2026 年 04 月 08 日

华鑫证券 CHINA FORTUNE SECURITIES

英伟达推出 AVO 智能体技术，Gemma 4 开启端侧智能新纪元

—计算机行业周报

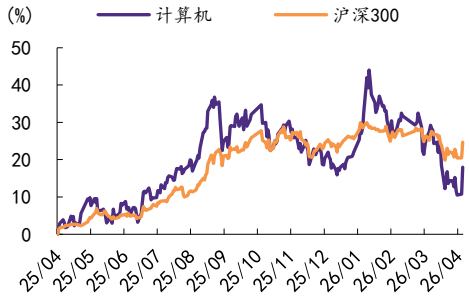
推荐(维持)

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)	-6.0	-8.3	22.5
沪深 300	-1.4	-3.0	25.9

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：英伟达（NVDA）：Blackwell 量产驱动业绩增长，数据中心仍是 AI 主线》2026-04-03

2、《计算机行业点评报告：DUOL：用户增长与订阅变现共振，AI 产品线继续抬升长期空间》2026-04-03

3、《计算机行业点评报告：Spotify：用户与盈利双升，广告修复与产品扩张共振》2026-04-02

投资要点

算力：算力租赁价格平稳，英伟达推出 AVO 智能体技术

2026 年 3 月 26 日，英伟达发布智能体式变异算子 AVO 技术，以自主编码智能体替代传统进化搜索方式，构建全流程自主优化闭环。该技术在 BlackwellB200GPU 上对注意力内核自主迭代优化，性能显著超越 cuDNN 与 FlashAttention-4，可快速迁移至 GQA 场景。

AI 应用：Discord 周访问量环比+3.52%，Gemma 4 开启端侧智能新纪元

2026 年 4 月 3 日，谷歌 DeepMind 发布了全新的开源模型系列 Gemma 4。Gemma 4 系列基于与 Gemini3 同源的技术打造，涵盖了从适合端侧设备的 2B 版本到可在单张计算卡上运行的高性能 31B 版本，四个尺寸全部开源，试图实现对不同部署场景的全覆盖。

AI 融资动向：Starcloud 完成 1.7 亿美元 A 轮融资

2026 年 3 月，太空数据中心企业 Starcloud 完成 1.7 亿美元 A 轮融资，估值达 11 亿美元，成为 YC 成长最快的独角兽，累计融资 2 亿美元。本轮由 Benchmark 与 EQT 领投，资金用于卫星研发、制造基地建设、团队扩张及发射采购。

投资建议

2026 年 3 月 31 日，智谱发布上市后首份 2025 年全年业绩财报。财报显示，公司全年实现总收入 7.24 亿人民币，同比增长 131.9%，稳居国内大模型公司收入规模首位，综合毛利率达 41%，远超行业水准；MaaS 商业飞轮全面运转，MaaSAPI 平台实现 ARR17 亿元（约 2.5 亿美金），同比提升 60 倍，毛利率同比提升近 5 倍至 18.9%，盈利能力显著改善。业务拆分看，企业级通用大模型业务实现收入 3.66 亿元，占总收入的 50.4%。企业级智能体业务收入从上年的 0.47 亿元增至 1.66 亿元，增长 248.8%，收入占比达 22.9%。开放平台及 API 平台业务从上年的 0.48 亿元增至 1.90 亿元，增幅高达 292.6%，收入占比提升至 26.3%。截至 2026 年 3 月，智谱平台的注册企业及用户已突破 400 万，服务全球超 218 个国家及地区，其 GLMCodingPlan 付费开发者超 24.2 万，2026 年 3 月推出的 ClawPlan 上线 20 天订阅用户突破 40 万。当前的 GLM 模型已

全面部署于 Google VertexAI、AWSBedrock、Fireworks、Cerebras 等全球顶尖云服务商，并入驻 OpenRouter、Vercel 等国际主流模型聚合平台，智谱已成为国内付费 Token 消耗量最高的厂商之一。从技术层面来看，公司依托自研 Slime 框架提升异步强化学习效率，为 GLM-5-Turbo 模型研发提供支撑；同时 GLM-5 实现国产芯片软硬协同优化，通过量化策略降低显存占用与部署成本，在国产硬件平台达成国际顶级芯片等效推理性能，形成技术与算力自主可控的一体化体系。公司的模型性能优势显著，2026 年一季度 API 提价 83%后调用量仍保持增长，客户需求持续旺盛。本次智谱最新财报验证了模型能力向商业价值的有效转化，其 API 大幅提价后调用量仍增，验证了优质模型是应用层付费的基础。随着行业从同质化接入转向依托强基座的差异化落地，头部应用的盈利空间将被打开，业绩兑现与毛利改善的确定性将显著提升。基于此，我们维持对 AI 应用板块的看好。

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	2026-04-08	EPS			PE			投资评级
		股价	2024	2025E	2026E	2024	2025E	2026E	
300757.SZ	罗博特科	475.28	0.41	-0.30	0.30	1159.22	-1584.27	1584.27	买入
301196.SZ	唯科科技	104.48	1.76	2.53	3.34	59.36	41.30	31.28	买入
603859.SH	能科科技	38.82	0.78	0.96	1.18	49.77	40.44	32.90	买入
688615.SH	合合信息	179.59	4.01	3.24	4.22	50.54	55.43	42.56	买入

