2026 年 04 月 27 日

华鑫证券 CHINA FORTUNE SECURITIES

OpenAI 发布 GPT-5.5 旗舰大模型，DeepSeek V4 发布

—计算机行业周报

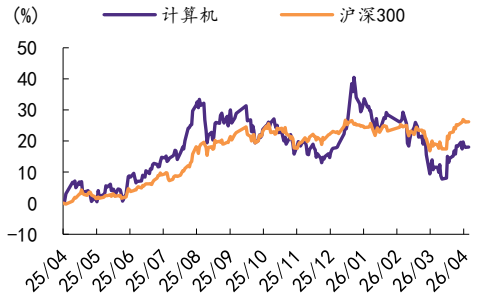
推荐(维持)

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)	5.8	-10.0	17.5
沪深 300	6.0	1.4	26.0

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：ClaudeOpus4.7 深夜上线，Terafab 项目加速落地》2026-04-21

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

3、《计算机行业周报：英伟达推出 AVO 智能体技术，Gemma4 开启端侧智能新纪元》2026-04-08

投资要点

算力：算力租赁价格平稳，OpenAI 发布 GPT-5.5 旗舰大模型

2026 年 4 月 24 日，OpenAI 正式推出旗舰大模型 GPT-5.5，主打 AI 智能体原生能力，综合性能登顶各类基准榜单，全面超越竞品模型。该模型在代码开发、办公实操、前沿科研等领域实现大幅升级，可承接复杂工程与高端科研任务。

AI 应用：Kimi 周访问量环比+10.80%，DeepSeek V4 发布

2026 年 4 月 24 日，DeepSeek 正式发布了其最新一代模型 DeepSeekV4 系列。该系列模型的核心突破集中在百万级别上下文窗口的普及化，同时在智能体能力、世界知识掌握以及逻辑推理性能方面达到了开源社区的新高度。与此同步公开的还有详细的技术报告，供研究者和开发者查阅。

AI 融资动向：Cognition 拟融资数亿美元

2026 年 4 月，AI 编程公司 Cognition 正寻求数亿美元新一轮融资，目标估值 250 亿美元。公司主打全球首个自主 AI 软件工程师 Devin，可完成规划、编码、调试、部署全流程。公司成立不足三年，ARR 快速增长至 7300 万美元，估值从 3.5 亿美元一路攀升。

投资建议

2026 年 4 月 24 日，DeepSeek 正式推出 V4 模型预览版并同步开源。本次版本迭代核心集中于超长上下文能力升级、推理架构优化及国产化算力适配，整体运行效率与成本控制能力实现显著提升。该模型同步推出 pro 与 flash 两大版本；其中 pro 总参数 1.6T、激活参数 49B、预训练数据 33T，flash 总参数 284B、激活参数 13B、预训练数据 32T，两款模型统一搭载 100 万词元超长上下文窗口，综合性能对标行业顶尖水平，能够以更低的内存资源消耗，稳定处理大篇幅长文本处理任务。V4 在底层架构层面完成全面升级，融合压缩稀疏注意力（CSA）与重度压缩注意力（HCA）混合架构，搭配 mHC 训练稳定机制及 Muon 主训练优化器，部分模块沿用 AdamW 架构协同运作，大幅优化长上下文场景下的推理运行效率。对比前代产品，v4-pro 在百万级上下文场景中，单 token 推理浮点运算量缩减至原有 27%，KV 缓存占用仅为原先 10%；

flash 优化效果更为突出，单词元推理浮点运算量下降至前代的 10%，KV 缓存占用压缩至 7%，整体运行成本得到有效控制。与此同时，deepseek 进一步推进国产化算力适配布局，深度对接华为昇腾 950 超级节点量产规划，随着 2026 年下半年该芯片产品实现大规模供货落地，v4-pro 的 API 调用定价有望迎来明显下调，进一步提升商业化落地性价比。

此次发布的 V4 深度适配华为昇腾体系，通过底层架构重构优化推理调度与缓存占用，整体算力运行效率大幅改善。当前高端算力供给紧缺，持续约束大模型服务吞吐能力。伴随下半年昇腾 950 的批量交付，高端大模型商业化面临的算力瓶颈与成本压力将有效缓解。头部模型厂商加速推进国产芯片兼容适配，持续完善算力自主可控生态，芯模协同的共振下，国产算力产业链有望进入放量周期。

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	2026-04-27	EPS			PE			投资评级
		股价	2024	2025E	2026E	2024	2025E	2026E	
300757.SZ	罗博特科	518.50	0.41	-0.30	0.30	1264.63	-1728.33	1728.33	买入
301196.SZ	唯科科技	117.37	1.76	2.53	3.34	66.69	46.39	35.14	买入
603859.SH	能科科技	39.49	0.78	0.96	1.18	50.63	41.14	33.47	买入
688615.SH	合合信息	128.00	4.01	3.24	4.22	50.54	39.51	30.33	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，OpenAI 发布 GPT-5.5 旗舰大模型

