计算机行业深度报告

国产化训练从 0 到 1 里程碑，战略意义大于性能意义

增持（维持）

2026 年 04 月 30 日

证券分析师 王紫敬
执业证书：S0600521080005
021-60199781
wangzj@dwzq.com.cn

投资要点

- **重要意义：**国产开源大模型在国产算力训练适配领域以及百万级上下文能力实现了里程碑式突破。
- **DeepSeek V4 首次由华为昇腾芯片参与训练。**DeepSeek V4 Flash 是首个公开说明训练侧使用国产算力的通用大模型，通过三大核心设计实现了去英伟达化的技术布局。**（1）引入 MXFP4 量化感知训练**，对 MoE 专家权重与索引器 QK 路径实现 FP4 量化，降低了对 NVIDIA FP8 生态的绑定，可无缝适配华为昇腾、寒武纪等国产芯片；**（2）采用 TileLang 领域专用语言开发底层算子，**脱离 CUDA 生态强绑定，可跨硬件平台编译，降低向国产芯片的迁移成本；**（3）自研 MegaMoE2 融合内核，**实现专家并行的细粒度通信计算重叠，已在华为昇腾平台完成适配跑通，解决了国产硬件环境下 MoE 模型的通信瓶颈。
- **性能表现：整体跻身全球第一梯队，多项核心指标比肩甚至超越国际顶级闭源模型。（1）知识储备：**DeepSeek-V4-Pro-Max 在 SimpleQA-Verified 基准上取得 57.9 分，大幅领先其他主流开源模型；中文 SimpleQA 得分达 84.4，大幅缩小与 Gemini-3.1-Pro 的差距，MMLU-Pro、GPQA Diamond 等教育知识基准均领跑开源赛道。**（2）推理与代码能力：**Pro-Max 版本 Codeforces 评分达 3206，位列人类选手排行榜第 23 名，LiveCodeBench Pass@1 达 93.5，IMOAnswerBench 得分 89.8 仅略逊于 GPT-5.4；Flash 版本 Codeforces 评分也达到 3052，推理性能追平 GPT-5.2 等闭源模型。**（3）Agent 能力：**V4 Pro-Max 的 SWE-bench Verified 任务解决分数达 80.6，与 Claude Opus 4.6 基本持平，Terminal Bench 2.0、MCPAtlas Public 等基准均处于开源模型第一梯队。**（4）长上下文能力：**1M token 场景下，MRCR、CorpusQA 得分分别为 83.5、62.0，超越 Gemini-3.1-Pro，且 128K 上下文内检索能力保持高度稳定。**（5）中文创作：**其功能性写作对 Gemini-3.1-Pro 胜率达 62.7%，创意写作质量胜率高达 77.5%，仅在高难度多轮约束场景略逊于 Claude Opus 4.5。
- **模型技术架构：CSA+HCA+mHC 进一步压缩推理成本。（1）首创 CSA+HCA 交替的混合注意力架构。**通过分层 KV 缓存压缩与稀疏注意力结合，在 1M token 上下文场景下，Pro 版本单 token 推理 FLOPs 仅为 V3.2 的 27%，KV 缓存占用降至 10%，Flash 版本更是分别降至 10%与 7%，从底层解决了超长上下文的算力瓶颈；**（2）引入 mHC 流形约束超连接升级传统残差结构，**提升了深层模型的信号传播稳定性与表达能力，同时采用 Muon 优化器搭配预期性路由、SwiGLU 钳制技术，解决了万亿参数 MoE 模型训练的 Loss Spike 难题；**（3）采用领域专家独立训练+全词表在线蒸馏的后训练范式，**规避了多能力融合的性能退化问题。
- **投资建议：**DeepSeek V4 是大模型在训练侧使用国产算力从 0 到 1 的尝试。此前国产大模型采用国产算力均用于推理侧，而 DeepSeek 本次从模型内核到训练架构、到推理全流程均出现了国产算力的影子，是重要里程碑。因此，无论 DeepSeek V4 表现如何，战略意义均十分重要，对国产算力的训练适配前景才是关注的重点。**国产算力相关标的：禾盛新材、寒武纪-U、海光信息、中科曙光、摩尔线程-U、沐曦股份-U、华丰科技、航天电器等**，详见正文 P17【投资建议】。
- **风险提示：**大模型迭代节奏不及预期；国产算力软硬件生态适配进度不及预期；大模型行业市场竞争持续加剧；行业政策监管持续趋严

行业走势

相关研究

《Agent 时代 CPU 迎来重新定位，国产 CPU 有望量价齐升》

2026-04-26

《Token 时代下算力租赁行业重构》

2026-04-14

内容目录

图表目录

1. 与市场不同的观点：国产算力适配的探路者，

资本市场对 DeepSeek 的定位存在一定认知偏差，其核心定位已从挑战闭源模型，转向扛起开源龙头技术普惠与国产算力生态适配的使命。自 DeepSeek V3.13.2 版本起，其研发重心便从单一模型能力追赶，转向国产算力生态的底层适配与技术开源。模型训练端率先完成国产芯片的导入与适配，系统性解决国产芯片在大模型落地中的核心痛点，同时将核心技术全面开源，大幅降低行业对国产芯片的应用门槛与试错成本，从底层推动国产算力全链条的生态成熟与规模化应用，为国产算力产业创造长期市场需求。

在全球科技竞争加剧、潜在技术脱钩与高端算力出口管制持续升级的背景下，DeepSeek 与国内头部科技企业共同承担着筑牢中国 AI 大模型软硬件自主可控底线的核心使命。相较于短期性能与性价比的市场争议，其核心价值在于重塑国内 AI 产业对海外技术与算力的依赖格局，解决了 AI 产业自主发展“有与没有”的核心安全问题。

DeepSeek V4 的产业价值与长期布局，契合全球 AI 竞争的核心逻辑——生态主导权的争夺远重于短期性能差距。DeepSeek V4 预览版不仅实现了对海外主流闭源大模型的高性价比平替，**在保持全球顶级性能梯队的同时，将应用成本降至闭源模型的零头，既加速了国产算力的规模化落地，更验证了国产算力体系从训推全链路支撑顶级 AI 生产力的可行性。**美国白宫《Winning the AI Race: America’s AI Action Plan》明确提出，AI 生态系统规模将直接影响全球 AI 标准制定权，并决定相关国家能否获得广泛的经济与军事利益。性能差距可通过技术迭代逐步弥补，但 AI 生态的卡位与构建已刻不容缓，这正是 DeepSeek 长期布局的核心逻辑，也是当前资本市场普遍忽视的中长期价值锚点。

