-31

3、《计算机行业点评报告：小马智行（PONY）：Robotaxi 商业化加速兑现，全球版图扩张与成本下探》2026-05-29

投资要点

算力：算力租赁价格平稳，MiniMax 新一代旗舰大模型 MiniMax M3 正式发布

2026 年 6 月 1 日，MiniMax 发布新一代旗舰大模型 MiniMax M3。该模型在架构与能力上实现双重突破：采用 MSA 稀疏注意力机制，支持 100 万上下文窗口，预填充与解码速度较上代分别提升 9 倍和 15 倍以上；同时为原生多模态模型，支持图片、视频输入及电脑操作，在多模态测试集 OmniDocBench 中得分超越 Gemini 3.1 Pro，并在面向自主 Agent 的端到端评测框架 Claw-Eval 中取得最高分。

AI 应用：Character.AI 周访问量环比+11.57%，Gemma 4 12B 开启本地多模态 AI 新时代

2026 年 6 月 4 日，谷歌发布了轻量级多模态模型 Gemma 4 12B，一款能够在 16GB 内存的轻薄笔记本上全离线流畅运行的高性能模型。该模型抛弃了传统的编码器结构，能够直接处理文本、图像和音频输入。与此同时，DeepMind 的首席执行官 Demis Hassabis 也公开表示，Gemma 4 全系列模型的下载量已突破 1.5 亿次。

AI 融资动向：Ramp 完成 7.5 亿美元后估值达 440 亿美元

2026 年 6 月 4 日，AI 金融公司 Ramp 完成 7.5 亿美元融资，投后估值攀升至 440 亿美元，累计融资规模达 30 亿美元。本轮融资由 Iconiq Capital、新加坡主权财富基金 GIC 以及安大略省教师退休金计划联合领投，高盛成长股权、摩根士丹利投资管理及业客户，以 AI 驱动的企业信用卡及财务自动化平台为核心，推出自主 AI 智能体用于欺诈检测和交易审核，并于 2026 年 3 月上线新功能，支持 AI 智能体完成企业支付等操作。

投资建议

2026 年 6 月 2 日，英伟达宣布其 Spectrum-X 以太网硅光技术已全面量产。新一代 Spectrum-X 交换机基于光电一体封装技术（CPO）构建，支持其 VeraRubin 平台在数据中心实现横向扩展与跨区域扩展，为 AI 工厂部署提供网络支撑。公司通过与台积电、SPIL、TX 以太网硅光技术的量产，四家企业分别在硅光芯片制造、芯

片级封装测试、激光芯片与光模组、系统组装环节提供核心技术支持。作为英伟达全栈协同设计的典范，该技术相较传统收发器网络实现能效与 AI 集群正常运行时间均提升 5 倍，部署效率提升 30%，为百万 GPU 级 AI 工厂奠定了坚实的网络基础，目前已获得 CoreWeave、Lambda 及 Oracle Cloud Infrastructure 的率先采用。其大规模 CPO 部署突破了光互连在功耗、可靠性与部署时间方面的瓶颈，消除了制约 AI 集群规模扩张的关键障碍。光通信是英伟达战略布局的核心方向之一，在本周举办的 Computex2026 上，光互连领域龙头企业迈威尔科技首席执行官与黄仁勋同台出席。黄仁勋表示迈威尔科技有望成为下一家市值突破万亿美元的科技企业，并透露双方正进一步深化战略合作关系，共同打造支撑下一代人工智能数据中心运行的关键网络与连接基础设施体系。

2026 年截至 6 月，英伟达已密集对四家美国行业龙头企业进行大规模投资：3 月分别向迈威尔科技、Lumentum 及 Coherent 各注资 20 亿美元，其中与迈威尔科技的合作旨在将其定制 AI 芯片和网络技术整合进英伟达 NVLink 系统，对 Lumentum 和 Coherent 的投资则全面押注光互联技术与封装集成；5 月宣布与康宁达成多年期商业与技术合作伙伴关系，总投资上限 32 亿美元，支持其将美国光连接制造能力提升 10 倍、光纤产量提升超 50%，并在北卡罗来纳州和得克萨斯州新建三座先进制造工厂。英伟达作为全球人工智能产业绝对龙头，其全栈式技术协同与产业链资本投入，构成了行业长期增长最坚实的确定性基础。Spectrum-X 硅光互联平台全面量产及共封装光学(CPO)架构大规模商用，标志着光通信行业进入技术迭代与需求爆发的共振期。AI 算力建设需求的加速将驱动光通信板块景气度持续上行。

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）、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	2026-06-10 股价	EPS			PE			投资评级
			2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	627.01	-0.30	0.30	0.60	-2090.03	2090.03	1045.02	买入

301196.SZ	唯科科技	158.40	2.53	3.34	3.98	62.61	47.43	39.80	买入
603859.SH	能科科技	52.95	0.92	1.21	1.50	57.55	43.76	35.30	买入
688615.SH	合合信息	125.25	3.24	4.22	5.25	38.66	29.68	23.86	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，MiniMax 新一代旗舰大模型 MiniMax M3 正式发布