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 4 月 20 日至 4 月 26 日，周度 token 消耗量有所下降，调用量为 22T，环比上周 6.79%。在 tokens 规模 leaderboard 前五名中，moonshota 的 KimiK2.6 以 1.58Ttokens 位居榜首，Anthropic 的 ClaudeSonnet4.6 以 1.36Ttokens 位居第二，deepseek 的 DeepSeekV3.2 以 1.28Ttokens 位居第三；Anthropic 的 ClaudeOpus4.7 以 1.15T 位列第四；google 旗下的 Gemini3FlashPreview 以 1.04Ttokens 位居第五；

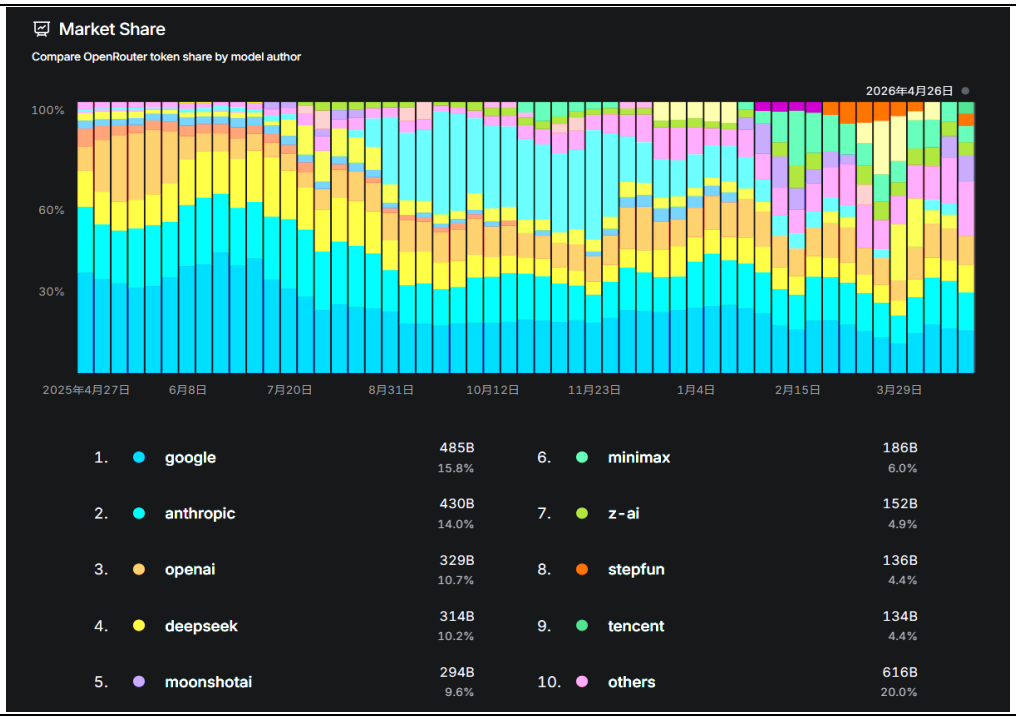
从市场份额维度来看，google 以 485Ttokens 占据 15.8%的份额，稳居首位；Anthropic 以 430B 占据 14.0%，位列第二；OpenAI、DeepSeek、moonshotai 则分别以 329B、314B、294Btokens，对应占据 10.7%、10.2%、9.6%的市场份额。

数据显示，谷歌云 AI 服务 Token 处理效率大幅升级，当前可实现每分钟 160 亿 Tokens 的运算吞吐量，较上季度增长 60%，强劲的算力底座为企业级大模型推理、多模态内容生成等高频场景提供支撑。蚂蚁百灵全新 Ling-2.6-flash 模型上线后市场反响热烈，依托轻量化高性能优势，短期调用规模爆发式增长，周度 Token 调用量涨幅突出，上线以来，其调用量持续增长，连续多日位列 Trending 榜首，日均 tokens 调用量达 100B 级别，周增长超 5000%。DeepSeek-V4 系列模型上线后，官方推出限时折扣举措，大幅下调 API 调用 Token 计价成本，极致性价比驱动用户需求集中释放。数据显示，调价后 DeepSeek-V4-Pro 单日 Token 调用量环比大幅攀升，短期实现数倍增长。4 月 25 日，DeepSeekV4-Pro 的调用量为 136 亿 Token，较前一日（4 月 24 日）增长近四倍。