图1：DeepSeek V4 Flash 由昇腾参与训练，Pro 正在进行国产算力训练适配

数据来源：大国 AI 论坛，东吴证券研究所

2. DeepSeek V4：首个实现国产算力训练适配的顶级通用大模型

长期以来，全球顶级通用大模型的训练环节高度依赖 NVIDIA GPU 与 CUDA 生态，国产算力在大模型领域的应用大多局限于推理侧，难以进入核心的训练环节。核心瓶颈在于模型底层架构、算子开发、量化体系均深度绑定 NVIDIA 技术栈，向国产芯片迁移的成本极高、性能损耗显著。而 DeepSeek V4 Flash 版本首次实现了通用大模型训练侧的国产算力适配，通过三大核心技术设计，开始由纯适配英伟达的训练框架转向国产算力，为国产大模型的国产化训练奠定了技术基础。

2.1. MXFP4 量化感知训练：打破 NVIDIA 浮点生态强绑定

2.1.1. 什么是 MXFP4

MXFP4（Microscaling Data Formats 4-bit）是由行业厂商联合推出的微缩放 4 位浮点数格式，核心是通过细粒度的块级缩放因子，实现低精度的数值表示，在大幅降低显存占用与计算量的同时，最小化精度损失。**相较于传统的 INT4 量化，MXFP4 浮点数格式在动态范围、精度保留上具备天然优势，更适配大模型 Transformer 结构的数值分布特征。**

2.1.2. MXFP4 在 DeepSeek V4 中的具体应用环节

根据 DeepSeek V4 技术文档，模型在训练与推理全流程中，将 MXFP4 量化应用于两大核心组件，实现了显存与计算量的双重优化：

MoE 专家权重量化：MoE 结构的专家权重是大模型显存占用的核心来源，DeepSeek V4 对所有路由专家参数采用 MXFP4 量化，在训练阶段通过量化感知训练（QAT）让模型适配精度损失，推理阶段直接使用原生 FP4 权重，大幅降低显存占用与内存加载开销。

CSA 索引器 QK 路径全链路量化：对压缩稀疏注意力（CSA）中闪电索引器的 Query-Key（QK）路径，实现了从激活值缓存、加载到矩阵乘法的全流程 FP4 精度执行，大幅加速了超长上下文下的注意力分数计算。

辅助量化优化：在 QAT 过程中，将索引器输出的索引分数从 FP32 量化至 BF16，使 top-k 选择器的执行速度提升 2 倍，同时 KV 条目的召回率仍保持 99.7%，几乎无精度损失。

2.1.3. MXFP4 助力国产化适配的核心逻辑

MXFP4 量化体系的设计，从底层动摇了大模型对 NVIDIA FP8 生态的强绑定，是

实现国产化适配的核心基石：

无损反量化实现跨平台兼容：DeepSeek V4 的 MXFP4 量化实现了 FP4 到 FP8 的无损反量化——FP8（E4M3）格式相比 FP4（E2M1）多 2 个指数位，具备更大的动态范围，只要 FP4 子块的缩放因子比值不超过阈值，细粒度缩放信息就能被 FP8 的动态范围完全吸收。这意味着整个 QAT 管线可以完全复用现有的 FP8 训练框架，无需修改反向传播管线。**国产 AI 芯片厂商正在加速补齐 FP8、MXFP8、MXFP4 等低精度计算能力。**华为新一代昇腾 950 系列已明确支持 FP8、MXFP8、MXFP4 等格式，寒武纪公开资料也披露其大模型训练方案具备原生 FP8 计算能力，为未来国产大模型适配打下基础。

为国产硬件预留优化空间：技术文档明确指出，尽管当前硬件上 FP4 × FP8 运算的峰值 FLOPs 与 FP8 × FP8 持平，**但在未来硬件架构上，FP4 × FP8 理论上可实现 1/3 的效率提升。**这一设计为国产芯片厂商提供了全新的技术迭代方向，国产厂商可基于 MXFP4 格式设计专属的计算加速单元，无需跟随 NVIDIA 的技术路线被动迭代。

降低国产硬件的适配门槛：MXFP4 量化大幅降低了模型对显存带宽、计算单元的要求，使显存与算力规格相对有限的国产芯片，也能承载万亿参数 MoE 模型的训练与推理，大幅拓宽了国产算力的应用边界。

2.1.4. MXFP4 当前存在的短板

从技术落地的实际情况来看，MXFP4 量化仍存在一定的局限性：

量化覆盖范围有限：当前 MXFP4 仅应用于 MoE 专家权重与 CSA 索引器 QK 路径，对于模型嵌入层、注意力核心计算、残差连接等核心模块，仍需采用 BF16/FP8 精度，无法实现全模型 FP4 量化，显存优化的范围存在局限。

权重分布约束性强：FP4 到 FP8 的无损反量化，依赖于权重缩放因子的比值满足阈值要求，对于极端分布的长尾权重参数，仍可能出现精度损失，进而影响模型在低概率知识、高难度推理任务上的表现。

硬件生态成熟度不足：包括国产芯片在内的当前主流硬件，对 FP4 × FP8 融合计算的硬件加速支持仍不完善，现阶段 MXFP4 的理论效率优势无法完全发挥，只能通过量化感知训练模拟实现，实际部署的加速效果受限于硬件生态的成熟度。

2.2. TileLang 领域专用语言：脱离 CUDA 生态的底层算子底座

2.2.1. 什么是 TileLang

TileLang 是由北大团队开发的一款面向现代神经网络内核开发的领域专用语言（DSL），核心价值是在张量计算的底层开发中，平衡开发效率与运行时性能。传统大

模型底层算子开发依赖 CUDA，开发者需要针对不同硬件平台重写算子，开发成本极高；而 TileLang 实现了开发中可实现跨平台编译优化，让开发者在同一代码库中即可完成算子的快速原型开发与深度性能优化，无需针对不同硬件架构重复开发。

2.2.2. DeepSeek 采用 TileLang 的核心战略意义

TileLang 的应用，是 DeepSeek V4 实现国产化适配的核心底层支撑，其最大意义在于彻底脱离了对 NVIDIA CUDA 生态的强绑定。具体体现在三大维度：