资料来源：Wind，华鑫证券研究

正文目录

1、 算力动态：算力租赁价格平稳，英伟达推出 AVO 智能体技术 4
1.1、 Tokens 跟踪 4
1.2、数据跟踪：全球云厂商集体涨价 5
1.3、产业动态：英伟达推出 AVO 智能体技术 6
2、 AI 应用动态：DISCORD 周访问量环比+3.52%，GEMMA 4 开启端侧智能新纪元 8
2.1、 周流量跟踪：Discord 周访问量环比+3.52% 8
2.2、 产业动态：从口袋到工作站全覆盖，Gemma 4 开启端侧智能新纪元 8
3、 AI 融资动向：GRANOLA 完成 1.25 亿美元 C 轮融资 11
4、 行情复盘 12
5、 投资建议 14
6、 风险提示 15

图表目录

图表 1：TOKENS 规模 LEADERBOARD 4
图表 2：市场份额占据示意 5
图表 3：EVO 与 AVO 对比差异示意图 6
图表 4：AVO 原理示意图 6
图表 5：MHA 结果对比示意图 7
图表 6：2026.3.28-2026.4.3AI 相关网站流量 8
图表 7：GEMMA 4 在 MODELPERFORMANCEVSSIZE 与 ARENAMODELRANKINGS 中的表现 9
图表 8：GEMMA 4 在多个测试中的表现 9
图表 9：GEMMA 4 包含的四款模型 10
图表 10：上周 AI 初创公司融资动态 11
图表 11：上周（2026.3.30-2026.4.3 日）指数日涨跌幅 12
图表 12：上周（2026.3.30-2026.4.3 日）AI 算力指数内部涨跌幅度排名 12
图表 13：上周（2026.3.30-2026.4.3 日）AI 应用指数内部涨跌幅度排名 13
图表 14：FICONTEC2025 年年中至今公告订单 14
图表 15：重点关注公司及盈利预测 15

1、算力动态：算力租赁价格平稳，英伟达推出 AVO 智能体技术

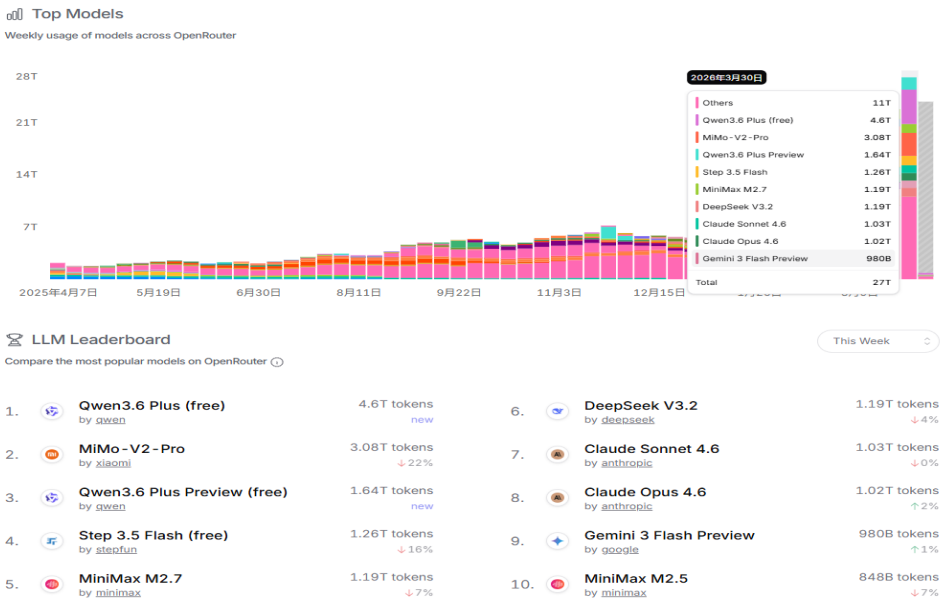
1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 3 月 30 日至 2026 年 4 月 5 日，周度 token 消耗量有所上升，调用量为 27T，环比上周 18.94%。在 tokens 规模 leaderboard 前五名中，qwen 的 Qwen3.6Plus(free)以 4.6Ttokens 位居榜首，xiaomi 旗下的 MiMo-V2-Pro 以 3.08Ttokens 位居第二；qwen 的 Qwen3.6Plus(free)以 1.64T 位列第三；stepfun 的 Step3.5Flash 的以 1.26T 位列第四；minimax 旗下的 MiniMaxM2.7 以 1.19Ttokens 位居第五。

从市场份额维度来看，qwen 以 1.78Ttokens 占据 39.6%的份额，稳居首位；google 以 430B 占据 9.5%，位列第二；OpenAI、Anthropic、Minimax 则分别以 397B、372B、328Btokens，对应占据 8.8%、8.2%、7.3%的市场份额。

