爱建证券 AIJIAN SECURITIES

行业研究 / 行业点评

2026 年 05 月 14 日

DeepSeek V4 发布,国产算力乘风起航

——电子行业专题报告

行业及产业

电子

强于大市

一年内行业指数与沪深 300 指数对比走势：

资料来源：聚源数据，爱建证券研究所

相关研究

《电子行业周报：NVIDIA 入局 Corning 锁定光互联核心产能，筑牢 AI 算力供应链根基》2026-05-11

《电子行业周报：HBM 迭代筑牢 Samsung AI 芯片壁垒》2026-05-06

《电子行业周报：DeepSeek 发布并同步开源 V4 大模型》2026-04-27

《电子行业周报：AI 需求持续高景气，先进制程产能加速扩张》2026-04-20

《电子行业周报：Samsung 2026Q1 业绩指引创历史新高》2026-04-13

证券分析师

许亮

S0820525010002

0755-83562506

xuliang@ajzq.com

联系人

朱俊宇

S0820125040021

021-32229888-25520

zhujunyu@ajzq.com

投资要点：

- **事件：**2026 年 4 月 24 日，DeepSeek 发布并同步开源了其全新大模型产品 DeepSeek V4 预览版。该版本的核心亮点为 1M 超长上下文能力，同时在 Agent 交互、世界知识储备与推理性能上，均实现了开源大模型领域的全面领先。

- **DeepSeek 以开源为核心战略，凭借极致成本控制与技术迭代快速崛起。**2023 年 11 月，DeepSeek 发布首个开源代码大模型 DeepSeek Coder，支持多语言生成与调试，性能超越 CodeLlama。同期开源通用大模型 DeepSeek LLM 67B，对标 LLaMA2 70B，中英文任务表现领先。2024 年 12 月，通用模型迭代至 DeepSeek-V3，训练成本仅 550 万美元，性能对标国际闭源模型，生成速度提升 3 倍。2025 年 1 月 20 日，发布第一代推理模型 DeepSeek-R1-Zero 和 DeepSeek-R1，补齐复杂逻辑推理能力短板。此后公司持续迭代核心能力，2026 年 4 月 24 日，正式发布 DeepSeek V4 预览版并同步开源。

- **本次发布的 DeepSeek V4 包含两款 MoE 模型，全系拥有 1M 超长上下文能力，同时在 Agent 交互、世界知识储备等方面表现突出。**DeepSeek-V4-Pro 总参数 1.6 万亿、激活参数 49B，主打高性能研发场景，综合能力对标行业顶级闭源大模型；DeepSeek-V4-Flash 总参数 2840 亿激活参数 13B，侧重轻量化部署与低成本高吞吐，可满足大规模日常推理需求。技术层面，模型创新融合 CSA 与 HCA 混合注意力架构，通过 KV 压缩与稀疏注意力协同优化长上下文推理效率，在 1M 上下文长度下，V4-Pro 单 Token 计算量仅为前代 V3.2 的 25%，KV 缓存占用进一步降至 10%，并首次将百万级超长上下文能力设为全系标配，为大模型规模化商用奠定坚实基础。性能方面，DeepSeek V4 Pro 综合表现已比肩行业顶级闭源大模型，其中 Agent 与编程能力位居开源第一梯队，世界知识储备领先同类产品，可全面支撑编程开发、工具调用、数学推理等高阶复杂任务。价格端具备显著竞争优势，据 DevTk.AI，V4-Pro 调用价格显著优于同级别 Claude Sonnet 4.6；V4-Flash 主打高吞吐场景，性价比优势更为突出，显著降低了企业级 AI 应用的落地成本与门槛。

- **寒武纪、摩尔线程、沐曦股份、海光信息等厂商相继完成 DeepSeek-V4 系列模型的部署。**寒武纪依托自研算子库对模型核心模块专项加速，深度优化热点算子并在 vLLM 框架中全面支持混合并行技术，充分释放硬件底层算力；摩尔线程在旗舰级 MTT S5000 GPU 上完成全链路工程化适配，其原生支持 FP8 的硬件架构可高效匹配模型“FP4+FP8”混合精度策略，相比 BF16/FP16 传统精度降低 50%显存带宽压力，形成差异化优势；沐曦联合 FlagOS 与 KernelSwift 智能算子迁移系统完成 Day 0 适配，通过核心算子优化实现国产芯片端平均 3.4 倍推理加速，大幅缩短模型适配周期；海光依托自研 DTK 异构计算平台与集成超 2000 个算子的 DAS 软件系统，对模型实现全栈深度调优，达成业界领先的计算效率。

- **投资建议：**DeepSeek V4 模型实现超长上下文、推理及 Agent 能力全面升级，叠加 CSA/HCA 混合注意力架构带来显著成本与性能优势，同时华为昇腾、寒武纪、摩尔线程、沐曦股份、海光信息等国产算力厂商完成快速适配，芯模协同生态持续完善，有望带动 AI 大模型及国产算力产业链需求加速释放。建议关注国产 AI 芯片产业链的投资机会。