跨平台迁移成本大幅降低：传统大模型的底层算子几乎全部基于 CUDA 开发，国产芯片想要适配，需要逐一对标 CUDA 算子进行重写和优化，不仅迁移周期长，性能往往无法达到原生水平。而 TileLang 作为跨平台 DSL，基于其开发的融合内核，可在 NVIDIA GPU、华为昇腾 NPU 等不同硬件平台上直接编译优化，无需重写核心代码，让 DeepSeek V4 向国产算力平台的迁移效率实现了量级提升。

开发效率与运行性能的完美平衡：DeepSeek V4 的创新架构原本会产生数百个细粒度的 Torch ATen 算子，调用开销大、性能损耗显著。通过 TileLang 开发融合内核，替代了绝大多数原生算子，实现了最优的运行性能。同时，TileLang 的 Host Codegen 技术，将主机端的运行时检查、参数校验等逻辑从 Python 侧移至生成的主机代码中，把 CPU 端的校验开销从数十到数百微秒降至 1 微秒以内，解决了国产硬件平台上 Python 调度开销大的核心痛点。

全流程数值一致性保障：TileLang 默认关闭 fast-math 优化，提供符合 IEEE-754 标准的数值内联函数，同时对齐 CUDA 工具链的代数简化与降级规则，可实现与手写 CUDA 内核位级一致的输出。这一特性保证了 DeepSeek V4 在预训练、后训练、推理全流程中，跨不同硬件平台的数值一致性，解决了“国产芯片训练、英伟达推理”的跨平台部署数值漂移难题，实现了训练与推理的硬件解耦。

架构创新的快速迭代支撑：TileLang 集成了 Z3 SMT 求解器，实现了形式化整数分析，可处理张量索引的复杂算术运算，解锁向量化、屏障插入、代码简化等高级优化。这让 DeepSeek 团队可以快速迭代 CSA、HCA、mHC 等创新架构的算子原型，同时保证在国产硬件上的性能表现，为国产算力平台上的大模型架构创新提供了底层工具支撑。

2.3. 自研 MegaMoE2 融合内核：解决国产硬件 MoE 通信瓶颈

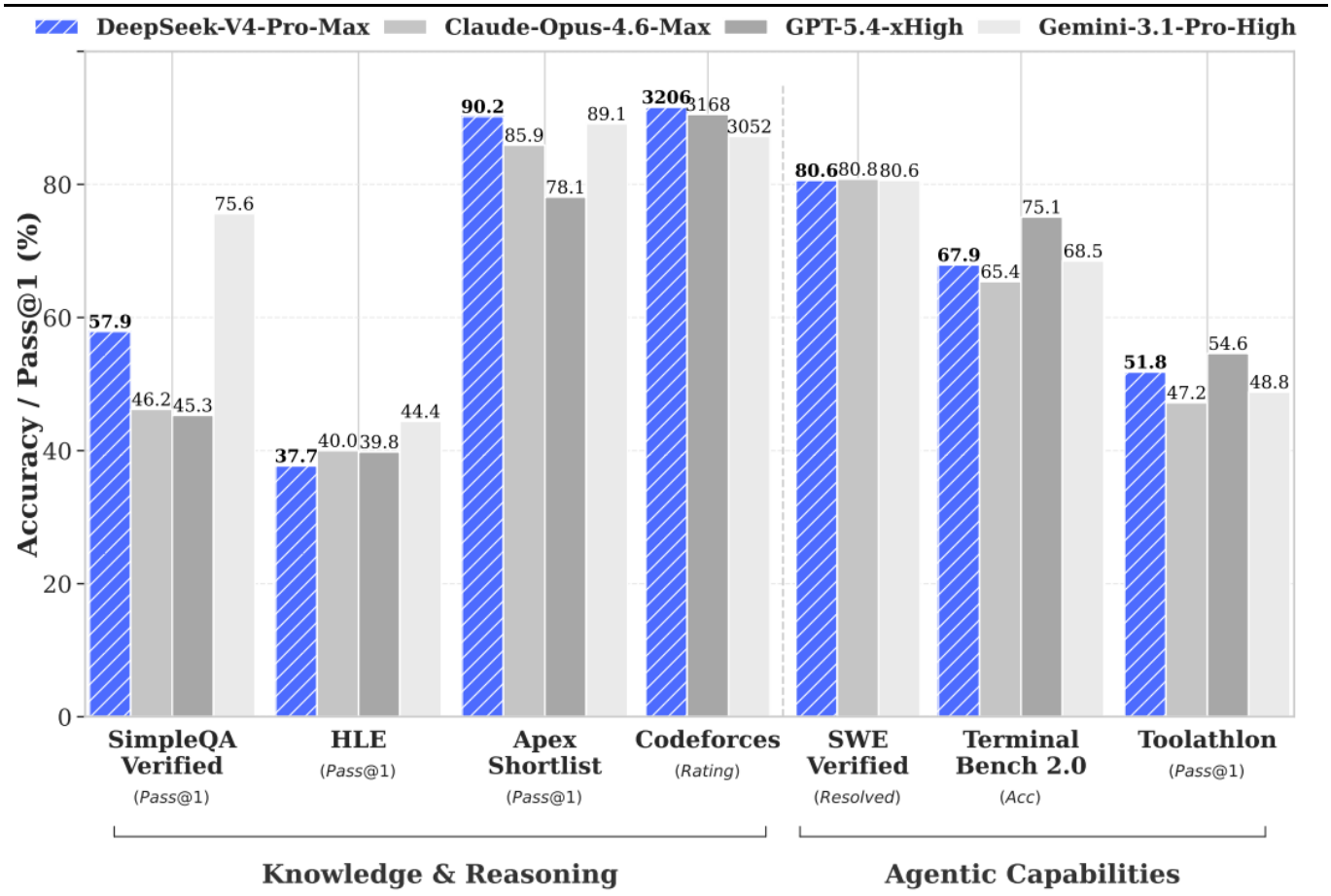
MoE（混合专家）模型的训练核心瓶颈在于专家并行的跨节点通信。传统专家并行方案中，通信与计算串行执行，对硬件的互联带宽与延迟要求极高，而国产硬件的互联能力往往是短板，难以承载万亿参数 MoE 模型的训练。

DeepSeek V4 自研的 MegaMoE2 融合内核，实现了专家并行的细粒度通信-计算重叠。MegaMoE2 通过将专家拆分调度为多个波次，在稳态下实现当前波次的计算、下一波次的 token 传输、已完成专家的结果发送三者并行执行，理论加速比可达 1.92 倍。该内核已在华为昇腾平台完成适配跑通，在通用推理负载上实现 1.50~1.73 倍的速度提升，在 RL 推理等延迟敏感场景中，最高提速可达 1.96 倍，彻底解决了国产硬件环境下 MoE 模型的通信瓶颈，让万亿参数 MoE 模型在国产算力上的训练成为可能。