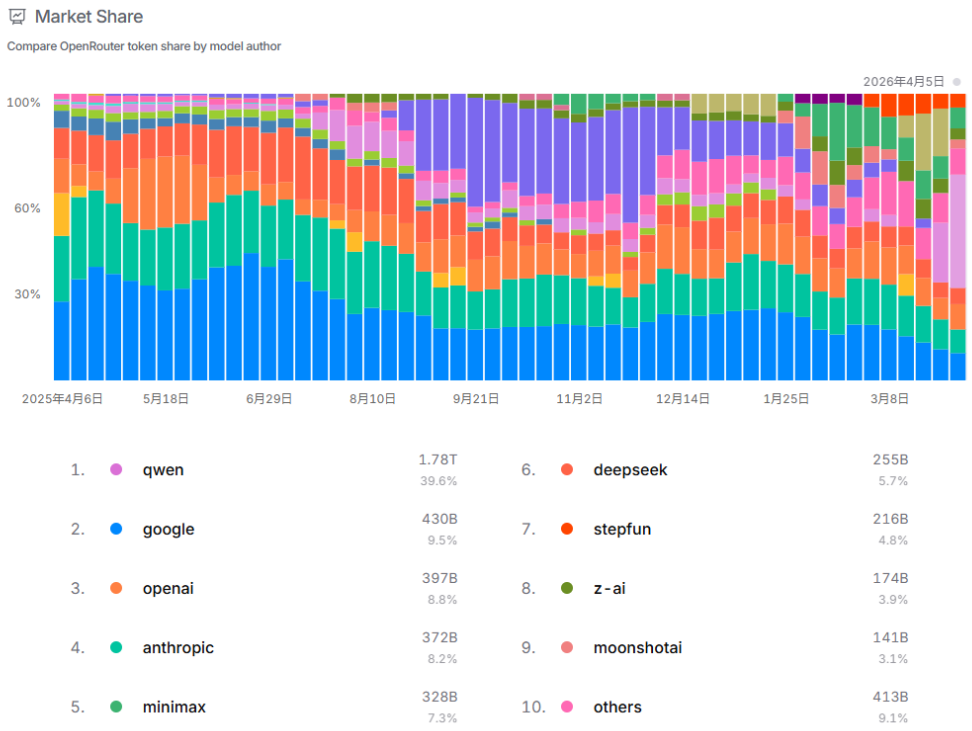
据 OpenRouter 最新数据，中国 AI 大模型周调用量已连续五周超过美国，在 3 月 30 日-4 月 5 日这一周达到 12.96 万亿 Token，环比涨幅 31.48%，而同期美国为 3.03 万亿 Token，中国是美国的 4.28 倍；该时段全球大模型周调用量榜单中，前 6 名均为中国模型。小米 MiMo-V2-Pro 于 3 月 31 日公布最新成绩，在 OpenRouter 平台周 Token 消耗量达 4.19 万亿，拿下日、周、月三榜榜首；该模型在 TextArena 评测跻身全球前五，小米宣布未来三年 AI 领域投入超 600 亿元。4 月 1 日，Qwen3.6-Plus 发布仅 1 天，日调用量突破 1.4 万亿 Token，成为 OpenRouter 首个单日[illegible]Token 的模型，刷新全球单日单模型调用纪录，其智能体与编程能力表现突出。两款国产模型接连破纪录，彰显中国大模型在全球应用市场的领先地位。

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：全球云厂商集体涨价

近期国内云计算市场迎来大规模调价浪潮，3 月以来阿里云、腾讯云、百度智能云等头部厂商相继上调 AI 算力、云存储及大模型服务价格，部分产品涨幅最高超 400%，AI 算力紧缺与硬件成本上涨推动行业进入集体涨价周期，价格战逻辑逐步转向成本传导式定价。在此行业背景下，火山引擎与京东云接连释放明确信号，选择逆势坚守价格稳定，与行业趋势形成鲜明对比。

京东云方面，公司延续 3 月公开承诺，全系核心云产品价格保持稳定，不新增涨价项；同时对数据库、中间件等多款 PaaS 产品推出专项优惠，平均降幅超 16%，最高降幅达 40%，依托低成本运营体系支撑普惠低价策略，承诺 7×24 小时技术支持等服务质量不打折。

火山引擎方面，4 月 2 日，公司总裁谭待在媒体采访中回应近期云涨价潮时明确表示，涨价只是部分厂商的行为。火山引擎对每一代模型的定价均经过严谨设计，定价后长期保持稳定，企业客户更应关注端到端完成业务的整体成本，而非单纯追求单 Token 低价，低能力模型即便单价低，也会因 Token 消耗过量造成浪费。

背景来看，2026 年 3 月，阿里云、百度智能云、腾讯云已相继官宣上调 AI 算力、存储及大模型服务价格，行业进入结构性调价周期。此次火山引擎、京东云的差异化定价策略，标志着国内云计算市场告别单一价格竞争，转向价值定价、成本可控的高质量发展新方向。

1.3、产业动态：英伟达推出 AVO 智能体技术

2026 年 3 月 26 日，英伟达发布全新研究成果，正式提出智能体式变异算子（AVO），该技术以自主编码智能体替代传统进化搜索中的固定变异、交叉与人工启发式方法，实现了无需人工干预的自主进化优化，在 GPU 算子优化领域展现出超人类智能的表现，成为软件工程“盲编程”发展的重要里程碑。

传统基于大语言模型的进化搜索框架，仅能将模型作为固定流程内的候选代码生成器，无法主动查阅资料、测试代码、解读反馈与提前修正策略，难以适配需深度迭代的顶级硬件优化任务。AVO 突破这一局限，将自主导向的代码智能体升级为进化操盘手，智能体可自主查阅历史方案、领域知识库与评估工具，自主规划、实施、测试并修正代码，在长周期运行中持续迭代改进。