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：阿里云与 DeepSeek 同步更新

近期，阿里云与 DeepSeek 先后发布重磅更新，从开发工具定价到大模型服务成本，为开发者带来了影响深远的行业变化。

阿里云率先宣布，自 2026 年 4 月 30 日 10:00 起，旗下 QoderTeams 版将调整新购价格与额度配置：每席位每月的月额度将提升至 3000Credits，同时定价调整为 300 元/席位/月，为团队级 AI 开发提供了更明确的成本预期。同时，阿里云百炼平台已首发上线 DeepSeek-V4-pro 与 DeepSeek-V4-flash 两款模型，API 价格与 DeepSeek 官网保持一致，百万 Tokens 输入最低仅需 1 元，百万 Tokens 输出最低 2 元；依托阿里云智能计算灵骏平台的算力支撑，实现了 DeepSeekV4 预览版的 Day0 稳定适配运行。

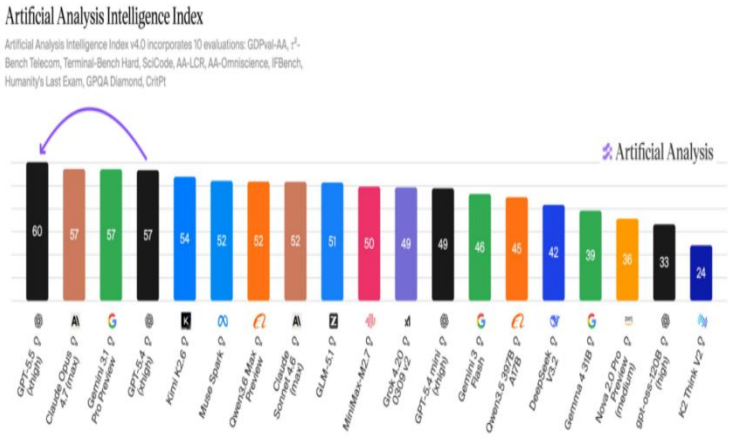
DeepSeek4 月 25 日推出限时优惠，宣布 DeepSeek-V4-Pro 开启 2.5 折限时降价活动，活动持续至 5 月 5 日 23:59。调价后，DeepSeek-V4-Pro 的百万 Tokens 输入（缓存命中）价格低至 0.25 元，输入（缓存未命中）为 3 元，输出价格为 6 元，较原价实现了大幅下调。

1.3、产业动态：OpenAI 发布 GPT-5.5 旗舰大模型

2026 年 4 月 24 日，OpenAI 正式发布旗下新一代旗舰大模型 GPT-5.5，该模型定位为 Agent 时代的原生智能大脑，专为实际工作与智能体驱动打造，已全面上线 ChatGPT 与 Codex 平台。GPT-5.5 在各项核心基准测试中斩获全榜第一，全方位超越 ClaudeOpus4.7、Gemini3.1Pro 等竞品，刷新全球大模型性能纪录，同时实现 token 效率与推理速度的双重优化，成为 OpenAI 迄今最强、最全能的 AI 模型。

GPT-5.5 在智能体编程领域实现突破性领先，Terminal-Bench2.0 全链路 Agent 工程测试中斩获 82.7%的得分，较 GPT-5.4 提升 7.6 个百分点，大幅领先 ClaudeOpus4.7 的 69.4%；OpenAI 内部 Expert-SWE 长周期编程任务测试中，模型以 73.1%的得分超越前代的 68.5%，可高效完成中位耗时 20 小时的复杂编程任务；业界通用的 SWE-BenchPro 测试中，GPT-5.5 取得 58.6%的成绩，OpenAI 指出 ClaudeOpus4.7 在该测试中存在部分问题子集过拟合迹象。依托 Codex 能力，GPT-5.5 可完成端到端全流程编程任务，从零实现 3D 交互应用、游戏开发等复杂项目，具备极强的系统形态理解与代码问题排查能力，且完成同等任务的 token 消耗量显著降低。