3. 性能表现：跻身全球第一梯队，百万级上下文实现商用级突破

DeepSeek V4 系列在知识储备、推理代码、Agent 能力、长上下文、中文创作五大核心维度，均实现了开源模型的新标杆，多项指标比肩甚至超越国际顶级闭源模型。其中百万级上下文能力的突破，更是为长程 Agent 任务的落地奠定了核心基础。

图2：DeepSeek V4-Pro Max 在各项指标上与主要竞争对手对比



数据来源：DeepSeek V4 技术论文，东吴证券研究所

图3：DeepSeek V4 与其他竞争对手各项评分详细对比

	Benchmark (Metric)	Opus-4.6 Max	GPT-5.4 xHigh	Gemini-3.1-Pro High	K2.6 Thinking	GLM-5.1 Thinking	DS-V4-Pro Max
Knowledge & Reasoning	MMLU-Pro (EM)	89.1	87.5	**91.0**	87.1	86.0	87.5
	SimpleQA-Verified (Pass@1)	46.2	45.3	**75.6**	36.9	38.1	57.9
	Chinese-SimpleQA (Pass@1)	76.4	76.8	**85.9**	75.9	75.0	84.4
	GPQA Diamond (Pass@1)	91.3	93.0	**94.3**	90.5	86.2	90.1
	HLE (Pass@1)	40.0	39.8	**44.4**	36.4	34.7	37.7
	LiveCodeBench (Pass@1)	88.8	-	91.7	89.6	-	**93.5**
	Codeforces (Rating)	-	3168	3052	-	-	**3206**
	HMMT 2026 Feb (Pass@1)	96.2	**97.7**	94.7	92.7	89.4	95.2
	IMOAnswerBench (Pass@1)	75.3	**91.4**	81.0	86.0	83.8	89.8
	Apex (Pass@1)	34.5	54.1	**60.9**	24.0	11.5	38.3
	Apex Shortlist (Pass@1)	85.9	78.1	89.1	75.5	72.4	**90.2**
Long	MRCR 1M (MMR)	**92.9**	-	76.3	-	-	83.5
	CorpusQA 1M (ACC)	**71.7**	-	53.8	-	-	62.0
Agentic	Terminal Bench 2.0 (Acc)	65.4	**75.1**	68.5	66.7	63.5	67.9
	SWE Verified (Resolved)	**80.8**	-	80.6	80.2	-	80.6
	SWE Pro (Resolved)	57.3	57.7	54.2	**58.6**	58.4	55.4
	SWE Multilingual (Resolved)	**77.5**	-	-	76.7	73.3	76.2
	BrowseComp (Pass@1)	83.7	82.7	**85.9**	83.2	79.3	83.4
	HLE w/ tools (Pass@1)	53.1	52.0	51.6	**54.0**	50.4	48.2
	GDPval-AA (Elo)	1619	**1674**	1314	1482	1535	1554
	MCPAtlas Public(Pass@1)	**73.8**	67.2	69.2	66.6	71.8	73.6
	Toolathlon (Pass@1)	47.2	**54.6**	48.8	50.0	40.7	51.8

数据来源：DeepSeek V4 技术论文，东吴证券研究所

图4：DeepSeek V4 系列内部各项指标横比

	Benchmark (Metric)	DeepSeek-V4-Flash Non-Think	DeepSeek-V4-Flash High	DeepSeek-V4-Flash Max	DeepSeek-V4-Pro Non-Think	DeepSeek-V4-Pro High	DeepSeek-V4-Pro Max
Knowledge & Reasoning	MMLU-Pro (EM)	83.0	86.4	86.2	82.9	87.1	87.5
	SimpleQA-Verified (Pass@1)	23.1	28.9	34.1	45.0	46.2	57.9
	Chinese-SimpleQA (Pass@1)	71.5	73.2	78.9	75.8	77.7	84.4
	GPQA Diamond (Pass@1)	71.2	87.4	88.1	72.9	89.1	90.1
	HLE (Pass@1)	8.1	29.4	34.8	7.7	34.5	37.7
	LiveCodeBench (Pass@1-COT)	55.2	88.4	91.6	56.8	89.8	93.5
	Codeforces (Rating)	-	2816	3052	-	2919	3206
	HMMT 2026 Feb (Pass@1)	40.8	91.9	94.8	31.7	94.0	95.2
	IMOAnswerBench (Pass@1)	41.9	85.1	88.4	35.3	88.0	89.8
	Apex (Pass@1)	1.0	19.1	33.0	0.4	27.4	38.3
	Apex Shortlist (Pass@1)	9.3	72.1	85.7	9.2	85.5	90.2
Long	MRCR 1M(MMR)	37.5	76.9	78.7	44.7	83.3	83.5
	CorpusQA 1M(ACC)	15.5	59.3	60.5	35.6	56.5	62.0
Agentic	Terminal Bench 2.0 (Acc)	49.1	56.6	56.9	59.1	63.3	67.9
	SWE Verified (Resolved)	73.7	78.6	79.0	73.6	79.4	80.6
	SWE Pro (Resolved)	49.1	52.3	52.6	52.1	54.4	55.4
	SWE Multilingual (Resolved)	69.7	70.2	73.3	69.8	74.1	76.2
	BrowseComp (Pass@1)	-	53.5	73.2	-	80.4	83.4
	HLE w/ tools (Pass@1)	-	40.3	45.1	-	44.7	48.2
	MCPAtlas Public (Pass@1)	64.0	67.4	69.0	69.4	74.2	73.6
	GDPval-AA (Elo)	-	-	1395	-	-	1554
	Toolathlon (Pass@1)	40.7	43.5	47.8	46.3	49.0	51.8

数据来源：DeepSeek V4 技术论文，东吴证券研究所

3.1. 知识储备：开源模型新标杆，大幅缩小与闭源模型差距

DeepSeek-V4-Pro-Max 在事实知识与专业知识两大维度，均领跑全球开源赛道。在衡量参数化事实知识的 SimpleQA-Verified 基准上，模型取得 57.9 分，远超 Kimi-K2.6（36.9）、GLM-5.1（38.1）等主流开源模型；在中文 SimpleQA 基准上，模型得分达 84.4，直接逼近 Gemini-3.1-Pro 的 85.9，大幅缩小了开源模型与闭源头部模型的中文知识差距。