图表 3：EVO 与 AVO 对比差异示意图

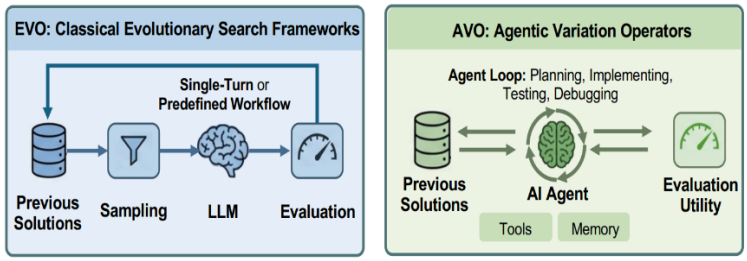


Figure 1: **EVO vs AVO**: Comparison between prior evolutionary search frameworks (e.g. FunSearch, AlphaEvolve, and related LLM-augmented evolutionary approaches) and the proposed Agentic Variation Operator. **Left:** Prior approaches follow a fixed pipeline where the LLM is confined to a single-turn generation step or a predefined workflow, with sampling and evaluation controlled by the framework. **Right:** AVO replaces this pipeline with an autonomous AI agent that iteratively plans, implements, tests, and debugs across long-running sessions, with direct access to previous solutions, evaluation utilities, tools, and persistent memory.

资料来源：机器之心，华鑫证券研究

在技术实现上，AVO 以深度智能体替代固定变异流程，依托规划能力、持久记忆与工具调用能力，形成完整自主闭环：智能体可读取 CUDA 编程指南、PTX 架构文档等专业资料，基于性能分析器定位瓶颈，自主完成代码修改、编译运行、正确性校验、性能评估与问题修复，全程无需人工介入。

图表 4：AVO 原理示意图

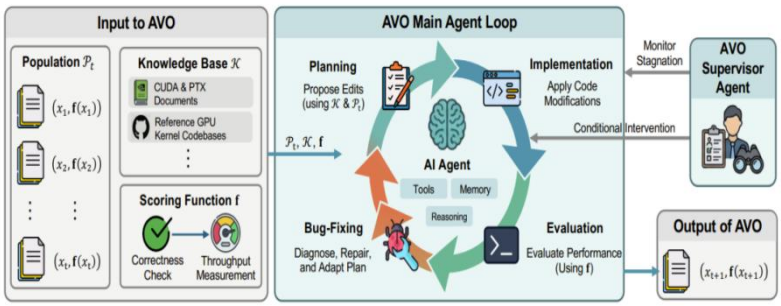


Figure 2: Illustration of the Agentic Variation Operator (AVO).

资料来源：机器之心，华鑫证券研究

在 NVIDIABlackwellB200GPU 的多头注意力（MHA）内核优化任务中，AVO 在无人工干预的情况下连续自主运行 7 天，探索超 500 个优化方向，迭代出 40 个内核版本，在 BF16 精度

下实现 1668TFLOPS 的吞吐量，较英伟达 cuDNN 库最高提升 3.5%，较 FlashAttention-4 最高提升 10.5%。该优化技术具备出色泛化能力，仅需 30 分钟自主适配即可迁移至分组查询注意力（GQA）场景，性能较 cuDNN 最高提升 7.0%，较 FlashAttention-4 最高提升 9.3%。

图表 5：MHA 结果对比示意图

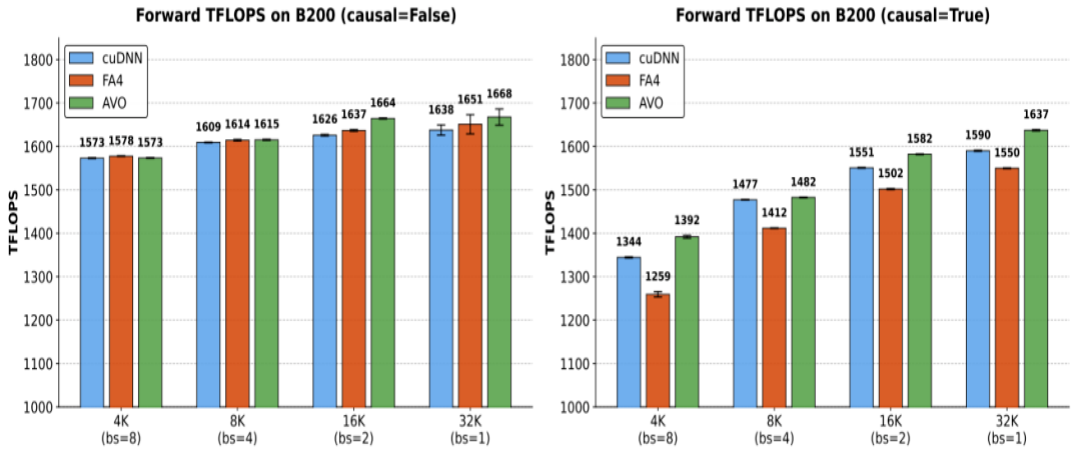


Figure 3: Multi-head attention forward-pass prefilling throughput (TFLOPS) on NVIDIA B200 with head dimension 128, 16 heads, and BF16 precision. Batch size and sequence length are varied with a fixed total of 32k tokens.