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 5 月 31 日至 6 月 7 日，周度 Token 消耗量有所上升，调用量为 36.1T，环比上周增加 13.52%。在 Tokens 规模 Leaderboard 前五名中，DeepSeek 的 DeepSeek V4
位居第二，Minimax 的 Minimax M3 以 2.5T tokens 位居第三；Xiaomi 的 MiMo-V2.5 以 2.19T tokens 位列第四；OpenRouter 旗下的 Owl Alpha 以 1.95T tokens 位居第五；

从市场份额维度来看，DeepSeek 以 6.75T tokens 占据 18.7%的份额，稳居首位；Anthropic 以 5.29T tokens 占据 14.6%，位列第二；Google、MiniMax、Xiaomi 则分别以 4.29T、3.05T、2.95T tokens，对应占据 11.9%、8.4%、8.2%的市场份额。

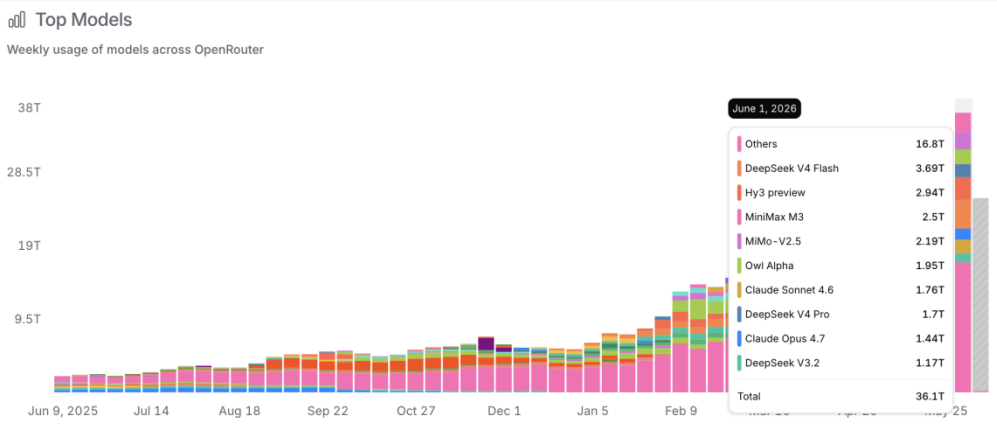
6 月 1 日，微软旗下人工智能编程工具 GitHub Copilot 正式实施计费模式调整，从固定额度订阅制转向按 Token 用量计费。新模式下，Copilot 按照用户实际 Token 消耗（包括输入、输出及缓存）核算成本，具体费率则按照不同模型的 API 定价执行。

6 月 5 日，腾讯云总经理、TokenHub 负责人高航于 AI 产业应用大会上透露，公司大模型服务平台 TokenHub 上线三个月以来，每月连续增长呈翻倍态势，现日 Token 消耗量已突破 5 万亿。

6 月 5 日，华为云发布四大基础设施新品之一——AICS 灵衢智算集群。该集群基于华为独立自主的所有算力芯片和全栈全国产化的算力硬件系统，支持 10 万卡级规模，总算力达 200E
务可用性高达 99.95%。

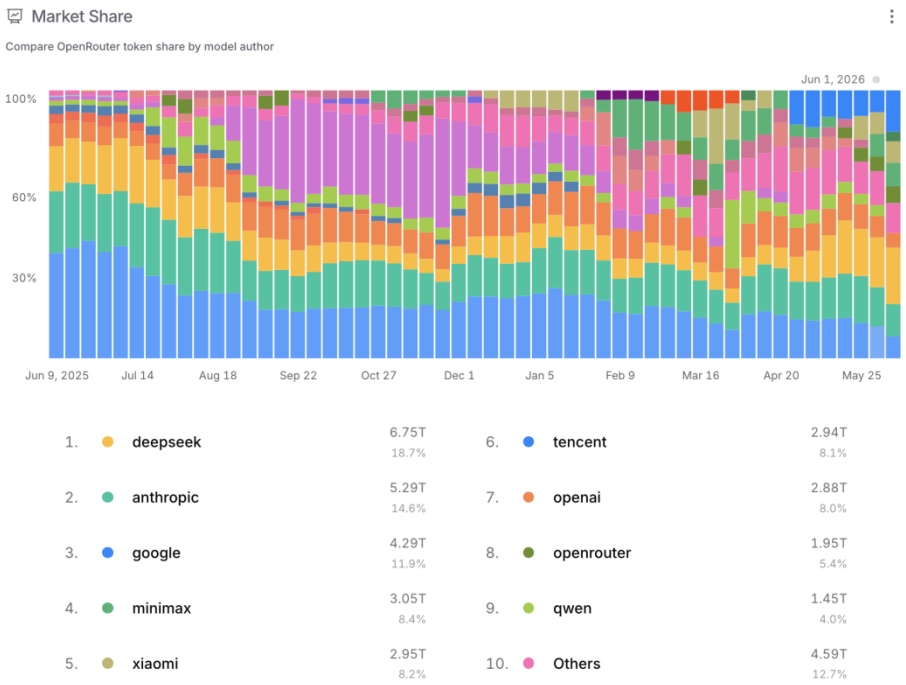
近期，中国信通院联合多家机构将于 6 月 16 日在北京举办高质量 Token 服务研讨会。该会议旨在系统性地提升 Token 服务能力，从组织机制和能力建设两方面入手，汇聚行业力量，为 Token 服务的规范化与高质量发展提供支撑。