在高等教育级别的专业知识基准上，Pro-Max 版本在 MMLU-Pro、GPQA Diamond 分别取得 87.5、90.1 的得分，均位列开源模型第一。其中 GPQA Diamond 是面向研究生级别的多学科知识基准，90.1 的得分意味着模型在专业学术知识上的能力已接近人类专家水平，仅小幅落后于 Gemini-3.1-Pro 的 94.3。

3.2. 推理与代码能力：开源模型首次追平闭源头部水平

在代码与数学推理这两大硬核能力维度，DeepSeek V4 实现了开源模型的历史性突破。在 Codeforces 代码竞赛基准上，Pro-Max 版本取得 3206 的 Elo 评分，位列人类选手排行榜第 23 名，是开源模型首次在该基准上追平 GPT-5.4 等顶级闭源模型；即便是轻量化的 Flash-Max 版本，Codeforces 评分也达到 3052，推理性能追平 GPT-5.2、Gemini-3.0-Pro 等闭源模型，实现了小参数规模下的推理能力飞跃。

在其他核心基准上，Pro-Max 版本的 LiveCodeBench Pass@1 达 93.5，超越 Claude Opus 4.6（88.8）、Gemini-3.1-Pro（91.7），登顶该基准榜单；在 IMOAnswerBench 国际数学奥林匹克基准上，模型取得 89.8 的得分，仅略逊于 GPT-5.4 的 91.4，大幅领先其他开源竞品。

3.3. Agent 能力：达到闭源模型同级水平，开源赛道第一梯队

Agent 能力是大模型落地企业级场景的核心，DeepSeek V4 在真实场景的 Agent 基准上表现亮眼。在面向真实软件工程的 SWE Verified 代码 Agent 基准上，Pro-Max 版本的 SWE-bench Verified 任务解决分数达 80.6，与 Claude Opus 4.6 完全持平，超越了所有主流开源模型；在 Terminal Bench 2.0 终端操作基准上，模型取得 67.9 分，接近 GPT-5.4 的 75.1，处于开源模型第一梯队。

在通用工具调用场景中，模型在 MCPAtlas Public、Toolathlon 基准上分别取得 73.6、51.8 的 Pass@1 得分，均位列开源模型前列，证明其工具调用能力并非仅适配自研框架，而是具备极强的泛化能力，可适配各类自定义工具与场景。

3.4. 长上下文能力：百万 token 原生支持，解决长程任务核心瓶颈

长上下文能力是 DeepSeek V4 最核心的突破之一，模型原生支持 100 万 token 的上下文窗口，是当前开源大模型中极少数能常态化支持百万级上下文的 MoE 模型，其不仅实现了纸面参数的突破，更在实际基准测试中展现了商用级的可用性。

从技术效率来看，在 1M token 上下文场景下，DeepSeek-V4-Pro 的单 token 推理 FLOPs 仅为前代 DeepSeek-V3.2 的 27%，KV 缓存占用降至 10%；Flash 版本更是将单 token 推理 FLOPs 降至 **V3.2 的 10%，KV 缓存占用仅为 7%。**这一效率突破，让百万级上下文的推理不再是实验室技术，而是可以低成本常态化商用的能力，解决了此前长上下文推理成本过高、延迟过大的行业痛点。

从基准测试表现来看，模型在 OpenAI MRCR 长上下文多针检索基准上，**1M token 场景下取得 83.5 的 MMR 得分，**超越 Gemini-3.1-Pro 的 76.3；在面向真实场景的 CorpusQA 百万 token 语料级问答基准上，**模型取得 62.0 分，同样超越 Gemini-3.1-Pro 的 53.8。**MRCR 8-needle 测试显示，模型在 128K 上下文以内的检索能力保持高度稳定，平均 MMR 均在 0.9 以上，即便扩展到 1M token，平均 MMR 依然保持在 0.84，远高于同类长上下文模型，证明其超长上下文的信息留存能力具备实际商用价值。

长上下文能力的突破，从底层解决了 Agent 长程任务链拆解的核心痛点。传统 Agent 的最大瓶颈，在于复杂长周期任务中容易出现上下文遗忘与任务目标偏移：对于代码库全量重构、数十万字行业报告撰写、跨多文档法律尽调等复杂任务，往往需要拆解为数十上百个子任务，**传统 128K 以内的上下文窗口，无法承载全量的任务背景、历史交互记录与中间执行结果，**导致 Agent 在多轮执行中频繁出现信息丢失、重复执行、目标偏离等问题，任务失败率极高。

而 DeepSeek V4 的 1M token 原生上下文，约对应 75 万字的中文文本，足以一次性承载完整的中型代码库、上百份行业研报、全量企业规章制度与长周期任务执行日志。

Agent 可在单次上下文窗口内完成**任务理解、拆解、执行、校验、复盘**全流程，无需频繁对上下文进行截断与摘要，从根本上避免了信息丢失。同时，模型针对 Agent 场景优化了交错思考机制，在工具调用场景下完整保留所有轮次的推理内容，不会因新的用户消息刷新推理轨迹，搭配百万级上下文，让 Agent 可以在长周期任务中持续维护完整的思考链与执行记录，实现真正的长程任务闭环。这也是模型能在 SWE Verified、Terminal Bench 2.0 等长程 Agent 基准上取得优异表现的核心原因。

3.5. 中文创作能力：全面超越国际竞品，仅高难度场景略逊头部闭源模型

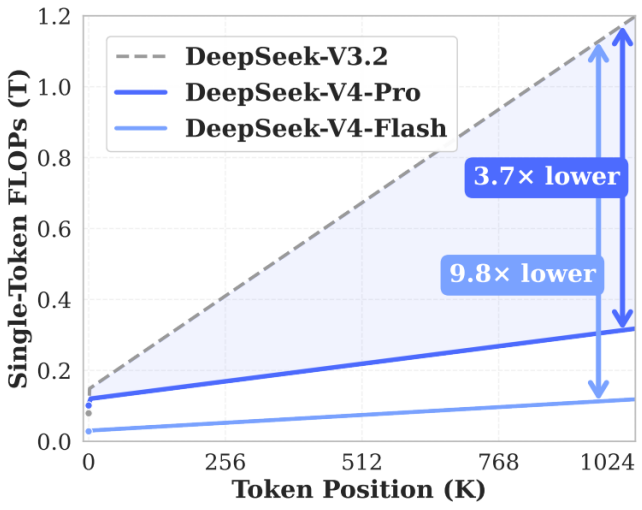
中文创作是 DeepSeek 系列的传统优势，V4 版本进一步扩大了这一领先性。在覆盖 7 大类、3170 个样本的功能性写作测试中，DeepSeek-V4-Pro 对 Gemini-3.1-Pro 的整体胜率达 62.7%，其中邮件书信类胜率 73.29%，技术文本类胜率 75.86%，办公报告类胜率 66.41%，几乎在所有功能性写作场景中均占据优势。