资料来源：机器之心，华鑫证券研究

AVO 智能体实现了硬件级底层推理，优化覆盖寄存器分配、指令流水线调度、负载分布等内核设计核心层面。其关键优化技术包括三项：一是无分支累加器重缩放，通过消除条件分支、替换轻量级内存屏障，消除 warp 同步开销，使非因果注意力吞吐量提升 8.1%；二是纠错与张量核心（MMA）流水线重叠，重构执行依赖关系，将顺序执行转为交叠执行，减少硬件空闲等待；三是跨 warp 组寄存器重新平衡，依据性能分析数据重新分配 2048 个寄存器预算，解决运算组寄存器不足导致的内存溢出问题，进一步提升 2.1%性能。

AVO 不仅突破了 GPU 算子优化的瓶颈，更为 AI 芯片、深度学习底层生态及各类算力密集型科学工程领域的自动化系统优化，开辟了全新技术路径。

2、AI 应用动态：Discord 周访问量环比+3.52%，Gemma 4 开启端侧智能新纪元

2.1、周流量跟踪：Discord 周访问量环比+3.52%

本期（2026.3.28-2026.4.3）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1254.0M）、Bing（812.9M）和 Gemini（604.7M），访问量环比增速第一为 Discord（3.52%）；平均停留时长前三位分别为 Character.AI（00:17:12）、Discord（00:10:54）和 Kimi（00:08:16）;平均停留时长环比增速第一为 QuillBot（0.60%）。

图表 6：2026.3.28-2026.4.3AI 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	1254.0	-2.03%	5:51	0.00%
Bing	搜索	微软	812.9	-0.68%	7:55	-0.21%
Gemini	聊天机器人	谷歌	604.7	1.04%	7:11	-0.69%
Canva	在线设计	Canva	209.9	-3.67%	5:56	0.00%
Github	代码托管	微软	148.1	2.49%	6:38	0.00%
Discord	游戏社区	微软	141.3	3.52%	10:54	0.46%
Character.AI	聊天机器人	Character.AI	42.92	3.22%	17:12	-0.39%
NotionAI	文本/笔记	Notion	41.14	0.34%	8:01	0.00%
Perplexity	AI 搜索	Perplexity	36.05	-2.33%	4:45	0.35%
DeepI	翻译工具	DeepL	27.37	-2.91%	2:25	0.00%
QuillBot	释义工具	QuillBot	10.92	-0.27%	2:49	0.60%
Kimi	聊天机器人	Moonshot AI	9.27	-4.38%	8:16	-2.55%
文心一言	聊天机器人	百度	0.54	-13.06%	2:33	-2.55%

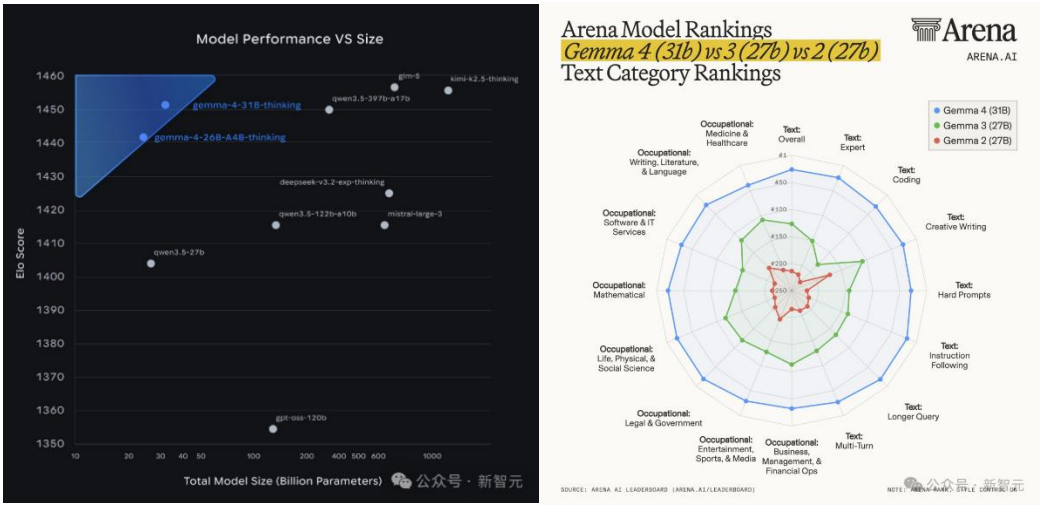
资料来源：similarweb,华鑫证券研究

2.2、产业动态：从口袋到工作站全覆盖，Gemma 4 开启端侧智能新纪元

2026 年 4 月 3 日，谷歌 DeepMind 发布了全新的开源模型系列 Gemma 4。Gemma 4 系列基于与 Gemini3 同源的技术打造，涵盖了从适合端侧设备的 2B 版本到可在单张计算卡上运行的高性能 31B 版本，四个尺寸全部开源，试图实现对不同部署场景的全覆盖。

Gemma 4 在各项能力上出现了显著提升，其中 31B Dense 在 Arena AI 文本榜单上取得了开源第三的成绩，Elo 评分达到 1452。排在前面的两个模型参数量分别超过 600 亿和 1000 亿，这意味着 Gemma 4 仅以 31B 的规模进入了原本属于更大体量模型的竞争区间。此外，26B 参数的 MoE 模型在推理时仅激活 38 亿参数，Elo 评分达到 1441，位列开源第六。这种参数效率的提升，使得小体量模型能够与数倍于自身规模的对手抗衡。