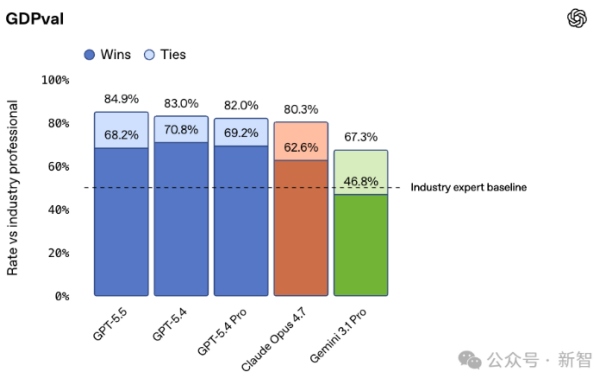
图表 3：ArtificialAnalysis 得分对比示意图



资料来源：新智元，华鑫证券研究

在知识型工作场景中，GPT-5.5 展现出顶尖的实用价值，GDPval 职业知识工作评估中斩获 84.9%，领先 ClaudeOpus4.7 的 80.3%与 Gemini3.1Pro 的 67.3%；OSWorld-Verified 真实电脑环境操作测试得分 78.7%，与 ClaudeOpus4.7 基本持平；复杂客服工作流测试 Tau2-bench 中，模型在无微调提示词的情况下达到 98.0%的准确率。OpenAI 内部超 85%的员工每周跨部门使用 Codex，公关、财务、市场等团队借助 GPT-5.5 实现数据分析、报表自动化、文档处理等工作，大幅提升办公效率。同时，Codex 可直接与 Web 应用交互，完成界面测试、屏幕截取、文件审阅等操作，跨工具上下文流转能力显著提升。

图表 4：GDPval 得分示意图



资料来源：新智元，华鑫证券研究

科研领域，GPT-5.5 实现重大突破，成功协助发现并验证拉姆齐数的全新数学证明，填

补组合数学领域研究空白；多阶段科学数据分析测试 GeneBench 中得分 25.0%，较 GPT-5.4 提升 6 个百分点；生物信息学测试 BixBench 以 80.5%的成绩位列已公开模型榜首；陶哲轩参与策划的前沿数学测试 FrontierMathTier4 中，模型取得 35.4%的高分，远超 GPT-5.4 的 27.1%与 ClaudeOpus4.7 的 22.9%，在高难度科研任务中优势悬殊。实际应用中，GPT-5.5 可快速完成基因数据分析、代数几何应用构建等工作，将人类团队数月的工作量压缩至数小时。

图表 5：Opus4.7 与 GPT-5.5 一图对比

Evaluations

Coding

Eval	GPT-5.5	GPT-5.4	GPT-5.5 Pro	GPT-5.4 Pro	Claude Opus 4.7	Gemini 3.1 Pro
SWE-Bench Pro (Public) *	58.6%	57.7%	-	-	64.3%	54.2%
Terminal-Bench 2.0	82.7%	75.1%	-	-	69.4%	68.5%
Expert-SWE (Internal)	73.1%	68.5%	-	-	-	-

*Labs have noted evidence of memorization on this eval

Professional

Eval	GPT-5.5	GPT-5.4	GPT-5.5 Pro	GPT-5.4 Pro	Claude Opus 4.7	Gemini 3.1 Pro
GDPval (wins or ties)	84.9%	83.0%	82.3%	82.0%	80.3%	67.3%
FinanceAgent v1.1	60.0%	56.0%	-	61.5%	64.4%	59.7%
Investment Banking Modeling Tasks (Internal)	88.5%	87.3%	88.6%	83.6%	-	-
OfficeQA Pro	54.1%	53.2%	-	-	43.6%	18.1%

Computer use and vision

Eval	GPT-5.5	GPT-5.4	GPT-5.5 Pro	GPT-5.4 Pro	Claude Opus 4.7	Gemini 3.1 Pro
OSWorld-Verified	78.7%	75.0%	-	-	78.0%	-
MMMU Pro (no tools)	81.2%	81.2%	-	-	-	80.5%
MMMU Pro (with tools)	83.2%	82.1%	-	-	-	-

Tool use

Eval	GPT-5.5	GPT-5.4	GPT-5.5 Pro	GPT-5.4 Pro	Claude Opus 4.7	Gemini 3.1 Pro
BrowseComp	84.4%	82.7%	90.1%	89.3%	79.3%	85.9%
MCP Atlas**	75.3%	70.6%	-	-	79.1%	78.2%
Toolathlon	55.6%	54.6%	-	-	-	48.8%
Tau2-bench Telecom*** (original prompts)	98.0%	92.8%	-	-	-	-

资料来源：新智元，华鑫证券研究

此次 GPT-5.5 的发布，距 Anthropic 推出 ClaudeOpus4.7 仅 8 天，标志着全球 AI 大模型竞赛进入全新阶段。OpenAI 将 GPT-5.5 定位为全新计算机工作方式的核心载体，聚焦自主任务规划、多工具调用、跨平台协同的通用智能体能力，行业竞争核心已从单纯跑分转向 AIAgent 化办公的实际应用与生态定义。

2、 AI 应用动态：Kimi 周访问量环比+10.80%，DeepSeek V4 发布

2.1、 周流量跟踪：Kimi 周访问量环比+10.80%

本期（2026.4.17-2026.4.23）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1290.0M）、Bing（810.2M）和 Gemini（659.2M），访问量环比增速第一为 Kimi（10.80%）；平均停留时长前三位分别为 Character.AI（00:16:08）、Discord（00:11:02）和 Kimi（00:08:26）;平均停留时长环比增速第一为文心一言（3.40%）。