在创意写作维度，模型在 2837 个测试样本中，对 Gemini-3.1-Pro 的指令遵循胜率达 60.0%，写作质量胜率高达 77.5%。其中同人小说、小说故事类的写作质量胜率分别达到 83.25%、80.77%。仅在高难度多轮约束、复杂指令跟随场景下，模型对 Claude Opus 4.5 的胜率为 45.9%，略逊于闭源头部模型。

4. 技术架构：底层创新实现效率与能力的双重突破

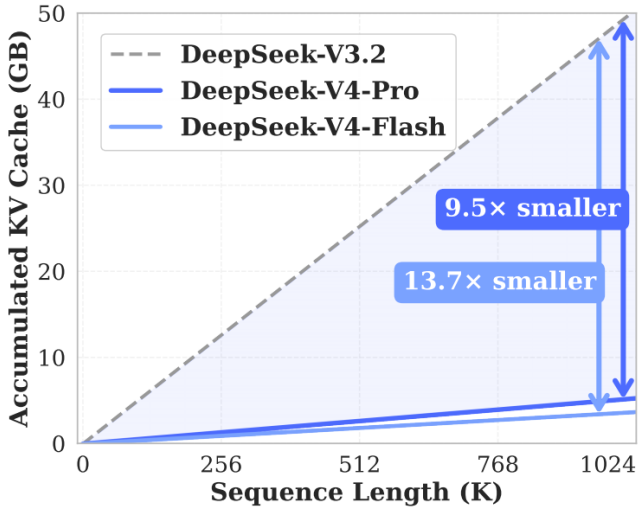
DeepSeek V4 的能力突破，源于架构层面的三大核心创新：**CSA+HCA 混合注意力架构、mHC 流形约束超连接、Muon 优化器与创新训练范式，**从底层解决了超长上下文效率、深层大模型训练稳定性两大行业核心痛点。在这些技术架构创新加持下，**DeepSeek-V4-Pro 的单 token 推理 FLOPs 仅为前代 DeepSeek-V3.2 的 27%，KV 缓存占用降至 10%；**Flash 版本更是将单 token 推理 FLOPs 降至 **V3.2 的 10%，KV 缓存占用仅为 7%。**

图5：DeepSeek V4 与 V3.2 的计算量对比



数据来源：DeepSeek V4 技术论文，东吴证券研究所

图6：DeepSeek V4 与 V3.2 的显存容量对比

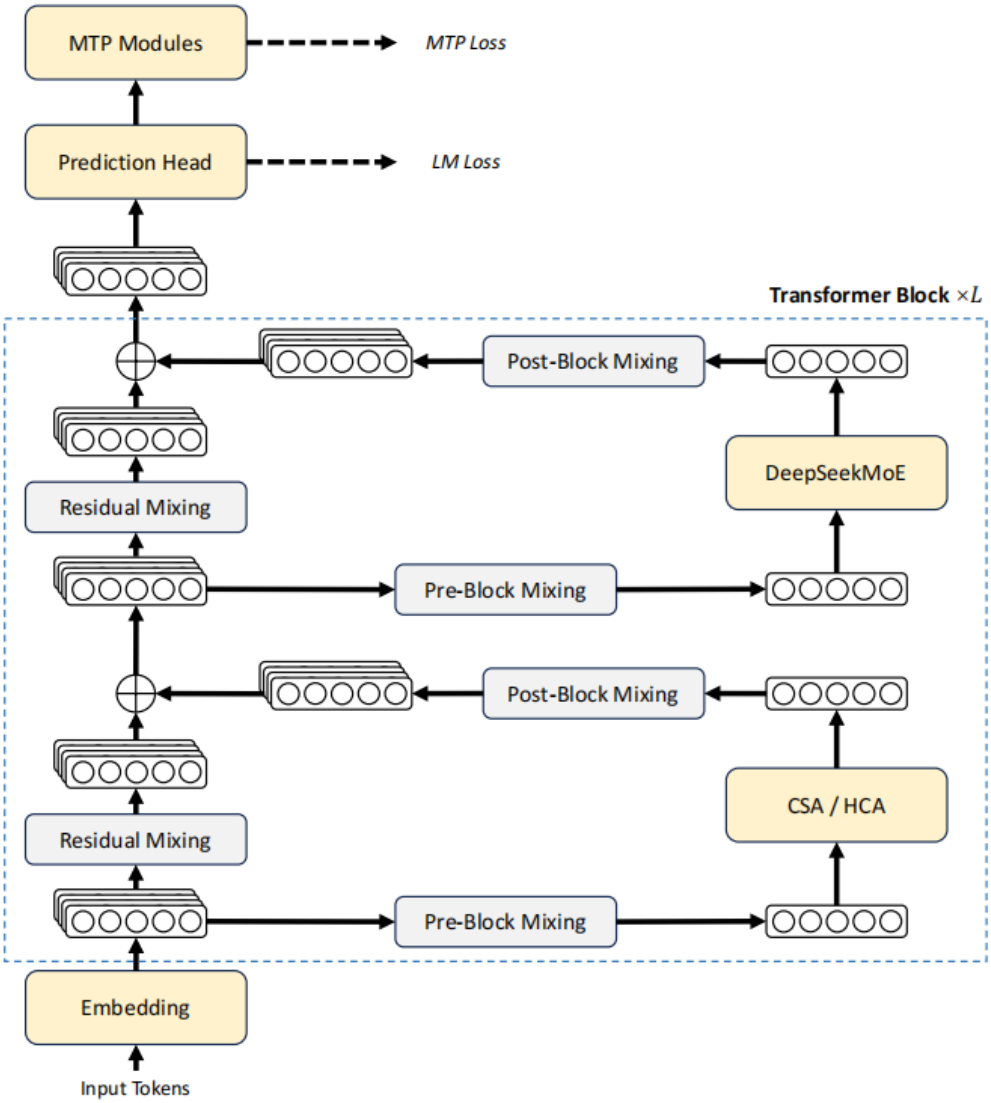


数据来源：DeepSeek V4 技术论文，东吴证券研究所

4.1. CSA+HCA 混合注意力架构：彻底打破超长上下文的算力瓶颈

传统 Transformer 的自注意力机制为二次计算复杂度，随着上下文长度增长，计算量与 KV 缓存占用呈平方级上升，当上下文扩展到百万 token 时，传统注意力的计算成本会达到无法商用的水平。DeepSeek V4 首创 CSA 与 HCA 交替的混合注意力架构，通过“分层压缩+稀疏选择”的组合设计，将注意力计算复杂度从 $O(n^2)$降至 $O(n)$，从底层解决了超长上下文的算力瓶颈。