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：腾讯云下调 DeepSeek-V4 系列模型价格，阿里云上线 Qwen3.7-Plus

2026 年 6 月 3 日，腾讯云智能体开发平台针对 DeepSeek-V4 系列模型价格进行下调，其中，最高降幅达 97.5%。本次调整仅涉及价格变更，不涉及模型服务能力变动。

具体来看，DeepSeek-V4-Pro 模型的推理输入、输出价格均下调 75%，分别降至 0.003 元/千 tokens 和 0.006 元/千 tokens；缓存命中价格下调 97.5%，降至 0.000025 元/千 tokens。DeepSeek-V4-
0.00002 元/千 tokens。价格调整后，腾讯云 DeepSeek-V4 系列模型调用价格已同 DeepSeek 官方售价全面持平。

此前，腾讯云已针对部分模型开展过一系列价格调整。其中，GLM 5、MiniMax 2.5、Kimi 2.5 由免费转为正式商用，并根据模型调用按量计费。Tencent HY2.0 Instruct、Tencent HY2.0 Think 的输入、输出价格则进行了上调。

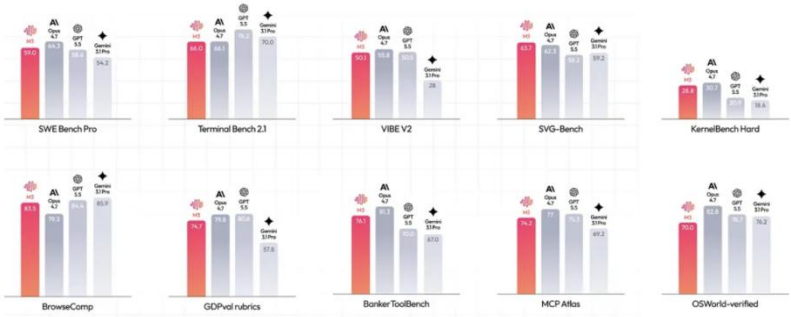
除此以外，上周（5 月 27 日）小米正式宣布旗下 MiMo-V2.5 系列 API 价格永久下调。其中，MiMo-V2.5-Pro 输入缓存命中降幅高达 99%，定价与 DeepSeek V4-Pro 完全对齐。

与此同时，阿里通义千问于 6 月 2 日发布千问 3.7 系列多模态大模型 Qwen3.7-Plus。该模型补齐了 Qwen3.7 系列的视觉识别能力，实现了多模态混合智能体的新突破，不仅能看懂图片和视频，还能深度推理、自我编程、调用工具、验证测试并自主迭代。当前，该模型已在阿里云百炼平台上线，支持 OpenAI 兼容 API 与 Anthropic 协议，并于 Qwen Studio 开放在线体验。

1.3、产业动态：MiniMax 新一代旗舰大模型 MiniMax M3 正式发布

2026 年 6 月 1 日，MiniMax 正式发布其新一代旗舰大模型 MiniMax M3。实验结果显示，该模型已在多个衡量编程与 Agent 能力的基准测试中达到前沿水平。在真实软件工程能力基准测试 SWE-Bench Pro 中，该模型取得 59.0%的得分，其表现已小幅度超过 GPT-5.5 和 Gemini 3.1 Pro，接近 Claude Opus 4.7。

图表 3：MiniMax M3 基准测试结果横向对比图



资料来源：智东西，华鑫证券研究

在架构创新与模型能力上，MiniMax M3 实现了双重突破。一方面，该模型采用了全新的稀疏注意力架构 MSA（MiniMax Sparse Attention），将上下文窗口扩展至 100 万 tokens 大小，计算速度大幅提升，并且与上一代全注意力模型 M2 相比，M3 在预填充阶段实现了 9 倍以上的加速，在解码阶段实现了 15 倍以上的加速。另一方面，该模型作为一款原生多模态模型，支持图片、视频输入，并具备操控电脑桌面的能力，在多模态测试集 OmniDocBench 中，该模型得分超越 Gemini 3.1 Pro，并在面向自主 Agent 的端到端评测框架 Claw-Eval 中取得最高分。