图表 6：2026.4.17-2026.4.23AI 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	-0.39%	5:51	0.29%
Bing	搜索	微软	810.2	-0.42%	7:36	-1.08%
Gemini	聊天机器人	谷歌	659.2	0.67%	7:12	-0.23%
Canva	在线设计	Canva	234.6	1.12%	5:53	-0.56%
Github	代码托管	微软	145.6	-1.09%	6:33	-0.25%
Discord	游戏社区	微软	138.4	0.14%	11:02	0.30%
Character.AI	聊天机器人	Character.AI	41.71	-3.34%	16:08	-2.81%
NotionAI	文本/笔记	Notion	40.55	-0.61%	8:03	0.00%
Perplexity	AI 搜索	Perplexity	36.13	-2.06%	4:48	-0.35%
DeepI	翻译工具	DeepL	27.76	-0.72%	2:28	0.00%
QuillBot	释义工具	QuillBot	11.47	0.44%	2:50	0.00%
Kimi	聊天机器人	Moonshot AI	10.75	10.80%	8:26	0.80%
文心一言	聊天机器人	百度	0.55	-1.72%	2:32	3.40%

资料来源：similarweb,华鑫证券研究

2.2、 产业动态：DeepSeek V4 发布，开源模型再攀性能与效率新高

2026 年 4 月 24 日，DeepSeek 正式发布了其最新一代模型 DeepSeekV4 系列。该系列模型的核心突破集中在百万级别上下文窗口的普及化，同时在智能体能力、世界知识掌握以及逻辑推理性能方面达到了开源社区的新高度。与此同步公开的还有详细的技术报告，供研究者和开发者查阅。

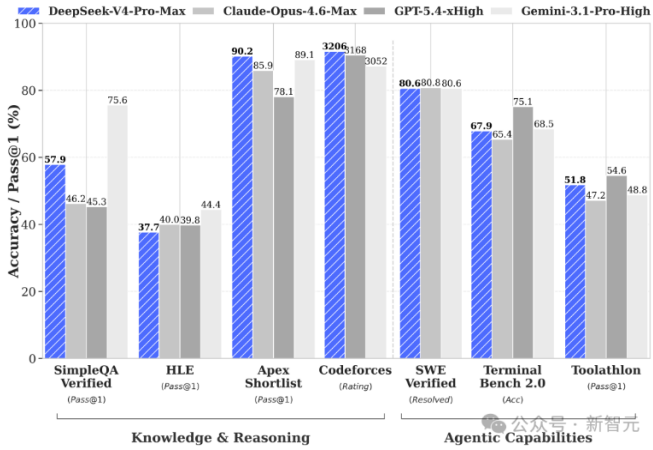
DeepSeekV4 系列包含两个不同定位的版本。其一是参数规模庞大的 DeepSeek-V4-Pro，总参数达到 1.6T，激活参数为 49B，性能上对标业内顶尖的闭源模型。其二是 DeepSeek-V4-Flash，总参数 284B，激活参数 13B，设计上更注重高效率和经济性。根据评测，Pro 版本在 Agent 相关任务上表现出色，尤其在编码方面，其体验被认为超越了 Sonnet4.5，交付质量也非常接近 Opus4.6（非思考模式），已成为公司内部智能体编程的首选。此外，该模型拥有深厚的世界知识储备，在知识测评上明显领先于其他开源产品，与 Gemini-Pro-3.1 间的差距已非常小。在逻辑推理方面，无论是数学、STEM 领域还是高难度的竞赛代码任务，Pro 版本不仅在开源社区中位居榜首，也具备了挑战全球最强闭源模型的实际能力。

图表 7：DeepSeek-V4-Pro 与 DeepSeek-V4-Flash

模型	参数	激活	预训练数据	上下文长度	开源	API 服务	网页端/APP 访问方式
deepseek-v4-pro	1.6T	49B	33T	1M	✓	✓	专家模式
deepseek-v4-flash	284B	13B	32T	1M	✓	✓	快速模式

资料来源：新智元，华鑫证券研究

图表 8：DeepSeek-V4-Pro 在 Agent 能力和知识储备方面的表现



资料来源：新智元，华鑫证券研究

图表 9：DeepSeek-V4-Pro 与其他模型在多项测试中的对比

Benchmark (metric) Reasoning Effort		DS-V4-Pro Max	DS-V4-Flash Max	K2.6 Thinking	GLM-5.1 Thinking	Opus-4.6 Max	GPT-5.4 xHigh	Gemini-3.1-Pro High
Knowledge & Reasoning	MMLU-Pro (EM)	87.5	86.2	87.1	86.0	89.1	87.5	91.0
	SimpleQA-Verified (Pass@1)	57.9	34.1	36.9	38.1	46.2	45.3	75.6
	Chinese-SimpleQA (Pass@1)	84.4	78.9	75.9	75.0	76.2	76.8	85.9
	GPQA Diamond (Pass@1)	90.1	88.1	90.5	86.2	91.3	93.0	94.3
	HLE (Pass@1)	37.7	34.8	36.4	34.7	40.0	39.8	44.4
	LiveCodeBench (Pass@1)	93.5	91.6	89.6	-	88.8	-	91.7
	Codeforces (Rating)	3206	3052	-	-	-	3168	3052
	HMMT 2026 Feb (Pass@1)	95.2	94.8	92.7	89.4	96.2	97.7	94.7
	IMOAnswerBench (Pass@1)	89.8	88.4	86.0	83.8	75.3	91.4	81.0
	Apex (Pass@1)	38.3	33.0	24.0	11.5	34.5	54.1	60.9
	Apex Shortlist (Pass@1)	90.2	85.7	75.5	72.4	85.9	78.1	89.1
Long Context	MRCR 1M (MMR)	83.5	78.7	-	-	92.9	-	76.3
	CorpusQA 1M (ACC)	62.0	60.5	-	-	71.7	-	53.8
Agentic	Terminal Bench 2.0 (Acc)	67.9	56.9	66.7	63.5	65.4	75.1	68.5
	SWE Verified (Resolved)	80.6	79.0	80.2	-	80.8	-	80.6
	SWE Pro (Resolved)	55.4	52.6	58.6	58.4	57.3	57.7	54.2
	SWE Multilingual (Resolved)	76.2	73.3	76.7	73.3	77.5	-	-
	BrowseComp (Pass@1)	83.4	73.2	83.2	79.3	83.7	82.7	85.9
	HLE w/tools (Pass@1)	48.2	45.1	54.0	50.4	53.1	52.0	51.6
	GDPval-AA (Elo)	1554	1395	1482	1535	1619	1674	1314
	MCPAtlas Public (Pass@1)	73.6	69.0	66.6	71.8	73.8	67.2	69.2
	Toolathlon (Pass@1)	51.8	47.8	50.0	40.7	47.2	54.6	48.8

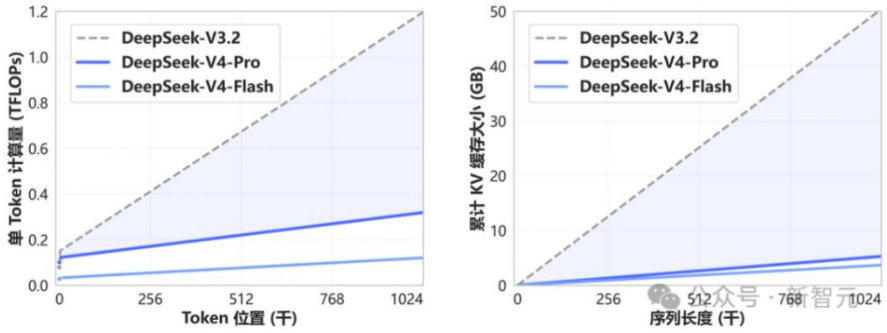
资料来源：新智元，华鑫证券研究

支撑这些性能提升的是三项底层技术创新。第一项是混合注意力机制，包含压缩稀疏注意力和重压缩注意力两种方式。前者对 KV 缓存进行基于 token 维度的压缩并结合 DSA 稀疏注意力，后者则采用更极致的压缩策略以维持稠密计算。两者结合形成了长短兼顾的策略，大幅降低了模型在处理百万级别上下文时的计算量和显存需求。第二项创新是流形约束超连接结构，它改进了传统的残差连接方式，能够增强信号传播的稳定性并提升模型的表达能力，使得深层网络依然保持强大的建模能力。第三项是引入了 Muon 优化器，它让训练过程的收敛速度更快且更加稳定。由于这些架构上的革新，DeepSeek-V4 在百万 token 这种极端长上下文场景下，Pro 版本单 token 的推理计算量仅为前一版本的 27%，KV 缓存的占用更是缩减了 10%。

相比之下，Flash 版本在高效率与性价比之间实现了更好的平衡。虽然它在世界知识的深度上略逊于 Pro 版本，但在逻辑推理水平上与之非常接近。凭借更小的参数规模和激活机制，Flash 版本能够提供响应速度更快、成本更低的 API 接入方式。在处理基础智能体任务时，Flash 版本与 Pro 版本表现相近，只有在面对极端复杂的任务时才会体现出差距。

在长上下文处理能力上，DeepSeek-V4 通过革命性的注意力机制设计，在 token 维度进行高效压缩，并配合 DSA 稀疏注意力技术，实现了全球领先的长文本处理性能。这种创新显著降低了对计算资源和显存的依赖。从发布之日起，一百万 token 的超长上下文已经成为 DeepSeek 官方服务的标准配置。