图7：DeepSeek V4 保留 Transformer 架构和 MTP 模块，同时引入 mHC、CSA+HCA



数据来源：DeepSeek V4 技术论文，东吴证券研究所

4.1.1. Compressed Sparse Attention（CSA，压缩稀疏注意力）

CSA 的核心逻辑是“先压缩，再稀疏选择”，分两步实现注意力效率的飞跃：

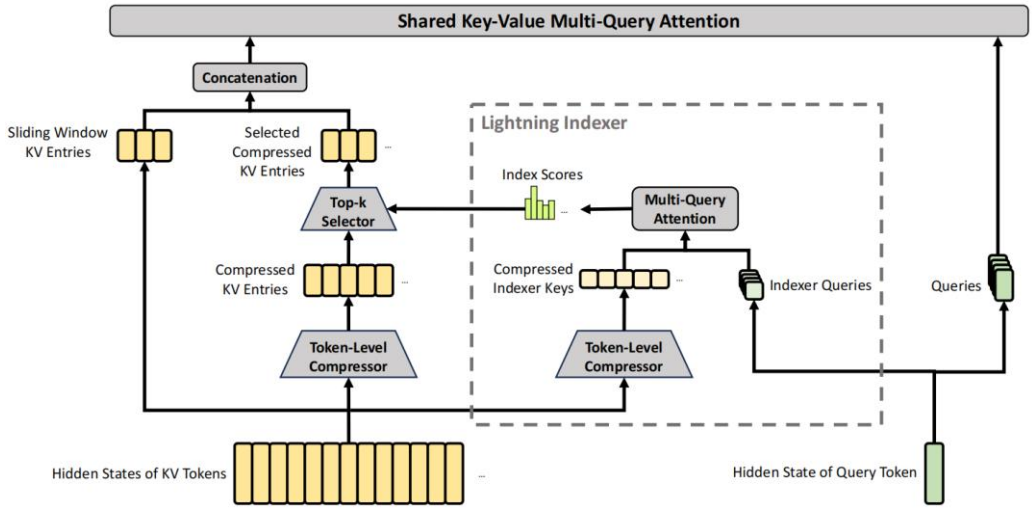
第一步：KV 缓存序列压缩：CSA 先将每 4 个 token 的 KV 缓存压缩为 1 个条目，序列长度直接压缩为原来的 1/4。具体而言，模型通过可训练的参数矩阵计算 KV 条目与对应的压缩权重，结合可学习的位置偏置，通过 Softmax 计算每个 token 的压缩权重，最终通过加权求和得到压缩后的 KV 条目；同时采用重叠压缩设计，每个压缩条目来自 2 倍压缩长度的 KV 条目，避免了边界信息丢失，保证了压缩后的信息完整性。

第二步：稀疏注意力精准选择：压缩完成后，CSA 通过闪电索引器，为每个查询 token 选择 top-k 个最相关的压缩 KV 条目，仅对选中的条目执行注意力计算。Pro 版本

的 top-k 设置为 1024，这意味着即便上下文长度达到 1M token，每个查询 token 仅需对 1024 个压缩 KV 条目执行计算，彻底打破了二次复杂度的瓶颈。

同时，CSA 还做了多项细节优化：新增滑动窗口注意力分支，为每个查询 token 保留最近 128 个未压缩的 KV 条目，保证局部细粒度依赖的建模；采用共享 KV 的多查询注意力（MQA）与分组输出投影策略，进一步降低计算量；对查询和 KV 条目执行 RMSNorm，避免注意力对数爆炸，提升训练稳定性；采用部分旋转位置编码（RoPE），保证相对位置信息的精准建模。

图8：CSA 的核心架构，系统将 KV 数量压缩至 1/m 倍，随后应用 DSA 机制进一步加速



数据来源：DeepSeek V4 技术论文，东吴证券研究所

4.1.2. Heavily Compressed Attention（HCA，重度压缩注意力）

HCA 的核心逻辑是“极致压缩，稠密注意力”，与 CSA 形成能力互补。HCA 采用 128 倍的极致压缩率，每 128 个 token 的 KV 缓存压缩为 1 个条目，1M token 的序列经过压缩后仅剩下不到 8000 个条目，此时稠密注意力的计算量已完全在可接受范围内，因此 HCA 不做稀疏选择，直接对压缩后的全量 KV 条目执行稠密注意力计算，保证了对全局序列的整体语义建模，避免稀疏注意力导致的全局信息丢失。

4.1.3. 混合注意力架构的设计价值

DeepSeek V4 在 Transformer 层中采用 CSA 与 HCA 交替的 interleaved 配置，一层使用 CSA，下一层使用 HCA，以此类推。这种设计实现了“全局语义建模”与“关键信息精准捕捉”的完美平衡：CSA 通过稀疏选择，适配长距离的稀疏依赖建模，精准捕捉超长序列中的关键信息；HCA 通过极致压缩的稠密注意力，实现对全局序列的整体语义理解。两者交替使用，让模型在极低的计算成本下，同时具备局部细节建模、关键信息

检索、全局语义理解的三重能力，这也是其能在 1M token 上下文下依然保持高准确率的核心原因。

4.2. mHC 流形约束超连接：升级残差结构，解决万亿模型训练稳定性难题

随着模型参数规模扩大到万亿级别，传统 Transformer 的残差连接会出现信号传播不稳定、梯度消失/爆炸的问题，MoE 模型训练过程中极易出现 Loss Spike（损失尖峰），导致训练崩溃。DeepSeek V4 引入的 Manifold-Constrained Hyper-Connections（**mHC，流形约束超连接**），对传统残差连接实现了全面升级，解决了深层 MoE 模型的训练稳定性与表达能力的双重问题。