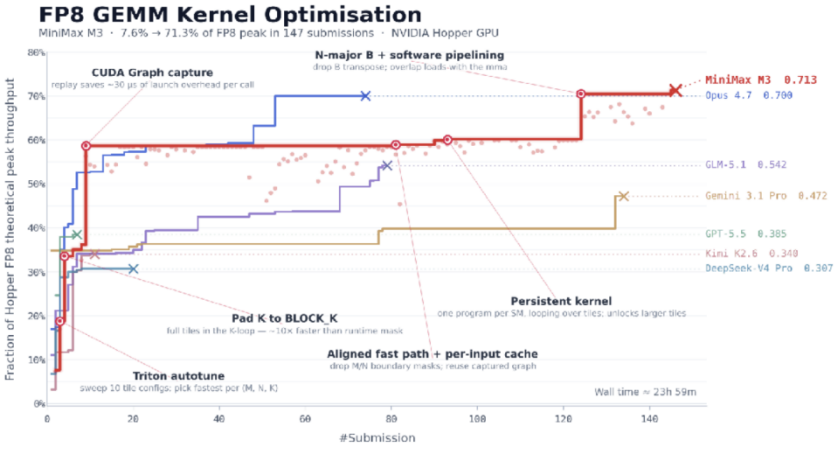
这种高效长上下文与原生多模态的结合，使该模型具备了驾驭超复杂任务的能力。在 PostTrainBench 后训练能力测试中，该模型根据研究团队下达的指令，于 12 小时内从零开始自主训练 4 个仅有预训练基座的模型。在全程没有人工干预的情况下，M3 独立完成了数据合成、模型训练、效果评估与迭代优化的完整闭环，最终让这 4 个模型在数学推理、工具调用、代码生成等五项任务上习得了基础能力，综合得分为 0.37，紧追 GPT-5.5 的 0.39 和 Opus 4.7 的 0.42，且其表现显著领先于其他参测模型。

图表 4：MiniMax M3 于 PostTrainBench 后训练能力测试中的表现情况

资料来源：智东西，华鑫证券研究

此外，M3 还展现出在底层硬件优化上的强大自主能力，连续工作 24 小时，历经 147 次 benchmark 提交和 1959 次工具调用，该模型完成对 CUDA 内核的优化，成功将 Hopper 件峰值利用率从首版的 7.6%提升至 71.3%，实现了相较于原始版本近 9.4 倍的加速。

图表 5：MiniMax M3 自动优化 CUDA 内核成果图



资料来源：智东西，华鑫证券研究

实测体验结果显示，该模型具备主动迭代需求、规划项目、持续反思纠错等能力，并能够针对视觉任务中的细节进行详尽描述，同时展现出不错的 Agentic 能力和扎实的多模态分析底子，但在具体任务交付上，不少结果的完成度仍然不够高。

价格方面，该模型的 API 调用价格以 512k 上下文为界分为两档，并提供优先调用和标准模式两种方案。其中，512k 以内上下文的调用享有 7 天限时五折优惠。优惠后，该模型标准模式下的输入价格为 2.1 元/百万 tokens，输出价格为 8.4 元/百万 tokens，缓存读取价格为 0.42 元/百万 tokens；优先模式下的输入价格为 3.15 元/百万 tokens，输出价格为 12.6 元/百万 tokens，缓存读取价格为 0.63 元/百万 tokens。

当前，MiniMax M3 已在 MiniMax Code、Token Plan 和 API 中上线，并将在未来 10 天内更新模型技术报告以及开源对应的模型权重。

图表 6：MiniMax M3 模型 API 调用价格表

标准 优先*

模型	输入价格 元/百万 tokens	输出价格 元/百万 tokens	缓存读取 元/百万 tokens
MiniMax-M3 ≤ 512k 输入 tokens 7天限时五折	~~4.20~~ 2.10	~~16.80~~ 8.40	~~0.84~~ 0.42
MiniMax-M3 > 512k 输入 tokens*	8.40	33.60	1.68

标准 优先*

模型	输入价格 元/百万 tokens	输出价格 元/百万 tokens	缓存读取 元/百万 tokens
MiniMax-M3 ≤ 512k 输入 tokens 7天限时五折	~~6.30~~ 3.15	~~25.20~~ 12.60	~~1.26~~ 0.63
MiniMax-M3 > 512k 输入 tokens	12.60	50.40	2.52

资料来源：智东西，华鑫证券研究

2、 AI 应用动态：Character.AI 周访问量环比+11.57%，Gemma 4 12B 开启本地多模态 AI 新时代

2.1、 周流量跟踪：Character.AI 周访问量环比+11.57%

本期（2026.5.29-2026.6.4）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1260.0M）、Bing（818.9M）和 Gemini（681.4M），访问量环比增速第一为 Character.AI（11.57%）；平均停留时长前三位分别为 Character.AI（00:14:12）、Discord（00:11:06）和 Kimi（00:08:28）;平均停留时长环比增速第一为文心一言（10.06%）。