图表 10：DeepSeek-V4 和 DeepSeek-V3.2 的计算量和显存容量随上下文长度的变化



资料来源：新智元，华鑫证券研究

模型还针对 Claude Code、OpenClaw、OpenCode、CodeBuddy 等主流智能体开发生态做了深度适配。在代码编写和自动生成文档等实际应用场景中，DeepSeek-V4 的生产效率获得了显著提升。

对于开发者而言，API 已经同步上线，只需要简单修改 model_name 即可接入两款新模型。追求性能的可以使用 deepseek-v4-pro，追求效率的则使用 deepseek-v4-flash。需要特别注意的是，原有的 deepseek-chat 和 deepseek-reasoner 这两个模型名将作为 V4 系列的过渡别名，分别指向 Flash 版本的非思考模式和思考模式，但这些旧名称会在 2026 年 7 月 24 日正式停用。

3、AI 融资动向：Cognition 拟融资数亿美元估值目标 250 亿美元

2026 年 4 月，AI 编程智能体公司 CognitionAI 正洽谈新一轮数亿美元巨额融资，计划将估值提升至 250 亿美元，融资谈判仍在进行中。公司 2023 年 8 月成立，凭借自主 AI 软件工程师 Devin 快速崛起，截至 2025 年 6 月 ARR 已达 7300 万美元，收购 Windsurf 后增长进一步提速，本轮融资将全力投入 Devin 的技术迭代与商业化拓展。

Cognition 核心产品 Devin 被誉为全球首个完全自主 AI 软件工程师，可自主完成任务规划、代码编写、调试、测试与全流程部署。执行任务前生成详细蓝图并给出置信度评分，通过多重检测降低错误率，兼顾代码质量与安全性。产品面向个人与企业提供分级服务，企业版支持定制化、增强安全与审计能力，已获戴尔、思科等科技巨头采用。

从 2024 年 3 月 A 轮估值 3.5 亿美元，到如今冲刺 250 亿美元，Cognition 在 AI 编程赛道实现估值跨越式增长。此次巨额融资计划，凸显资本市场对自主 AI 软件工程范式的高度认可，也将进一步巩固 Devin 在智能编程领域的领先地位，推动软件开发向全流程自主化加速演进。

图表 11：上周 AI 初创公司融资动态

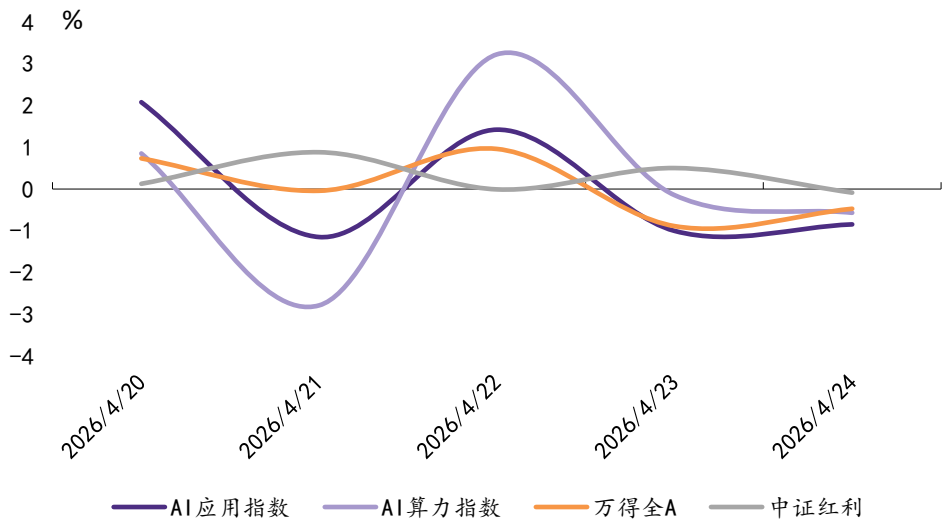
应用	应用类型	领投方	融资轮	融资额	目前累计融资额	目前估值
Cognition	AI 开发工具	FoundersFund	后续轮	5 亿美元	6.96 亿美元	102 亿美元
AfterQuery	AI 训练数据	AltosVentures	A 轮	3000 万美元	3050 万美元	3 亿美元
Factory	AI 开发平台	KhoslaVentures	C 轮	1.5 亿美元	2.2 亿美元	15 亿美元