- **风险提示：1）技术迭代不及预期风险；2）商业化落地放缓风险；3）行业竞争加剧风险。**

目录

图表目录

1. DeepSeek 发布 V4 新版本

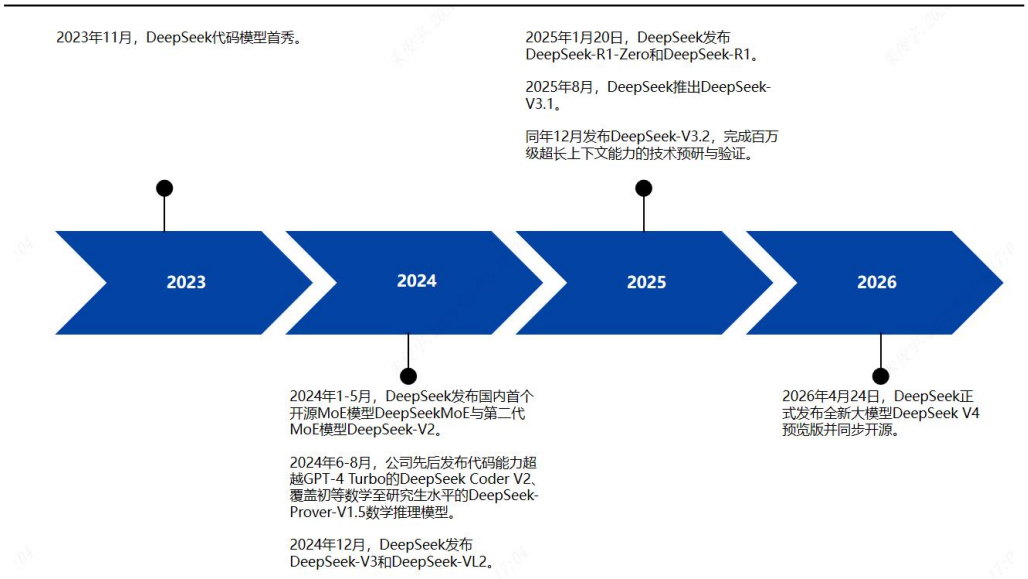
事件：2026 年 4 月 24 日，DeepSeek 发布并同步开源了其全新大模型产品 DeepSeek V4 预览版。该版本的核心亮点为 1M 超长上下文能力，同时在 Agent 交互、世界知识储备与推理性能上，均实现了开源大模型领域的全面领先。

1.1 DeepSeek 发展史梳理

2023 年 11 月，DeepSeek 代码模型首秀。主要包括：DeepSeek Coder：首个开源代码大模型，支持多语言生成与调试，且性能超越 CodeLlama，奠定了技术口碑。DeepSeek LLM 67B：通用大模型开源，对标 LLaMA2 70B，中英文任务表现领先。

2024 年 1-5 月，DeepSeek 发布国内首个开源 MoE 模型 DeepSeekMoE 与第二代 MoE 模型 DeepSeek-V2，完成细粒度专家共享架构、MLA 多头潜在注意力技术的核心突破，将推理成本压至 LLaMA3 的 1/4、API 定价低至 GPT-4 Turbo 的 1/70，大幅拉低 AI 使用成本；2024 年 6-8 月，公司进一步实现多领域性能跃升，先后发布代码能力超越 GPT-4 Turbo 的 DeepSeek Coder V2、覆盖初等数学至研究生水平的 DeepSeek-Prover-V1.5 数学推理模型。

图表 1：DeepSeek 发展史梳理



资料来源：DeepSeek Github，DeepSeek 微信公众号，爱建证券研究所

2024 年 12 月，DeepSeek 实现通用模型的迭代。DeepSeek-V3 发布，公司宣称训练成本仅 550 万美元，性能对标国际闭源模型，生成速度提升 3 倍。DeepSeek-VL2（2024 年 12 月）：多模态 MoE 模型，视觉能力显著提升。2025 年 1 月 20 日，DeepSeek 正式发布第一代推理模型 DeepSeek-R1-Zero 和 DeepSeek-R1。

2025 年 8 月，DeepSeek 推出 DeepSeek-V3.1，采用混合推理架构，原生支持智能体工具调用，实现向 AI Agent 方向的关键技术突破；同年 12 月发布 DeepSeek-V3.2，完成百万级超长上下文能力的技术预研与验证。2026 年 4 月 24 日，DeepSeek 正式发布全新大模型 DeepSeek V4 预览版并同步开源。

1.2 DeepSeek 模型架构创新

本次发布的 DeepSeek V4 包含两款混合专家（MoE）模型，分别适配不同场景需求。其中 DeepSeek-V4-Pro 主打高性能研发，总参数 1.6 万亿、激活参数 49B，聚焦尖端复杂任务，整体性能对标行业顶级闭源大模型；DeepSeek-V4-Flash 主打轻量化、低成本落地，总参数 2840 亿、激活参数 13B，主打高性价比部署。两款模型均原生搭载 1M 超长上下文能力，全面纳入 DeepSeek 官方标配服务体系。

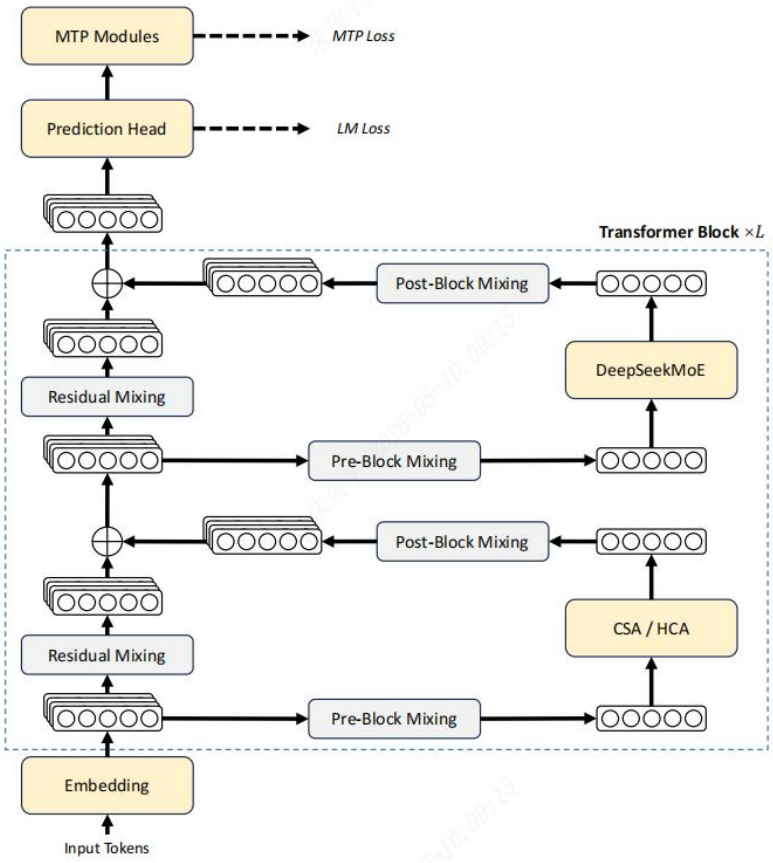
图表 2：DeepSeek V4 包含 Pro、Flash 系列

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：DeepSeek 微信公众号，爱建证券研究所

DeepSeek V4 创新性地融合了 CSA（Compressed Sparse Attention）与 HCA（Heavily Compressed Attention）两种技术，构建了高效的混合注意力架构。在显著降低长上下文推理显存占用的同时，大幅提升了模型的推理吞吐量。

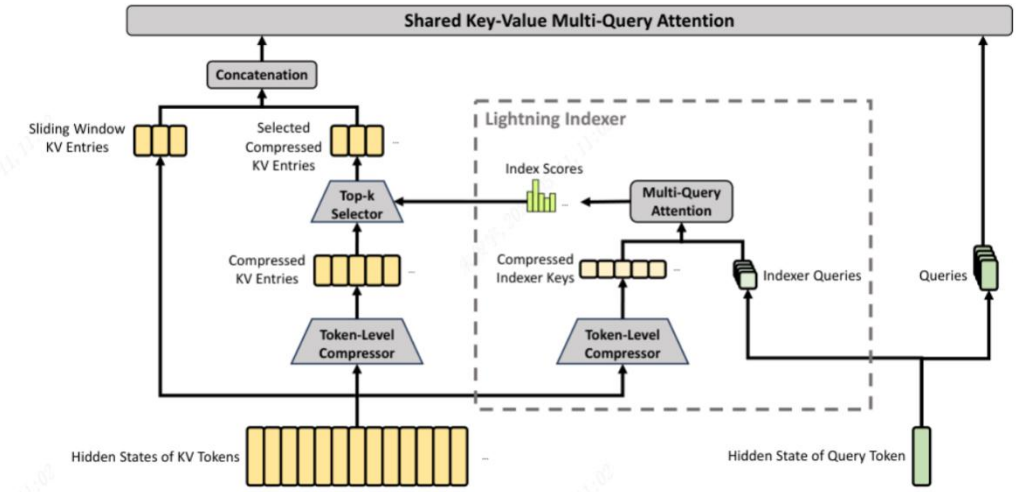
图表 3：DeepSeek V4 采取 CSA/HCA 新架构



资料来源：DeepSeek Towards Highly Efficient Million-Token Context Intelligence，爱建证券研究所

CSA 结合了压缩和稀疏注意力策略，它首先将每个 Token 的 KV 压缩成一个条目，然后应用 DeepSeek Sparse Attention（DSA），其中每个查询 Token 只关注注意力打分最高的 Top-k 个压缩 KV 条目，从而降低计算复杂度。