传统的 Hyper-Connections（HC）技术通过扩展残差流的宽度，为模型提供了额外的缩放维度，用极小的计算开销提升了表达能力，但无约束的残差映射矩阵会导致多层堆叠时出现数值爆炸，训练极不稳定。**而 mHC 的核心创新，是将残差映射矩阵约束到双随机矩阵的流形（Birkhoff 多面体）上，**从数学上保证了映射矩阵的谱范数不超过 1，残差变换是非扩张的，无论是前向传播还是反向传播，信号都不会出现无限制的放大，从根本上解决了多层堆叠时的数值不稳定问题。同时，双随机矩阵的集合在乘法下是闭合的，即便堆叠上百层 mHC，依然能保证训练稳定性。

在具体实现上，mHC 通过 Sinkhorn-Knopp 算法，将无约束的原始参数投影到双随机矩阵流形上，同时对输入、输出映射通过 Sigmoid 函数保证非负性与有界性，避免信号抵消风险；参数采用动态生成方式，分解为输入相关的动态分量与静态分量，进一步提升模型表达能力。通过融合内核、选择性重计算等工程优化，mHC 的额外耗时仅占流水线阶段的 6.7%，几乎不增加训练开销，却能大幅提升模型的训练稳定性与参数效率。

mHC 的应用，让 DeepSeek V4-Flash 即便只有 13B 激活参数，依然能超越前代 37B 激活参数的 DeepSeek-V3.2，实现了参数效率的飞跃。同时，其稳定的信号传播能力，也为超长上下文的建模提供了支撑，让注意力模块的信号能在数十层 Transformer 中稳定传播，避免了超长上下文下的信号衰减。搭配 Muon 优化器、预期性路由、SwiGLU 钳制技术，mHC 彻底解决了万亿参数 MoE 模型训练的 Loss Spike 难题，实现了 1.6T 参数模型的稳定训练，这也是模型能在国产算力平台上完成训练的重要基础。

4.3. 创新后训练范式：规避多能力融合的性能退化

DeepSeek V4 采用“领域专家独立训练+全词表在线蒸馏（OPD）”的两阶段后训练范式：先针对数学、代码、Agent、指令跟随等领域，分别训练独立的专家模型，每个专家先经过有监督微调（SFT）建立基础能力，再通过 GRPO 强化学习优化领域对齐能力；最终通过多教师在线蒸馏，将多个专家的能力融合到一个统一模型中。这一范式规避了

传统多能力融合中常见的性能退化问题，让模型同时具备多个领域的顶尖能力，实现了“全能力无短板”的模型优化目标。

5. 综合评价：开源模型标杆，国产化战略意义远超性能表现

综合来看，DeepSeek V4 整体表现符合市场预期，在百万级上下文、国产化适配层面超出市场预期，是当前全球开源大模型的标杆之作，但其在多模态能力、复杂任务指令跟随上，仍未达到市场的高预期。

从超预期的维度来看，模型在推理、代码能力上延续了 DeepSeek 系列的传统优势，且实现了对前代版本的全面超越；更重要的是，模型以原生 1M token 上下文支持、万亿参数模型的国产芯片训练适配，实现了两大历史性突破，成为首个完成该级别国产化适配的顶级开源大模型，为国产大模型与国产算力的协同发展打开了全新空间。同时，模型具备极高的商用性价比，Flash 版本定价仅 1 元/输入百万 token、2 元/输出百万 token，Pro 版本相较海外顶级闭源模型便宜约 60%，后续随着华为昇腾 950 产能释放，仍存在进一步的降价空间。

图9：DeepSeek V4 价格，后续随昇腾 950 节点放量有望大幅下降

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	

*受限于高端算力，目前 Pro 的服务吞吐十分有限，预计下半年昇腾 950 超节点批量上市后，Pro 的价格会大幅下调。

数据来源：DeepSeek 官方公众号，东吴证券研究所

从不及预期的维度来看，多模态能力并未随本次 V4 版本发布落地，与同期 GPT-5.5、Claude Opus 4.7 等模型的多模态升级形成差距，让部分用户产生了心理落差。同时，模型在实测中也暴露了部分短板：复杂 Agent 场景下对自定义工具的调用触发灵敏度不足，多约束条件的开发任务中存在未与用户确认便执行操作的问题，高难度多轮创意写作场景的表现仍落后于 Claude Opus 4.5，通用复杂任务的鲁棒性仍有优化空间。

从实际落地表现来看，模型在工程测评中表现令人满意：中文创作能力实测优异，长文档生成与续写效果显著优于多数竞品；代码开发能力在内部真实 R&D 任务中通过率达 67%，远超 Claude Sonnet 4.5，接近 Opus 4.5 水平；在 30 个覆盖 13 个行业的高级中文专业办公任务中，模型对 Opus-4.6-Max 的非损失率达到 63%，在任务完成度、内容质量上具备显著优势。

6. 投资建议

DeepSeek V4 的发布，是国产通用大模型在国产算力训练侧从 0 到 1 的里程碑式突破。在此之前，国产算力在大模型领域的应用大多局限于推理侧，核心训练环节始终被 NVIDIA 生态垄断，而 DeepSeek 本次从模型内核到训练架构、再到推理全流程，均实现了国产算力的深度适配，完成了真正意义上的国产化训练从 0 到 1 跨越。因此，无论 DeepSeek V4 的性能表现如何，其对国产算力生态的战略意义，都远大于模型本身的性能意义。

DeepSeek V4 证明了国产算力已经具备承载顶级通用大模型训练的能力，随着模型架构与国产硬件的持续协同优化，国产大模型的训推全流程国产化将成为行业趋势，国产算力产业链将迎来全新的发展机遇。

国产算力相关标的：

CPU 产业链核心配套标的：禾盛新材；

国产 AI 芯片与算力基础设施核心厂商：寒武纪-U、海光信息、中科曙光；

国产 GPU 新锐厂商：摩尔线程-U、沐曦股份-U；

算力硬件上游核心供应链：华丰科技、航天电器。

7. 风险提示

大模型迭代节奏不及预期：当前大模型技术迭代节奏若不及市场普遍预期，通用推理、多模态融合等核心能力突破缓慢，行业场景落地与商用闭环进度将明显滞后。

国产算力软硬件生态适配进度不及预期：国产算力供给稳定性与适配效率不足，直接拖累下游大模型商业化落地与行业应用规模化推广。

大模型行业市场竞争持续加剧：海内外科技巨头与初创企业扎堆入局，价格战与高额研发投入叠加，份额争夺持续挤压企业盈利增长空间。

行业政策监管持续趋严：数据安全、内容合规、隐私保护等要求不断升级，企业合规成本抬升，显著限制业务拓展与技术创新空间。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）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。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。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。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中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投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn