DeepSeek V4 发布点评

百万上下文进入普惠时代，国产算力成功适配需求爆发将至

优于大市

◆ 行业研究·行业快评 ◆ 互联网 ◆ 投资评级：优于大市（维持）

证券分析师： 张伦可 0755-81982651 zhanglunke@guosen.com.cn 执证编码：S0980521120004
证券分析师： 陈淑媛 021-60375431 chenshuyuan@guosen.com.cn 执证编码：S0980524030003
证券分析师： 张昊晨 zhanghaochen1@guosen.com.cn 执证编码：S0980525010001

事项：

事项：4月24日，DeepseekV4发布。根据DeepSeek的官方介绍，V4系列包含两个MoE模型：DeepSeek-V4-Pro总参数1.6T、激活参数49B，DeepSeek-V4-Flash总参数 284B、激活参数13B，两者均原生支持100万token上下文。

国信互联网观点：

Agent能力大幅提升，处于国内第一梯队，性能比肩全球顶级闭源模型。根据公司官方，在 Agentic Coding评测中，V4 Pro 已达到当前开源模型最佳水平。目前 DeepSeek V4 已成为公司内部员工使用的 Agentic Coding 模型，据评测反馈使用体验优于 Sonnet 4.5，交付质量接近 Opus 4.6 非思考模式，但仍与 Opus 4.6 思考模式存在一定差距。从第三方评测来看，Arena.ai 在 X 上将 V4 Pro 定性为"相较 DeepSeek V3.2 的重大飞跃"，在代码开源模型榜单中位列第 3 位、综合第 14 位。另一家测评方 Vals AI 表示，V4 在其 Vibe Code Benchmark 中是开源模型榜首。

Deepseek通过注意力层改进，推动百万上下文进入普惠时代。DeepSeek V4 开创了一种全新的注意力机制，在 token 维度进行压缩，结合 DSA 稀疏注意力，实现了全球领先的长上下文能力，并且相比于传统方法大幅降低了对计算和显存的需求。在 1M 上下文设置下，DeepSeek V4 Pro 的单 token 推理 FLOPs 只有 V3.2 的 27%，KV Cache 只有 10%；V4-Flash 更极端，分别压到 10%和 7%。价格方面，Deepseek 实现了高性价比。DeepSeek-V4-Pro 输入/输出百万 Token 价格 12 元/24 元；对比国内 1T 模型，比如小米输入/输出百万 Token 价格$1/ $3 （小于 256K 上下文）；$2/ $6（1M 上下文）。Deepseek-V4-Flash 价格更低，为输入/输出百万 Token 价格 1 元/2 元。

Deepseek与国产芯片进行适配，包括华为、寒武纪等。V4 在技术报中表示，在英伟达 GPU 和华为昇腾 NPU 两个平台上均验证了细粒度 EP（专家并行）方案。根据官网，下半年昇腾 950 超节点批量上市有望继续推动 V4 Pro 降价。根据华为云官方，昇腾一直同步支持 DeepSeek 系列模型，本次通过双方芯模技术紧密协同，实现昇腾超节点全系列产品支持 DeepSeek V4 系列模型。昇腾 950 通过融合 kernel 和多流并行技术降低 Attention 计算和访存开销，大幅提升推理性能，结合多种量化算法，实现了高吞吐、低时延的 DeepSeek V4 模型推理部署。昇腾 A3 超节点系列产品也全面适配，同时为便于用户快速微调，提供了基于昇腾 A3 超节点的训练参考实现。寒武纪 Day 0 适配 DeepSeek-V4。寒武纪已基于 vLLM 推理框架完成 285B DeepSeek-V4-flash 和 1.6T DeepSeek-V4-pro 的 Day 0 适配，适配代码已开源到 GitHub 社区。这一成果得益于寒武纪长期积累的自研 NeuWare 软件生态与芯片设计技术，也是寒武纪对芯片与算法联合创新持续投入的延续。

投资建议：密切关注国产模型和国产芯片。Deepseek 的进一步在架构设计层面“降本增效”，推动国产模型更普惠实现百万上下文长度，密切关注国产模型厂商进展。寒武纪、华为昇腾的 Day 0 适配表明，表明国产芯片在已达到商业可用的成熟度，密切关注国产算力进展。