图表 4：DeepSeek V4 CSA 核心架构

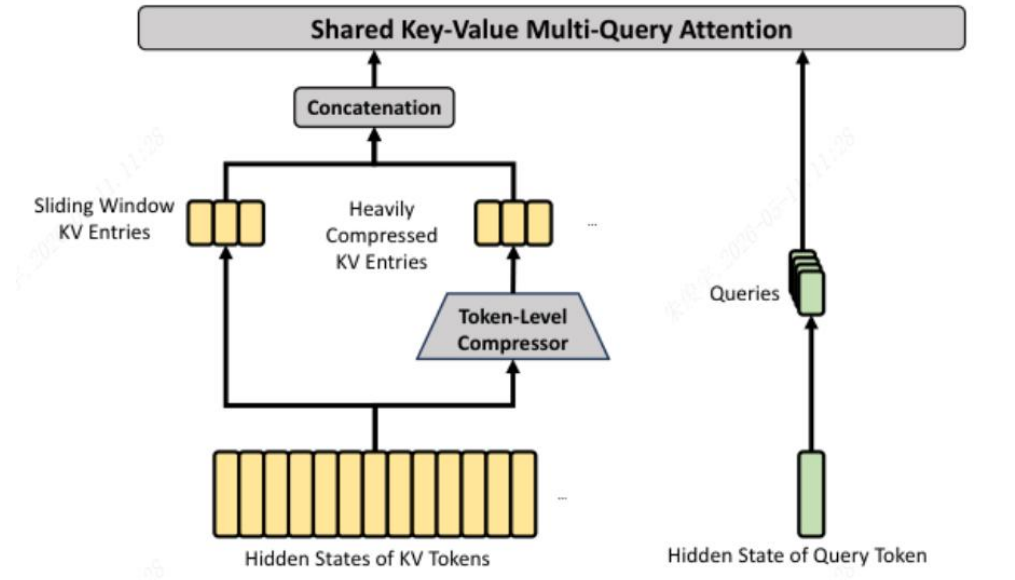


资料来源：DeepSeek Towards Highly Efficient Million-Token Context Intelligence，爱建证券研究所

HCA 采用与 CSA 同源的压缩思路，通过对 KV Cache 进行块级聚合，将每 m'个连续 Token 的 KV 缓存合并为单个紧凑条目，进一步提升了压缩比。

这种 CSA+HCA 的混合注意力架构，大幅优化了 DeepSeek-V4 系列的长上下文推理效率与显存占用。

图表 5：DeepSeek V4 HCA 核心架构

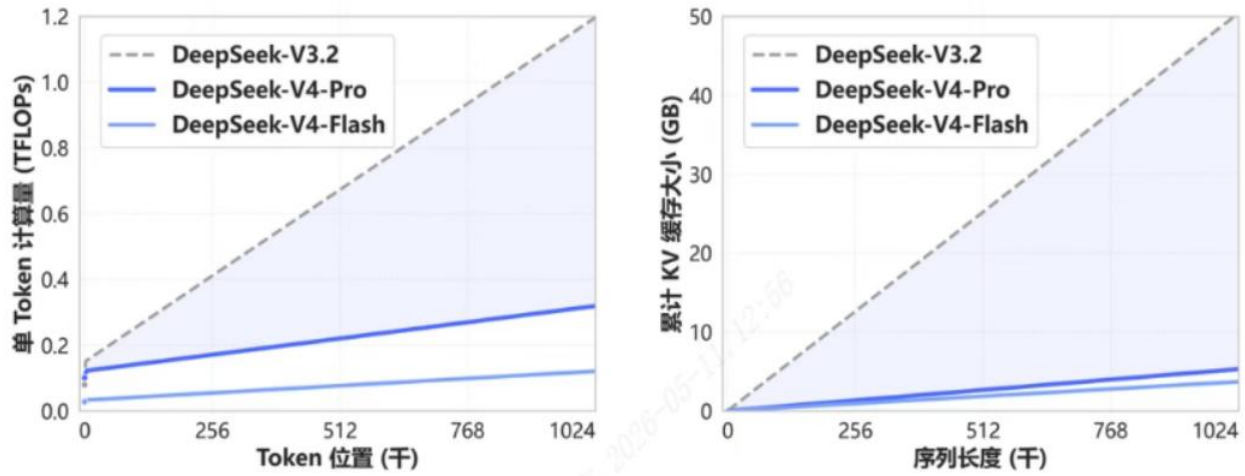


资料来源：DeepSeek Towards Highly Efficient Million-Token Context Intelligence，爱建证券研究所

从实测数据来看，在 1M 上下文长度下，DeepSeek-V4-Pro 的单 Token 计

算量仅为上一代 V3.2 的 25%，累计 KV 缓存占用更是仅为后者的 10%。正是依托这一突破性的效率提升，DeepSeek 首次将 1M 超长上下文能力从高端付费功能降维为全系列产品的标配，为大模型的规模化商用奠定了核心技术基础。

图表 6：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：DeepSeek 微信公众号，爱建证券研究所

1.3 DeepSeek V4 Pro 性能比肩顶级闭源模型

DeepSeek V4 Pro Agent 能力显著升级，兼具完备世界知识与世界级推理能力，综合性能已比肩行业顶级闭源大模型。

1）相较于前代模型，DeepSeek-V4-Pro 的 Agent 能力明显增强。在 Agentic Coding 评测中位列开源模型第一梯队；目前已成为 DeepSeek 内部主力开发模型。据官方评测，其使用体验优于 Claude Sonnet 4.5，代码交付质量接近 Claude Opus 4.6 非思考模式，但与 Opus 4.6 思考模式仍存在一定差距。

2）DeepSeek-V4-Pro 拥有完备的世界知识储备，在专项测评中大幅领先同类开源模型，整体表现仅稍逊于 Gemini-Pro-3.1 顶尖闭源模型。

3）模型聚焦自主编程、工具调用、数学及 STEM 等高阶复杂任务，依托超大参数量与高效激活推理机制，基准测试表现突出，可为复杂智能任务提供高性能求解方案。

图表 7：DeepSeek V4 系列性能卓越

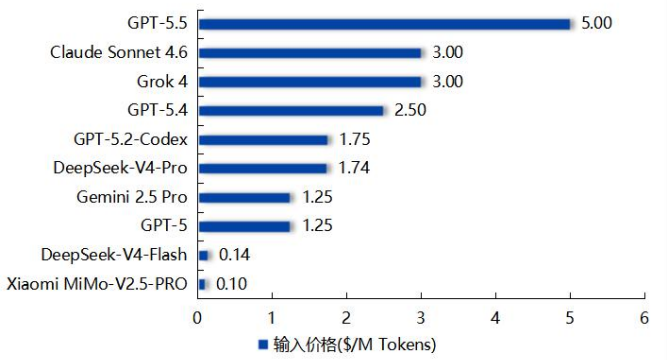
Benchmark (metric) Reasoning Effort		DS-V4-Pro Max	DS-V4-Flash Max	K2.6 Thinking	GLM-5.1 Thinking	Opus-4.6 Max	GPT-5.4 xHigh	Gemini-3.1-Pro High
Knowledge & Reasoning	MMLU-Pro (EM)	87.5	86.2	87.1	86.0	89.1	87.5	91.0
	SimpleQA-Verified (Pass@1)	57.9	34.1	36.9	38.1	46.2	45.3	75.6
	Chinese-SimpleQA (Pass@1)	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：DeepSeek 微信公众号，爱建证券研究所

通过上图，我们发现：编程能力是本次 DeepSeek-V4 迭代提升的亮点。DeepSeek-V4-Pro 在 SWE-bench Verified、LiveCodeBench、Terminal Bench 2.0 评测中分别取得 80.6、93.5、67.9 的成绩，整体位居开源模型前列。

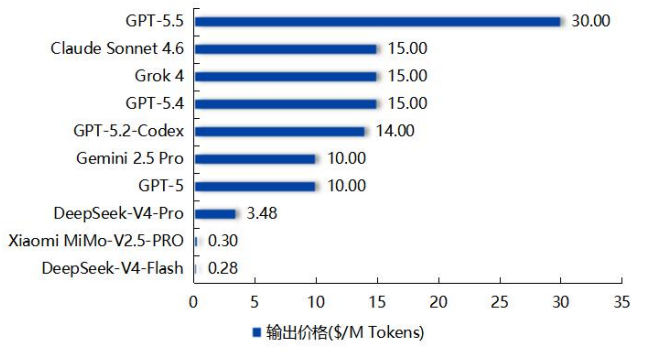
相较于市场主流核心大模型，DeepSeek 系列 API 价格优势显著。据 DevTk.AI 数据，DeepSeek-V4-Pro 输入、输出原价分别为 1.74 美元/百万 Tokens、3.48 美元/百万 Tokens（当前 DeepSeek-V4-Pro 模型开启限时 2.5 折优惠，优惠期至 2026/05/31；为客观反映长期定价竞争力，本次采用原价进行测算），价格显著优于同定位 1M 上下文商用模型 Claude Sonnet 4.6；V4-Flash 系列主打高性价比、高吞吐场景，输入、输出价格分别为 0.14 美元/百万 Tokens、0.28 美元/百万 Tokens。

图表 8：DeepSeek 输入价格优势明显



资料来源：DevTK.AI，爱建证券研究所

图表 9：DeepSeek 输出价格优势明显

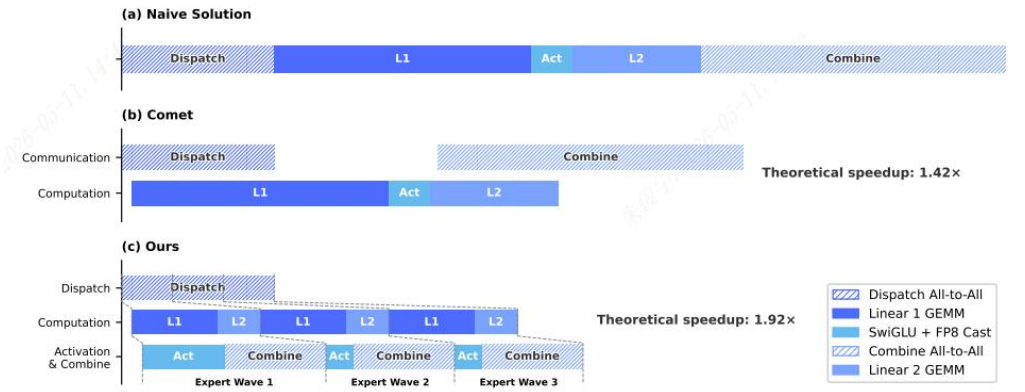


资料来源：DevTK.AI，爱建证券研究所

2. 国产算力厂商相继完成 DeepSeek-V4 系列模型的部署

相较于过往仅依托 NVIDIA CUDA 框架的运行模式，DeepSeek-V4 本次已在 NVIDIA GPU 与华为昇腾 NPU 双硬件平台，完成细粒度专家并行（EP）方案的落地验证。根据 DeepSeek-V4 报告，在通用推理场景下，相较传统非融合基线方案，推理性能实现显著提升；针对强化学习推演、高并发智能体服务等延迟敏感型业务场景，最高可达成 1.92 倍推理加速，适配性与工程落地性能优势显著。