资料来源：wind，Saasverse，华鑫证券研究

4、 行情复盘

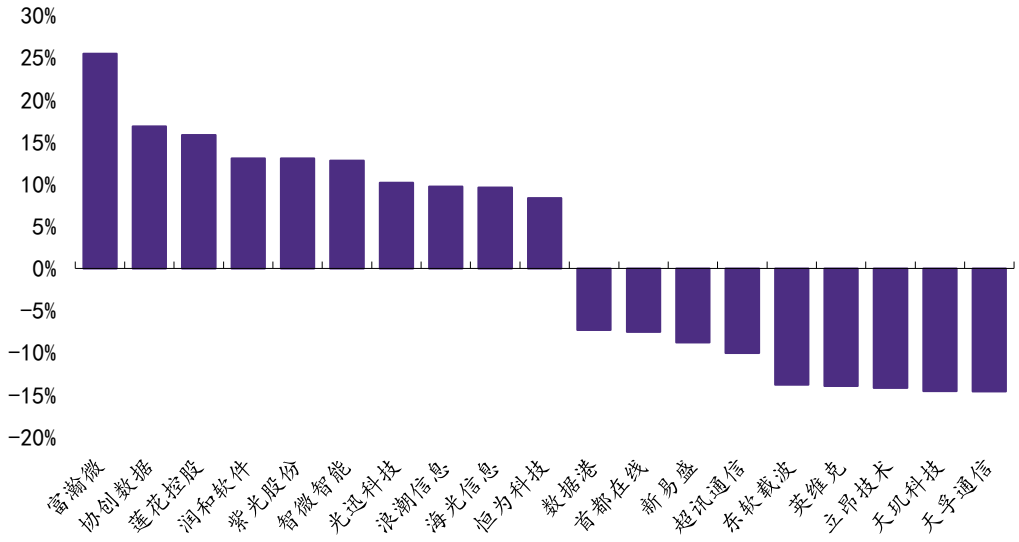
上周（2026.4.20-2026.4.24 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 2.08%/3.23%/0.96%/0.88%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-1.15%/-2.79%/-0.89%/-0.09%。AI 算力指数内部，富瀚微以 25.49%录得上周最大涨幅，天孚通信以-14.61%录得上周最大跌幅。AI 应用指数内部，富瀚微以 25.49%录得上周最大涨幅，真视通以-18.2%录得上周最大跌幅。

图表 12：上周（2026.4.20-2026.4.24）指数日涨跌幅



资料来源：wind，华鑫证券研究

图表 13：上周（2026.4.20-2026.4.24 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（2026.4.20-2026.4.24 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 4 月 24 日，DeepSeek 正式推出 V4 模型预览版并同步开源。本次版本迭代核心集中于超长上下文能力升级、推理架构优化及国产化算力适配，整体运行效率与成本控制能力实现显著提升。该模型同步推出 pro 与 flash 两大版本；其中 pro 总参数 1.6T、激活参数 49B、预训练数据 33T，flash 总参数 284B、激活参数 13B、预训练数据 32T，两款模型统一搭载 100 万词元超长上下文窗口，综合性能对标行业顶尖水平，能够以更低的内存资源消耗，稳定处理大篇幅长文本处理任务。V4 在底层架构层面完成全面升级，融合压缩稀疏注意力（CSA）与重度压缩注意力（HCA）混合架构，搭配 mHC 训练稳定机制及 Muon 主训练优化器，部分模块沿用 AdamW 架构协同运作，大幅优化长上下文场景下的推理运行效率。对比前代产品，v4-pro 在百万级上下文场景中，单 token 推理浮点运算量缩减至原有 27%，KV 缓存占用仅为原先 10%；flash 优化效果更为突出，单词元推理浮点运算量下降至前代的 10%，KV 缓存占用压缩至 7%，整体运行成本得到有效控制。与此同时，deepseek 进一步推进国产化算力适配布局，深度对接华为昇腾 950 超级节点量产规划，随着 2026 年下半年该芯片产品实现大规模供货落地，v4-pro 的 API 调用定价有望迎来明显下调，进一步提升商业化落地性价比。

此次发布的 V4 深度适配华为昇腾体系，通过底层架构重构优化推理调度与缓存占用，整体算力运行效率大幅改善。当前高端算力供给紧缺，持续约束大模型服务吞吐能力。伴随下半年昇腾 950 的批量交付，高端大模型商业化面临的算力瓶颈与成本压力将有效缓解。头部模型厂商加速推进国产芯片兼容适配，持续完善算力自主可控生态，芯模协同的共振下，国产算力产业链有望进入放量周期。

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 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 13.9 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	2026-04-27	EPS			PE			投资评级
		股价	2024	2025E	2026E	2024	2025E	2026E	
300757.SZ	罗博特科	518.50	0.41	-0.30	0.30	1264.63	-1728.33	1728.33	买入
301196.SZ	唯科科技	117.37	1.76	2.53	3.34	66.69	46.39	35.14	买入
603859.SH	能科科技	39.49	0.78	0.96	1.18	50.63	41.14	33.47	买入
688615.SH	合合信息	128.00	4.01	3.24	4.22	50.54	39.51	30.33	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。