风险提示：宏观经济波动风险、下游需求不及预期风险、核心技术水平升级不及预期的风险、AI 快速迭代平权化下竞争加剧等。

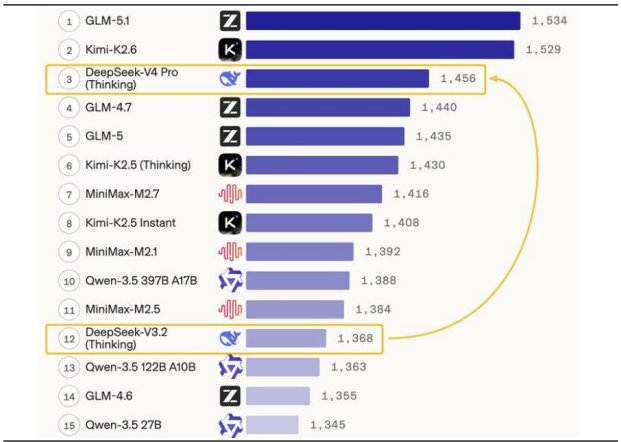
评论：

◆ Agent 能力大幅提升，处于国内第一梯队，性能比肩全球顶级闭源模型

Agent 能力大幅提高，交付质量接近 Opus 4.6 非思考模式。根据官网，在 Agentic Coding 评测中，V4-Pro 已达到当前开源模型最佳水平，并在其他 Agent 相关评测中同样表现优异。目前 DeepSeek-V4 已成为公司内部员工使用的 Agentic Coding 模型，据评测反馈使用体验优于 Sonnet 4.5，交付质量接近 Opus 4.6 非思考模式，但仍与 Opus 4.6 思考模式存在一定差距。

模型能力处于国内第一梯队。从第三方评测来看，Arena.ai 在 X 上将 V4 Pro（思考模式）定性为"相较 DeepSeek V3.2 的重大飞跃"，在代码开源模型榜单中位列第 3 位、综合第 14 位。另一家测评方 Vals AI 表示，V4 在其 Vibe Code Benchmark 中是开源模型榜首。

图1：Arena Code 开源模型榜单

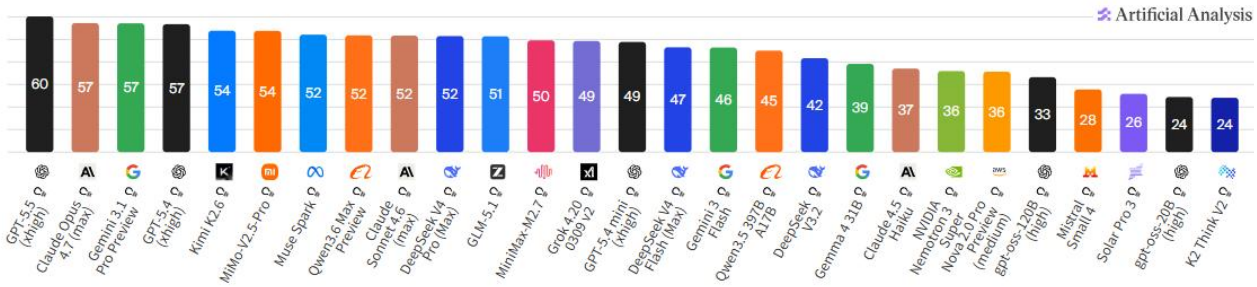


资料来源：Arena AI，国信证券经济研究所整理

图2：Vals AI 的 Vibe Code 开源模型榜单

资料来源：Vals AI，国信证券经济研究所整理

图3：全球模型智能水平排序



资料来源：Artificial Analysis，国信证券经济研究所整理

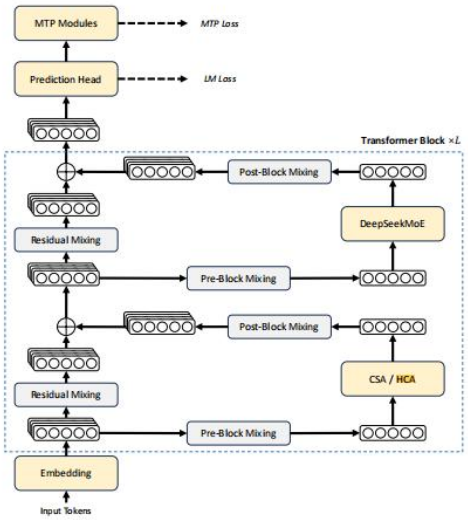
◆ Deepseek 通过注意力层改进，推动百万上下文进入普惠时代

DeepSeek V4 通过结构创新，实现了超高上下文效率。DeepSeek V4 开创了一种全新的注意力机制，在 token 维度进行压缩，结合 DSA 稀疏注意力（DeepSeek Sparse Attention），实现了全球领先的长上下文能力，并且相

比于传统方法大幅降低了对计算和显存的需求。根据腾讯科技，V4 的做法是把注意力拆成两种，交替叠用。

- ✓ CSA（压缩稀疏注意力）：把每 4 个 token 的 KV 缓存合并成一条摘要，再让每个 query 只在这些摘要里挑出最相关的 top-k 条去算注意力。相当于既压缩了“要看的内容”，又只挑“值得看的”去算。
- ✓ 另一种叫 HCA（重压缩注意力）：压缩率更激进，把每 128 个 token 合并成一条，但对剩下的摘要做稠密注意力，不做稀疏挑选。

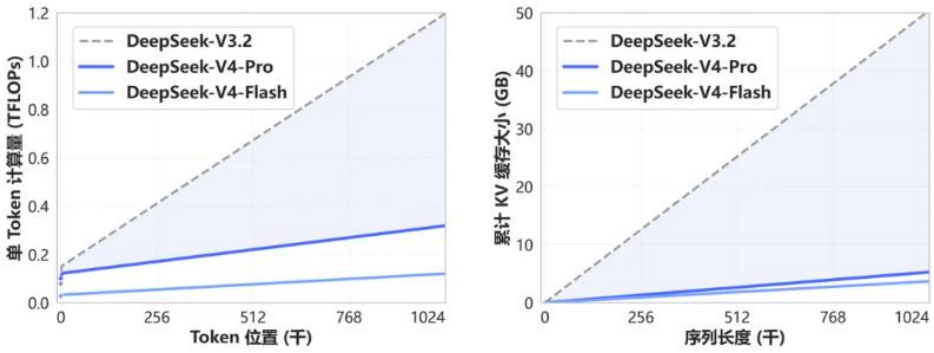
图4：Deepseek 模型架构设计图



资料来源：Deepseek 论文，国信证券经济研究所整理

在 1M 上下文设置下，DeepSeek V4-Pro 的单 token 推理 FLOPs 只有 V3.2 的 27%，KV Cache 只有 10%；V4-Flash 更极端，分别压到 10%和 7%。换句话说，上下文从 V3.2 的 128K 扩到 V4 的 1M，理论上放大了近 8 倍，但单 token 算力需求反而下降。

图5：DeepSeek V4 计算量和显存容量随上下文长度的变化



资料来源：Deepseek，国信证券经济研究所整理

价格方面，Deepseek 实现了高性价比。DeepSeek-V4-Pro 输入/输出百万 Token 价格 12 元/24 元；对比国内 1T 模型，比如小米输入/输出百万 Token 价格$1/ $3 （小于 256K 上下文）；$2/ $6（1M 上下文范围）。Deepseek-V4-Flash 价格更低，为输入/输出百万 Token 价格 1 元/2 元。

图6：全球模型 Token 价格比较

Model	Input ($/M)	Output ($/M)
DeepSeek V4 Flash	$0.14	$0.28
GPT-5.4 Nano	$0.20	$1.25
Gemini 3.1 Flash-Lite	$0.25	$1.50
Gemini 3 Flash Preview	$0.50	$3
GPT-5.4 Mini	$0.75	$4.50
Claude Haiku 4.5	$1	$5
DeepSeek V4 Pro	$1.74	$3.48
Gemini 3.1 Pro	$2	$12
GPT-5.4	$2.50	$15
Claude Sonnet 4.6	$3	$15
Claude Opus 4.7	$5	$25
GPT-5.5	$5	$30

资料来源：Simon Willison，36 氪，国信证券经济研究所整理

图7：Deepseek 的 Token 定价

API 访问模型名	输入（缓存命中）	输入（缓存未命中）	输出	上下文长度
deepseek-v4-pro	1 元	12 元	24 元	1M
deepseek-v4-flash	0.2 元	1 元	2 元	

资料来源：Deepseek，国信证券经济研究所整理

◆ Deepseek 与国产芯片进行适配，包括华为、寒武纪等

Deepseek V4 将华为昇腾写入硬件清单，下半年昇腾 950 超节点批量上市有望继续推动 V4 Pro 降价。V4 在技术报中表示，在英伟达 GPU 和华为昇腾 NPU 两个平台上均验证了细粒度 EP（专家并行）方案。这是 DeepSeek 官方第一次在正式文档中把华为昇腾和英伟达并列写进硬件验证清单。根据华为云官方，昇腾一直同步支持 DeepSeek 系列模型，本次通过双方芯模技术紧密协同，实现昇腾超节点全系列产品支持 DeepSeek V4 系列模型。昇腾 950 通过融合 kernel 和多流并行技术降低 Attention 计算和访存开销，大幅提升推理性能，结合多种量化算法，实现了高吞吐、低时延的 DeepSeek V4 模型推理部署。昇腾 A3 超节点系列产品也全面适配，同时为便于用户快速微调，提供了基于昇腾 A3 超节点的训练参考实现。

寒武纪 Day 0 适配 DeepSeek-V4。寒武纪已基于 vLLM 推理框架完成对 285B DeepSeek-V4-flash 和 1.6T DeepSeek-V4-pro 的 Day 0 适配，适配代码已开源到 GitHub 社区。这一成果得益于寒武纪长期积累的自研 NeuWare 软件生态与芯片设计技术，也是寒武纪对芯片与算法联合创新持续投入的延续。

◆ 投资建议

密切关注国产模型和国产芯片。Deepseek 进一步在架构设计层面“降本增效”，推动国产模型更普惠实现百万上下文长度，密切关注国产模型厂商进展等。寒武纪、华为昇腾的 Day 0 适配表明，表明国产芯片在已达到商业可用的成熟度，密切关注国产算力进展。

◆ 风险提示

宏观经济波动风险、下游需求不及预期风险、核心技术水平升级不及预期的风险、AI 快速迭代平权化下竞争加剧等。

免责声明

分析师声明

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价(或行业指数)相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

重要声明

关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券经济研究所

深圳
深圳市福田区福华一路 125 号国信金融大厦 36 层
邮编：518046　总机：0755-82130833

上海
上海浦东民生路 1199 弄证大五道口广场 1 号楼 12 层
邮编：200135

北京
北京西城区金融大街兴盛街 6 号国信证券 9 层
邮编：100032