图表 10：DeepSeek EP 方案示意图



资料来源：DeepSeek Towards Highly Efficient Million-Token Context Intelligence，爱建证券研究所

其中，昇腾平台的适配落地尤为亮眼：4 月 24 日 DeepSeek-V4 发布当日，昇腾即完成全系列 Day 0 适配。依托 CANN 架构，950 PR/DT 系列面向低时延场景实现 10-20ms 级推理，Atlas-A3 系列面向高吞吐场景实现 30ms 级推理，标志国产芯模协同实现里程碑突破。

图表 11：昇腾 Day 0 支持 DeepSeek-V4

资料来源：昇腾 CANN，爱建证券研究所

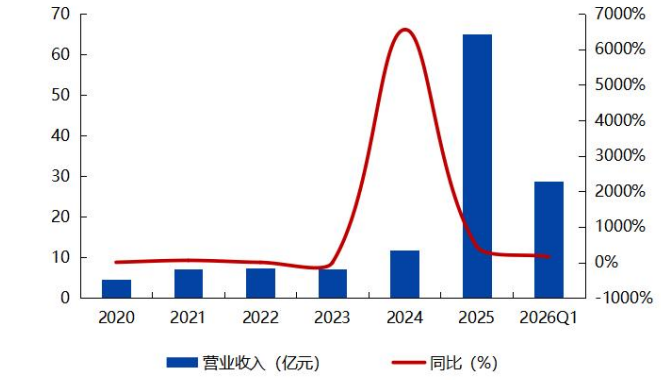
除华为昇腾外，寒武纪、摩尔线程、沐曦股份、海光信息等国产算力厂商也相继完成 DeepSeek-V4 系列模型的适配与推理部署，多平台兼容生态持续完善，为模型规模化落地提供了多元国产算力支撑。

2.1 寒武纪

寒武纪是国内领先的人工智能芯片研发设计企业，专注于 AI 核心芯片的研发、设计与销售，产品覆盖云服务器、边缘计算及终端设备，主要布局云端、边缘产品线与 IP 授权及软件业务。

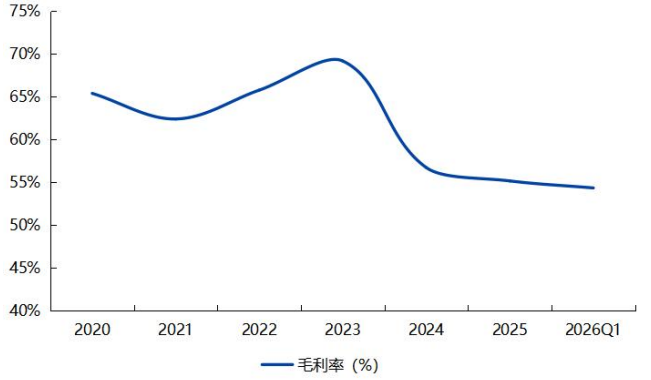
受人工智能行业算力需求的持续攀升，2025 年公司实现营业收入 64.97 亿元，同比+453.21%。值得注意的是，2026Q1 公司实现营收 28.85 亿元，同比+159.56%。2025 年公司毛利率达 55.15%，同比去年下降 1.56%，整体保持稳定。

图表 12：2020-2026Q1 寒武纪营业收入及同比



资料来源：公司公告，爱建证券研究所

图表 13：2020-2026Q1 寒武纪毛利率情况



资料来源：公司公告，爱建证券研究所

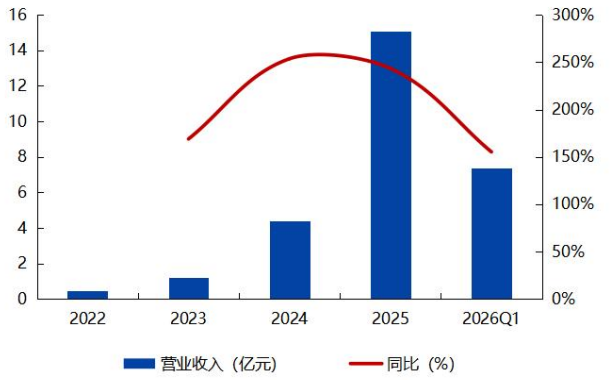
公司持续加强研发投入，聚焦人工智能芯片产品研发，持续强化产品核心竞争力。2025 年公司研发投入 11.69 亿元，研发投入占营收比例 17.99%。在硬件端，公司新一代智能处理器微架构和指令集持续研发。同时公司持续迭代训练平台和推理软件。

针对 DeepSeek-V4 全新架构，寒武纪已完成即时适配。公司依托自研 Torch-MLU-Ops 算子库对 Compressor、mHC 核心模块专项加速，通过 BangC 语言深度优化稀疏/压缩 Attention、GroupGemm 等热点算子，充分释放硬件底层算力。在 vLLM 推理框架中全面支持 5D 混合并行、通信计算并行、低精度量化及 PD 分离部署，在时延约束下实现最优 token 吞吐，显著提升端到端推理效率；同时借助 MLU 访存与排序加速能力、高互联带宽，有效加速稀疏 attention 等结构，降低通信开销，最大化分布式推理资源利用率。

2.2 摩尔线程

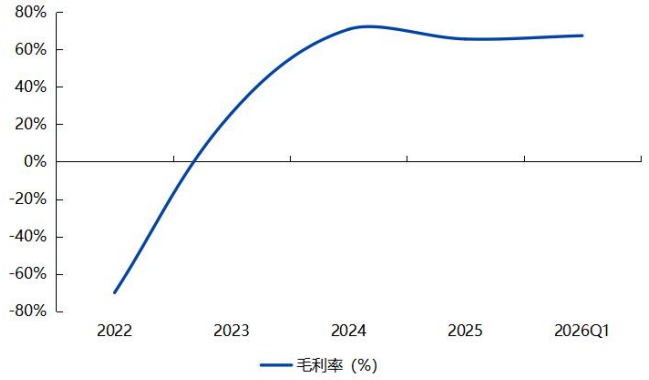
摩尔线程成立于 2020 年，以全功能 GPU 为核心，致力于向全球提供加速计算的基础设施和一站式解决方案，为各行各业的数智化转型提供强大的 AI 计算支持。2025 年，摩尔线程实现营业收入 15.06 亿元(同比+243.37%)，2022-2025 年复合增长率达 219.91%；2025 年公司毛利率为 65.57%，同比下滑 5.14 pct。

图表 14：2022-2026 Q1 摩尔线程营业收入及同比



资料来源：公司公告，爱建证券研究所

图表 15：2022-2026 Q1 摩尔线程毛利率情况



资料来源：公司公告，爱建证券研究所

摩尔线程产品矩阵呈现多元化特征，覆盖多场景需求。公司基于自主研发的 MUSA 架构，实现了单芯片架构同时支持 AI 计算加速、图形渲染、物理仿真和科学计算、超高清视频编解码的技术突破。截至 2025 年年报，公司主要产品线包括云端产品线、边缘与终端产品线。

图表 16：摩尔线程产品线梳理

资料来源：公司公告，爱建证券研究所

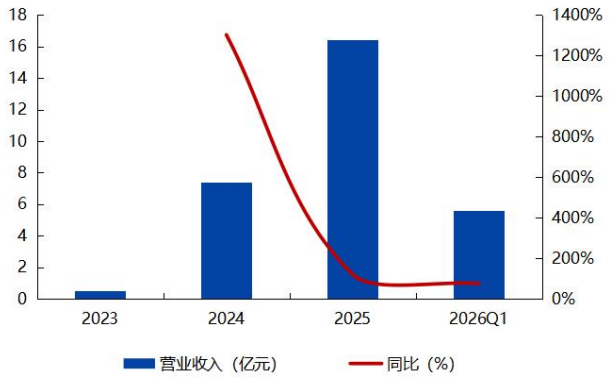
2026 年 4 月 24 日，公司携手智源众智及 FlagOS 社区宣布，已在旗舰级 AI 训推一体 GPU——MTT S5000 上完成 DeepSeek-V4-Flash 大模型的发布当日极速适配，全面支持全量核心算子深度优化与部署。2026 年 5 月，公司进一步宣布，依托 MTT S5000+MUSA 软件栈+SGLang 框架，已完成 DeepSeek-V4 完整运行验证，具备从底层硬件、热点算子支持到端到端部署验证的全链路工程化适配能力。

值得注意的是，DeepSeek-V4 首次采用“FP4+FP8”混合精度策略，而国内主流 AI 芯片仍以 BF16 为主。MTT S5000 作为国内率先原生支持 FP8 的全功能 GPU，搭载硬件级 FP8 Tensor Core，相比 BF16/FP16 可降低 50%显存带宽压力，理论算力翻倍，高效适配模型前沿精度设计。

2.3 沐曦股份

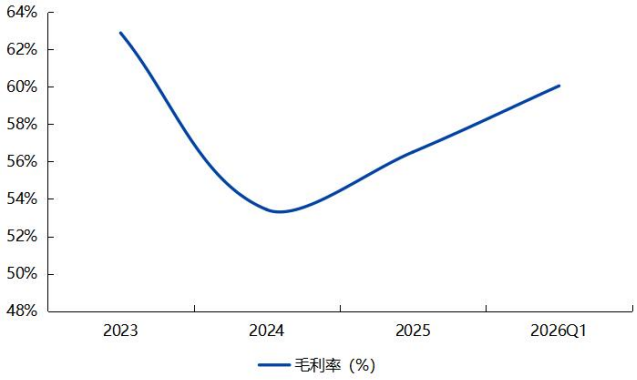
沐曦集成电路（上海）股份有限公司于 2020 年在上海成立，聚焦异构计算领域，打造全栈 GPU 芯片及解决方案。2025 年，沐曦股份实现营业收入 16.44 亿元（同比+121.26%）；2025 年公司毛利率为 56.51%，同比提高 3.08 pct。