图表 7：2026.5.29-2026.6.4 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	1260.0	3.03%	6:02	1.12%
Bing	搜索	微软	818.9	-0.70%	7:21	-0.68%
Gemini	聊天机器人	谷歌	681.4	3.26%	7:00	0.00%
Canva	在线设计	Canva	213.4	-0.88%	5:49	0.00%
Github	代码托管	微软	148.5	3.70%	6:26	-0.26%
Discord	游戏社区	微软	146.7	0.69%	11:06	0.15%
Character.AI	聊天机器人	Character.AI	39.33	11.57%	14:12	-0.35%
Perplexity	AI 搜索	Perplexity	28.90	-1.80%	4:26	-1.48%
DeepI	翻译工具	DeepL	25.87	-0.04%	2:29	2.05%
Kimi	聊天机器人	Moonshot AI	9.49	-6.77%	8:28	-0.20%
QuillBot	释义工具	QuillBot	9.08	-3.26%	2:54	1.16%
NotionAI	文本/笔记	Notion	8.646	-73.66%	7:50	-1.26%
文心一言	聊天机器人	百度	0.60	5.16%	2:55	10.06%

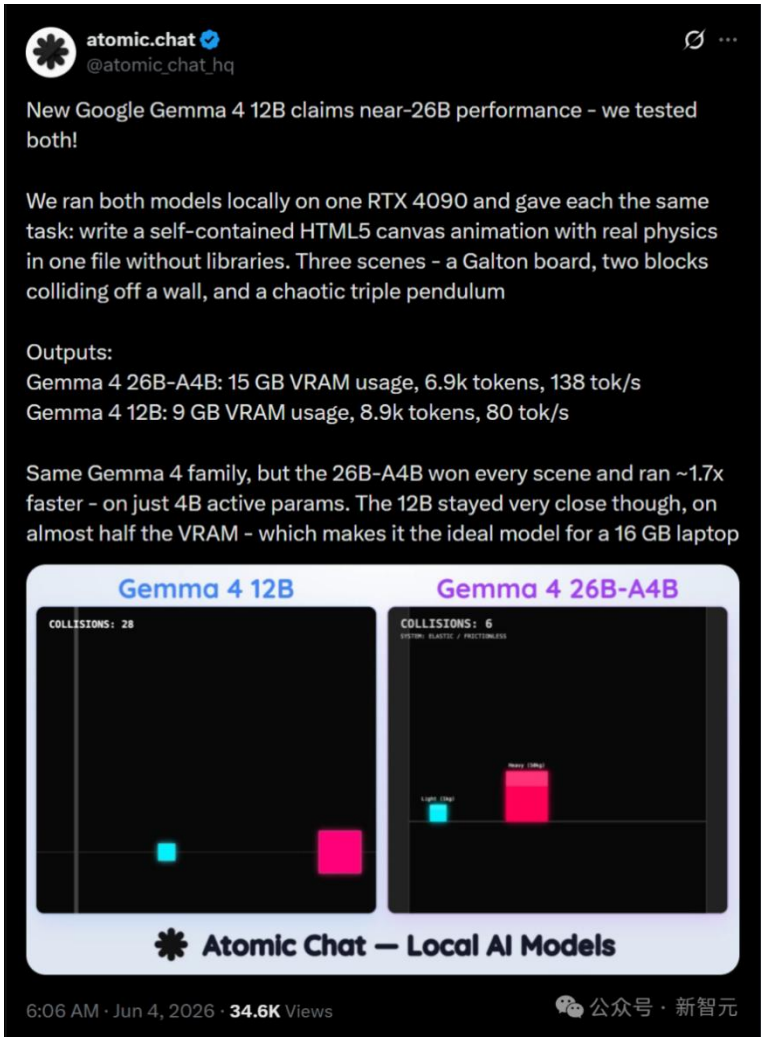
资料来源：similarweb,华鑫证券研究

2.2、 产业动态：下载量突破 1.5 亿次，Gemma 4 12B 开启本地多模态 AI 新时代

2026 年 6 月 4 日，谷歌发布了轻量级多模态模型 Gemma 4 12B，一款能够在 16GB 内存的轻薄笔记本上全离线流畅运行的高性能模型。该模型抛弃了传统的编码器结构，能够直接处理文本、图像和音频输入。与此同时，DeepMind 的首席执行官 Demis Hassabis 也公开表示，Gemma 4 全系列模型的下载量已突破 1.5 亿次。

在实际性能测试中，评测机构 atomic.chat 将 Gemma 4 12B 与前代模型 Gemma 4 26B-A4B 在单张 RTX 4090 显卡上进行了对比。测试任务要求模型在不依赖任何第三方库的情况下，仅凭自身推理能力生成包含高尔顿钉板、方块碰撞以及三摆系统等复杂物理动效的 HTML5 Canvas 代码。结果显示，虽然 Gemma 4 26B-A4B 模型以更高的 138token/s 的推理速度表现出了明显优势，但 Gemma 4 12B 以 80token/s 的速度同样成功完成了全部物理场景的代码生成，而其所占用的显存仅为 9GB，Gemma 4 26B-A4B 模型则占用了 15GB 的显存。这意味着，以往需要依赖云端 API 或昂贵双路工作站的复杂多模态物理推理任务，如今可以在普通的 MacBook 或者搭载消费级显卡的游戏本上离线完成。

图表 8：Gemma 4 12B 与 Gemma 4 26B-A4B 在单张 RTX 4090 显卡上的对比



资料来源：新智元，华鑫证券研究

Gemma 4 12B 之所以能够在较小的参数规模下实现如此强大的多模态理解能力，关键在于谷歌 DeepMind 采用了无编码器的统一架构设计。传统多模态模型通常需要视觉编码器将图像转换为向量、音频编码器将声音波形转换为向量，再将这些向量送入大语言模型进行处理。这种先编码再融合的方式存在延迟高、占用内存多、训练难度大等问题。而 Gemma 4 12B 直接处理原始的图像像素块和音频信号，视觉信息仅通过一个 35M 的超轻量级嵌入模块就能像文本 Token 一样流入模型的核心网络中，音频信号则被直接投影到与文本 Token 相同的维度空间。这种设计不仅大幅降低了端到端的延迟，还使得视觉、音频和文本共享同一套模型权重，开发者在使用 Hugging
就能同时更新所有模态的参数，极大方便了本地开发者的使用。

图表 9：Gemma 4 12B 的无编码器统一架构

资料来源：新智元，华鑫证券研究

图表 10：Gemma 4 12B 与传统多模态模型的对比

资料来源：新智元，华鑫证券研究

在应用场景层面，Gemma 4 12B 展现出了强大的 Agentic 能力。官方演示中的一个典型案例是，模型能够根据开发者的自然语言指令，自动生成完整的 Python 和 Gradio 桌面应用代码，构建出了一个带有图形界面的图像处理工具。这套应用背后的图像分析核心引擎调用了本地的 Gemma 4 12B 自身，形成了模型自己编写应用来调用自己的局面。另一个案例中，模型被输入一段长达五分钟的演讲视频，包含上千帧画面和原始音频，提示词要求模型分析视频中某个特定行为对应的深层含义。模型不仅完整消化了长达 256K token 的多模态上下文，还准确识别出了其中隐藏的视觉隐喻，这种深度的视频理解能力过去通常只有在顶级闭源模型中才能见到。

Gemma 4 12B 采用 Apache 2.0 开源协议，为商业化落地扫清了障碍，开发者可以自由修改、微调并将模型嵌入商业产品中而无需向谷歌支付任何费用。该模型已适配 LM Studio、Ollama、llama.cpp、MLX、vLLM 等多种主流本地部署工具，用户只需几条命令即可在自己的电脑上运行。

3、AI 融资动向：Ramp 完成 7.5 亿美元轮融资，投后估值达 440 亿美元

2026 年 6 月 4 日，AI 金融公司 Ramp 完成 7.5 亿美元
年 11 月的 320 亿美元攀升至 440 亿美元，累计融资规模已达 30 亿美元。本轮融资由 Iconiq Capital、新加坡主权财富基金 GIC 以及安大略省教师退休金计划联合领投，高盛成长股权、摩根士丹利投资管理及 Peter Thiel 旗下的

过去一年，Ramp 的估值增长了近两倍。当前，公司年化收入突破 15 亿美元，相较 2025 年 9 月的 10 亿美元，年化收入增长了 50%。收入的快速增长成为其估值持续重估的核心支撑。

Ramp 成立于 2019 年，最初专注于帮助初创企业管理费用报销，如今业务已拓展至企业支付、财务自动化及 AI 驱动的欺诈检测等领域。公司目前服务约 7 万家企业客户，较年初的 5 万家增长 40%，其中，新增客户相当一部分来自快速扩张的 AI 初创企业，公司则针对企业不断增长的模型调用、算力采购及相关支付支出的管理需求提供支持。

公司以面向企业的公司信用卡服务为核心，财务团队可通过设定支出规则来管控非必要开销。在此基础上，公司还提供高利率的企业银行账户服务 Treasury，以及供应商发票处理工具。2025 年 7 月，公司推出首款用于欺诈检测与交易审核的自主 AI 智能体，平台功能覆盖账单支付、采购、差旅、资金管理和会计自动化等综合场景。

公司自身也全面拥抱 AI 技术，不仅设立了 AI 研究实验室，还自主开发内部 AI 工具，并推动销售、客服、工程等团队广泛使用 AI。2026 年 3 月，公司上线新功能，允许 AI 智能体在平台上执行几乎所有原需人类用户完成的操作，包括发起企业信用卡支付。首席产品官 Geoff Charles 表示，未来企业间交易模式将演变为“智能体对智能体支付”，也就是说 AI 将替代人类自主执行从采购、审批到收付款等环节，人类则仅承担监督角色。

目前，Ramp 在全球拥有约 1700 名员工，并计划进一步扩大团队规模，持续加码 AI 驱动的企业服务市场。

图表 11：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
Ramp	AI 金融	Iconiq Capital、GIC、安大略省教师退休金计划				440 亿美元

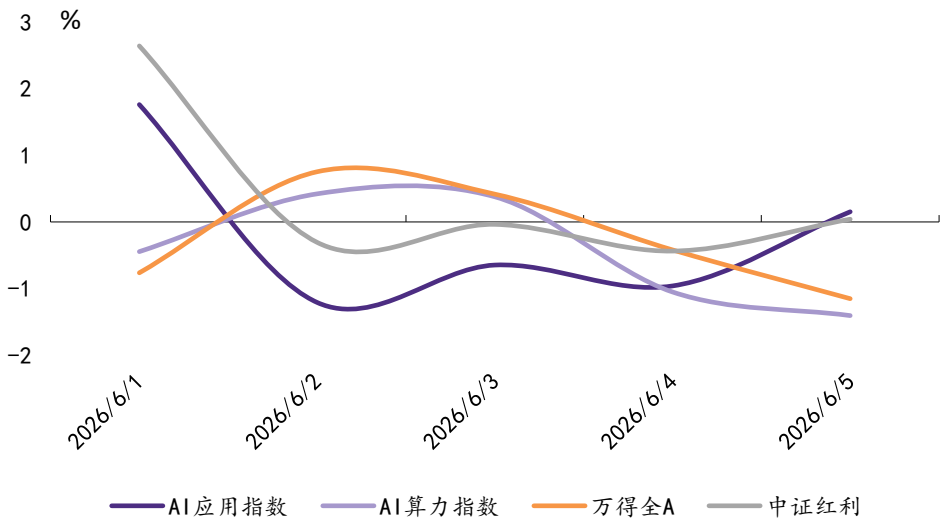
Supabase	AI 基础设施	GIC			美元	105 亿美元
Suno	AI 音乐生成	Bond Capital	D 轮	4 亿美元	超 8 亿美元	54 亿美元

资料来源：wind，Saasverse，华鑫证券研究

4、 行情复盘

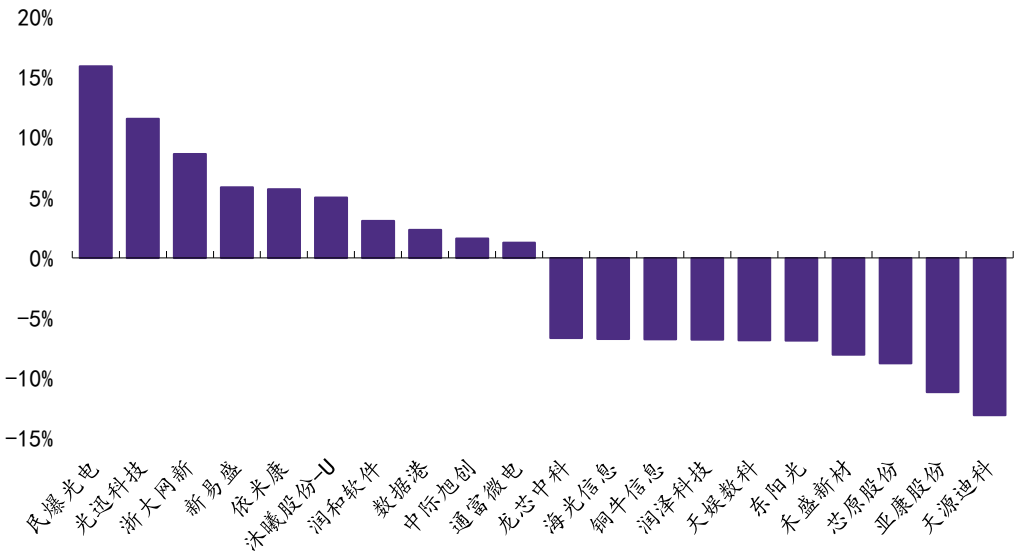
上周（2026.6.1-2026.6.5 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 1.76%/0.42%/0.75%/2.64%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-1.21%/-1.41%/-1.15%/-0.44%。AI 算力指数内部，民爆光电以 15.95%录得上周最大涨幅，天源迪科以-13.10%录得上周最大跌幅。AI 应用指数内部，美迪凯以 51.04%录得上周最大涨幅，电科数字以-14.10%录得上周最大跌幅。

图表 12：上周（2026.6.1-2026.6.5 日）指数日涨跌幅



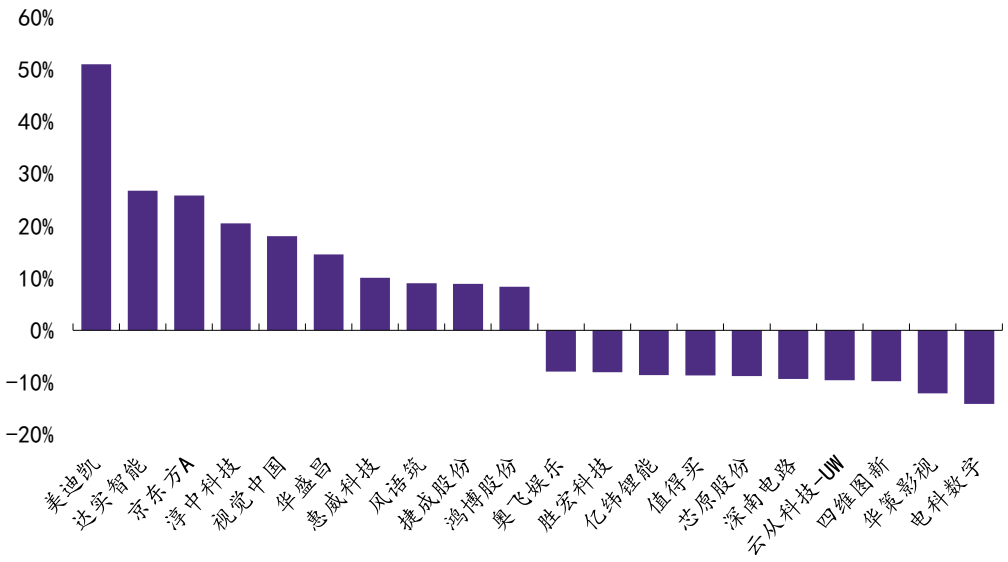
资料来源：wind,华鑫证券研究

图表 13：上周（2026.6.1-2026.6.5 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 14：上周（2026.6.1-2026.6.5 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 6 月 2 日，英伟达宣布其 Spectrum-X 以太网硅光技术已全面量产。新一代 Spectrum-X 交换机基于光电一体封装技术（CPO）构建，支持其 VeraRubin 平台在数据中心实现横向扩展与跨区域扩展，为大规模 AI 工厂部署提供网络支撑。公司通过与台积电、SPIL、Tower 及富士康的深度协同实现 Spectrum-X 以太网硅光技术的量产，四家企业分别在硅光芯片制造、芯片级封装测试、激光芯片与光模组、系统组装环节提供核心技术支持。作为英伟达全栈协同设计的典范，该技术相较传统收发器网络实现能效与 AI 集群正常运行时间均提升 5 倍，部署效率提升 30%，为百万 GPU 级 AI 工厂奠定了坚实的网络基础，目前已获得 CoreWeave、Lambda 及 Oracle Cloud Infrastructure 的率先采用。其大规模 CPO 部署突破了光互连在功耗、可靠性与部署时间方面的瓶颈，消除了制约 AI 集群规模扩张的关键障碍。光通信是英伟达战略布局的核心方向之一，在本周举办的 Computex2026 上，光互连领域龙头企业迈威尔科技首席执行官与黄仁勋同台出席。黄仁勋表示迈威尔科技有望成为下一家市值突破万亿美元的科技企业，并透露双方正进一步深化战略合作关系，共同打造支撑下一代人工智能数据中心运行的关键网络与连接基础设施体系。

2026 年截至 6 月，英伟达已密集对四家美国行业龙头企业进行大规模投资：3 月分别向迈威尔科技、Lumentum 及 Coherent 各注资 20 亿美元，其中与迈威尔科技的合作旨在将其定制 AI 芯片和网络技术整合进英伟达 NVLink 的投资则全面押注光互联技术与封装集成；5 月宣布与康宁达成多年期商业与技术合作伙伴关系，总投资上限 32 亿美元，支持其将美国光连接制造能力提升 10 倍、光纤产量提升超 50%，并在北卡罗来纳州和得克萨斯州新建三座先进制造工厂。英伟达作为全球人工智能产业绝对龙头，其全栈式技术协同与产业链资本投入，构成了行业长期增长最坚实的确定性基础。Spectrum-X 硅光互联平台全面量产及共封装光学(CPO)架构大规模商用，标志着光通信行业进入技术迭代与需求爆发的共振期。AI 算力建设需求的加速将驱动光通信板块景气度持续上行。

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 15：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	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	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 16：重点关注公司及盈利预测

公司代码	名称	2026-06-10	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	627.01	-0.30	0.30	0.60	-2090.03	2090.03	1045.02	买入
301196.SZ	唯科科技	158.40	2.53	3.34	3.98	62.61	47.43	39.80	买入
603859.SH	能科科技	52.95	0.92	1.21	1.50	57.55	43.76	35.30	买入
688615.SH	合合信息	125.25	3.24	4.22	5.25	38.66	29.68	23.86	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。