图表 7：Gemma 4 在 Model Performance VS Size 与 Arena Model Rankings 中的表现



资料来源：新智元，华鑫证券研究

从具体任务的表现来看，Gemma 4 相比前代几乎实现了全面压制。在数学推理方面，AIME 2026 测试中 Gemma 4 取得了 89.2%的成绩，而 Gemma 3 仅为 21.2%，提升了 68 个百分点。在编程能力上，LiveCodeBench 测试中 Gemma 4 达到 80%，前代只有 29.1%，差距明显。在智能体相关的 t2-bench 测试中，Gemma 4 获得了 86.4%，而 Gemma 3 仅有 6.6%。此外，在多语言推理和知识问答等基准测试中，Gemma 4 也普遍实现了约 40%的性能增长。

图表 8：Gemma 4 在多个测试中的表现

Benchmark		Gemma 4 31B IT Thinking	Gemma 4 26B A4B IT Thinking	Gemma 4 E4B IT Thinking	Gemma 4 E2B IT Thinking	Gemma 3 27B IT
Arena AI (text) As of 4/2/26		1452	1441	-	-	1365
MMMLU Multilingual Q&A	No tools	85.2%	82.6%	69.4%	60.0%	67.6%
MMMU Pro Multimodal reasoning		76.9%	73.8%	52.6%	44.2%	49.7%
AIME 2026 Mathematics	No tools	89.2%	88.3%	42.5%	37.5%	20.8%
LiveCodeBench v6 Competitive coding problems		80.0%	77.1%	52.0%	44.0%	29.1%
GPQA Diamond Scientific knowledge	No tools	84.3%	82.3%	58.6%	43.4%	42.4%
τ2-bench Agentic tool use	Retail	86.4%	85.5%	57.5%	29.4%	6.6%

资料来源：新智元，华鑫证券研究

Gemma 4 系列的四个模型各有侧重。每个尺寸都提供了 base 和 instruction-tuned 两个版本。其中 E2B 和 E4B 主要面向端侧设备，与谷歌 Pixel 团队、高通、联发科进行了联合优化，使得这些模型能够在手机、树莓派、Jetson Orin Nano 等设备上离线运行，且延迟接近零。而 31B 和 26B 版本则面向开发者工作站和服务器，31B 追求极致的输出质量，26B 则依靠 MoE 架构在推理时仅激活少量参数，从而实现更快的响应速度，特别适合需要低延迟的智能体应用场景。对于开发者而言，31B 模型的 bfloat16 权重可以放入一张 80GB 显存的 H100 加速卡中，量化后的版本可以在消费级显卡上运行。值得一提的是，Gemma 4 还支持 TurboQuant 压缩算法。

图表 9：Gemma 4 包含的四款模型

模型	有效参数	总参数	上下文窗口	定位
Gemma 4 E2B	23亿	51亿	128K	手机/IoT
Gemma 4 E4B	45亿	80亿	128K	手机/边缘设备
Gemma 4 26B MoE	38亿激活	260亿	256K	低延迟推理
Gemma 4 31B Dense	310亿	310亿	256K	最强质量/微调基座

资料来源：新智元，华鑫证券研究

在架构设计方面，Gemma 4 高效整合了三个经过验证的技术。首先是逐层嵌入技术，传统 Transformer 中每个 token 在输入层获得一个嵌入向量后，后续所有层都基于这个初始表示进行计算，这要求嵌入层一次性打包所有信息。逐层嵌入则为每一层都配备了一个专属的低维信号通道，每个 token 在每一层都能收到由 token 身份信息和上下文信息共同生成的定制化向量。由于 PLE 的维度远小于主隐藏层，计算开销很小，但每一层都获得了专属的调节能力，这一设计在小模型上效果尤其明显。其次是 KV 共享，最后几层不再自己计算 Key 和 Value 投影，而是直接复用前面层的 KV 张量，同类型的注意力层共享同一组 KV 状态。这使得推理时的显存占用和计算量都有所下降，对长上下文生成和端侧部署尤为有利。第三是交替注意力机制，模型交替使用局部滑动窗口注意力和全局全上下文注意力，小模型采用 512token 的滑动窗口，大模型采用 1024token。这三个设计的共同目标都是让每一个参数尽可能高效地被利用。

Gemma 4 全系列都能够处理图像和视频输入，其中 E2B 和 E4B 还额外兼容音频。相比上一代，视觉编码器做了两个关键升级：一是支持可变宽高比，不再强制裁切图片；二是提供了五档可配置的图像 token 预算，开发者可以根据场景在速度和精度之间自由取舍，低预算适合图像分类和描述，高预算适合 OCR 和文档解析。在实际测试中，Gemma 4 展示了多项多模态能力。在 GUI 元素检测任务中，给定一张网页截图并询问某个按钮的位置，四个尺寸的模型都能以 JSON 格式返回精确的边界框坐标，无需特殊提示词。在视频理解方面，用一段现场演唱会视频测试，E4B 不仅能准确描述舞台画面，还能从音轨中提取歌词主题；26B 和 31B 虽然不具备音频输入能力，但对纯视觉内容的理解同样准确，甚至能识别出屏幕上显示的赞助商品牌名。在音频转写任务中，E4B 对英文演讲的转写几乎完美，标点和断句都很自然，E2B 偶尔会出现幻觉但整体可用。在多模态函数调用方面，给出一张曼谷寺庙的照片并询问所在城市及当地天气，四个尺寸都正确识别出曼谷，并自动调用了天气查询工具。这种函数调用能力是从训练阶段就内置的，基于此前发布的 FunctionGemma 研究成果，能够处理多轮多工具的智能体工作流，与以往依靠提示词引导模型进行工具调用的方式有本质区别。

3、AI 融资动向：Granola 完成 1.25 亿美元 C 轮融资

2026 年 3 月 31 日，太空数据中心企业 Starcloud 宣布完成 1.7 亿美元 A 轮融资，估值达 11 亿美元，成为 YC 历史上成长最快的独角兽。本轮由 Benchmark 与 EQT 联合领投，麦格理资本、NFX、YC、776Ventures 等参投，叠加此前融资，公司累计融资达 2 亿美元。资金将用于 Starcloud-3 卫星研发、制造基地建设、团队扩张及发射合同采购，加速轨道算力基础设施落地。

Starcloud 专注近地轨道太空数据中心，攻克太空 AI 算力的能源与冷却瓶颈，利用太空无限太阳能实现可持续计算。公司已创造多项行业第一：成功将 NVIDIAH100 送入轨道、完成首次太空 AI 训练与轨道推理，算力提升百倍。年内将发射 Starcloud-2，搭载超大散热器，发电能力为初代 100 倍，可承载商业云与边缘负载，并开展太空比特币挖矿，已与 AWS、GoogleCloud、NVIDIA 建立合作。

作为太空计算赛道先行者，Starcloud 仅用 21 个月完成卫星从设计到发射，依托技术突破与高效执行力领跑市场。本轮融资后，公司将进一步扩大在轨算力规模，破解地球数据中心能耗与用地限制，为 AI 产业提供全新算力供给路径，引领云计算迈向太空时代。

图表 10：上周 AI 初创公司融资动态

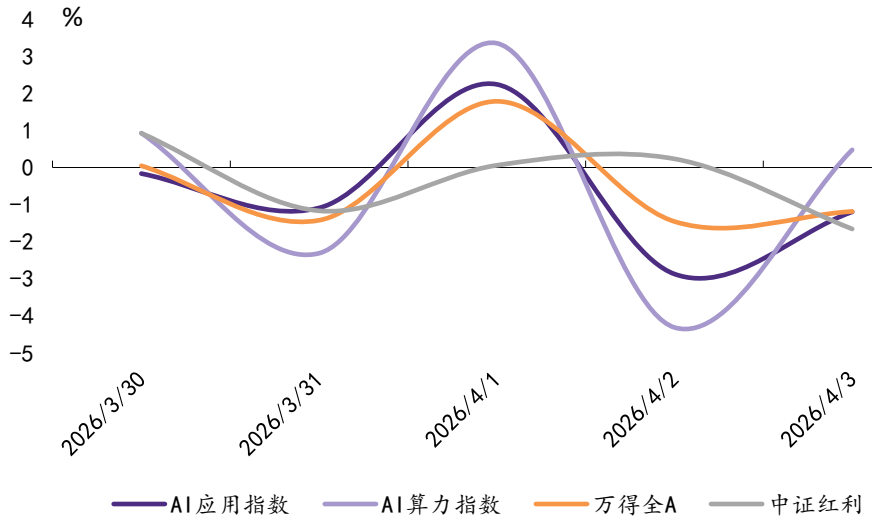
应用	应用类型	领投方	融资轮	融资额	目前累计融资额	目前估值
Starcloud	AI 算力基础设施	Benchmark、EQT	A 轮	1.7 亿美元	2 亿美元	11 亿美元
Replit	AI 技术服务	Georgian	D 轮	4 亿美元	4 亿美元	90 亿美元
GranolaAI	AI 企业服务	IndexVentures	C 轮	1.25 亿美元	1.92 亿美元	15 亿美元

资料来源：wind，Saasverse，华鑫证券研究

4、 行情复盘

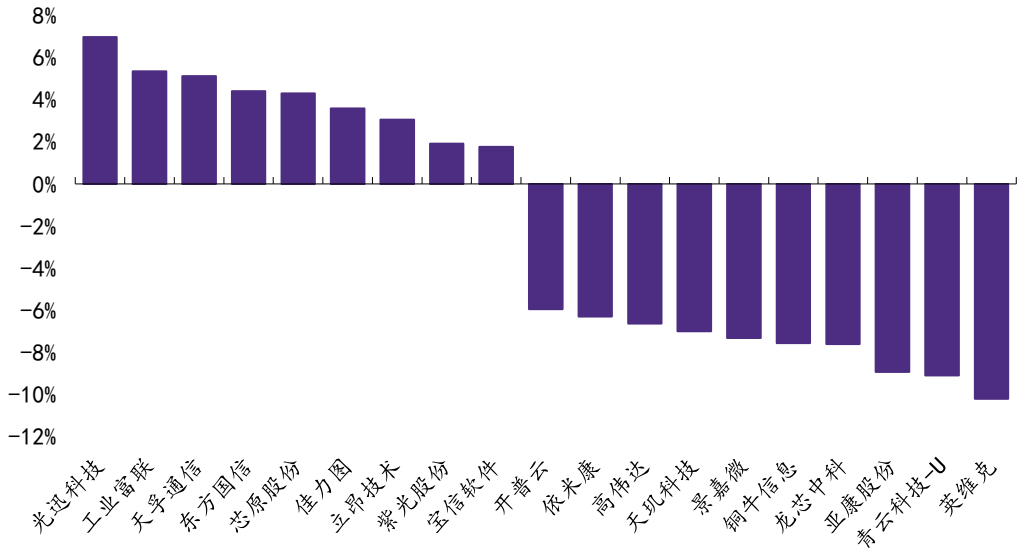
上周（2026.3.30-2026.4.3 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 2.24%/3.35%/1.78%/0.92%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-2.87%/-4.31%/-1.46%/-1.66%。AI 算力指数内部，光迅科技以 6.97%录得上周最大涨幅，英维克以-10.2%录得上周最大跌幅。AI 应用指数内部，华盛昌以 15.8%录得上周最大涨幅，亿纬锂能以-17.13%录得上周最大跌幅。

图表 11：上周（2026.3.30-2026.4.3 日）指数日涨跌幅



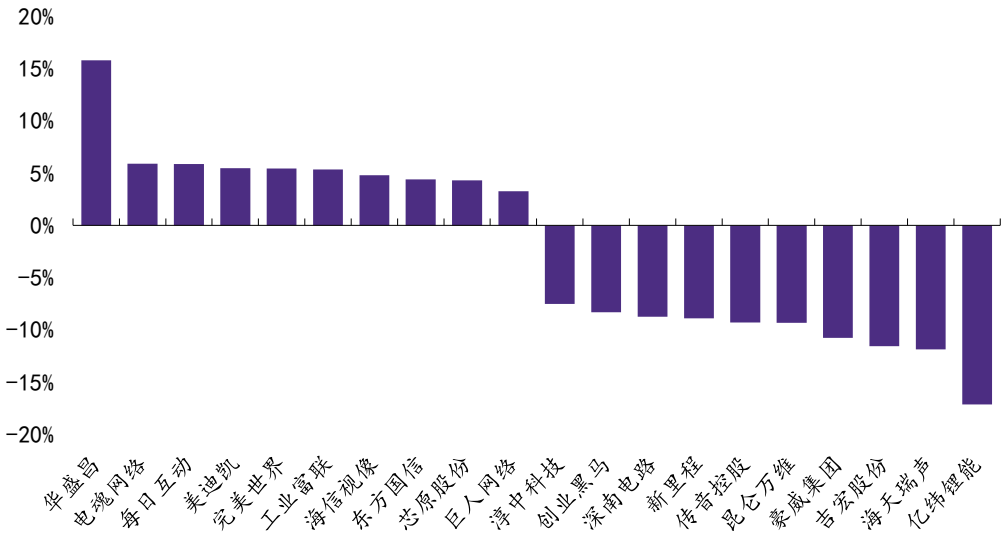
资料来源：wind，华鑫证券研究

图表 12：上周（2026.3.30-2026.4.3 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（2026.3.30-2026.4.3 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 3 月 31 日，智谱发布上市后首份 2025 年全年业绩财报。财报显示，公司全年实现总收入 7.24 亿人民币，同比增长 131.9%，稳居国内大模型公司收入规模首位，综合毛利率达 41%，远超行业水准；MaaS 商业飞轮全面运转，MaaSAPI 平台实现 ARR17 亿元（约 2.5 亿美金），同比提升 60 倍，毛利率同比提升近 5 倍至 18.9%，盈利能力显著改善。业务拆分看，企业级通用大模型业务实现收入 3.66 亿元，占总收入的 50.4%。企业级智能体业务收入从上年的 0.47 亿元增至 1.66 亿元，增长 248.8%，收入占比达 22.9%。开放平台及 API 平台业务从上年的0.48亿元增至1.90亿元，增幅高达292.6%，收入占比提升至26.3%。截至 2026 年 3 月，智谱平台的注册企业及用户已突破 400 万，服务全球超 218 个国家及地区，其 GLMCodingPlan 付费开发者超 24.2 万，2026 年 3 月推出的 ClawPlan 上线 20 天订阅用户突破40 万。当前的GLM模型已全面部署于Google VertexAI、AWSBedrock、Fireworks、Cerebras 等全球顶尖云服务商，并入驻 OpenRouter、Vercel 等国际主流模型聚合平台，智谱已成为国内付费 Token 消耗量最高的厂商之一。从技术层面来看，公司依托自研 Slime 框架提升异步强化学习效率，为 GLM-5-Turbo 模型研发提供支撑；同时 GLM-5 实现国产芯片软硬协同优化，通过量化策略降低显存占用与部署成本，在国产硬件平台达成国际顶级芯片等效推理性能，形成技术与算力自主可控的一体化体系。公司的模型性能优势显著，2026 年一季度 API 提价 83%后调用量仍保持增长，客户需求持续旺盛。本次智谱最新财报验证了模型能力向商业价值的有效转化，其 API 大幅提价后调用量仍增，验证了优质模型是应用层付费的基础。随着行业从同质化接入转向依托强基座的差异化落地，头部应用的盈利空间将被打开，业绩兑现与毛利改善的确定性将显著提升。基于此，我们维持对 AI 应用板块的看好。

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
	总金额			约 11.44 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	2026-04-08	EPS			PE			投资评级
		股价	2024	2025E	2026E	2024	2025E	2026E	
300757.SZ	罗博特科	475.28	0.41	-0.30	0.30	1159.22	-1584.27	1584.27	买入
301196.SZ	唯科科技	104.48	1.76	2.53	3.34	59.36	41.30	31.28	买入
603859.SH	能科科技	38.82	0.78	0.96	1.18	49.77	40.44	32.90	买入
688615.SH	合合信息	179.59	4.01	3.24	4.22	50.54	55.43	42.56	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&主题&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

陶欣怡：毕业于上海交通大学，于 2023 年 10 月加入团队。

倪汇康：金融学士，2025 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。