图表 17：2022-2026 Q1 沐曦股份营业收入及同比



资料来源：公司公告，爱建证券研究所

图表 18：2022-2026 Q1 沐曦股份毛利率情况



资料来源：公司公告，爱建证券研究所

公司旗下拥有曦思 N 系列（智算推理）、曦云 C 系列（通用计算）、曦彩 G 系列（图形渲染）等产品。就各产品性能来看，曦思 N 系列聚焦云端智算推理，凭借高带宽内存、强劲视频编解码能力及大显存与高算力，支撑大规模数据推理和超高清视频流处理，配套完整软件栈可高效部署智算任务；曦云 C 系列为自研架构通用 GPU，具备高精度算力与 MetaXLink 片间互联技术，支持多 GPU 无缝协同，依托自研 MXMACA 软件栈覆盖智算研发、数据分析等复杂场景；曦彩 G 系列专攻图形渲染加速，自研架构拥有出色的图形图像渲染与视频处理能力，作为国产全功能显卡兼容主 GPU 生态，可为云游戏、元宇宙提供高画质低延迟算力支撑。

图表 19：沐曦股份主要产品分类

芯片名	介绍	产品特点	应用场景
曦思 N 系列	曦思 N 系列是面向云端应用的智算推理产品，采用高带宽内存，提供强大的算力和领先的视频编解码能力。	高速显存； 澎湃算力； 领先的视频处理能力； 完整的软件栈	智算
曦云 C 系列	曦云 C 系列通用 GPU(GPGPU)芯片是针对智算及通用计算的完美解决方案	自主知识产权 GPGPU; 超强高精度及混合精度算力 片间互联 MetaXLink 无缝连接多 GPU 系统； 自主软件栈 MXMACA 提供全面生态解决方案	智算； 数据分析
曦彩 G 系列	曦彩 G 系列 GPU 是针对图形渲染加速的解决方案，沐曦自主知识产权架构提供卓越的图形图像渲染与视频处理能力	卓越的图形图像渲染与视频处理能力； 国产全功能显卡； 采取沐曦自主知识产权； 兼容主流 GPU 生态的完整软件栈	云游戏与元宇宙

资料来源：公司官网，爱建证券研究所

2026 年 4 月 24 日，沐曦股份宣布携手 FlagOS 已完成对 DeepSeek 最新开源的 DeepSeek-V4-Flash 模型的 Day 0 适配。双方通过高性能通用大模型算子库 FlagGems、独立并行策略、FP4→BF16 全链路精度转换三大核心技术，实现了该模型在多类主流国产芯片上的全量适配与开箱即用的推理部署方案。

同日，沐曦股份还联合上海人工智能实验室 KernelSwift 智能算子迁移系统，率先完成 DeepSeek-V4 核心算子的 Day 0 适配。公司实测数据显示，算子平均通过率约 80%；在 21 个核心算子上性能较 TileLang 提升 1.2 倍以上，国产芯片端平均正确性 75%+、推理加速 3.4 倍，人工修正后可达 100%正确性，显著缩短适配周期。

2.4 海光信息

海光信息是国内领先的高端处理器设计企业，坚持 CPU+DCU 双芯协同发展战略，持续推进产品迭代与技术升级。

公司核心产品包括海光通用处理器（CPU）和海光协处理器（DCU）。海光 CPU 全面兼容 x86 指令集及全球主流操作系统与应用软件，已实现电信、金融、互联网、教育、交通等关键行业的规模化应用；海光 DCU 采用 GPGPU 通用计算架构，广泛应用于大数据处理、人工智能、大模型训练与推理、高性能科学计算等领域。

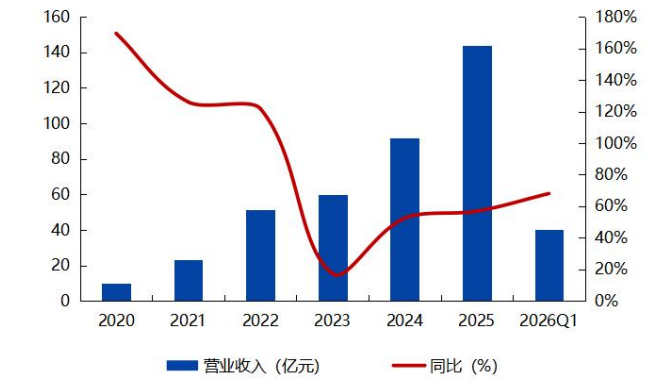
图表 20：海光信息主要产品

3000 系列 CPU	5000 系列 CPU	7000 系列 CPU	8000 系列 DCU
个人工作站、工控设备等产品	通用服务器	高端通用服务器、先进计算系统	科学计算、人工智能、大数据处理

资料来源：海光信息 2025 年报，爱建证券研究所

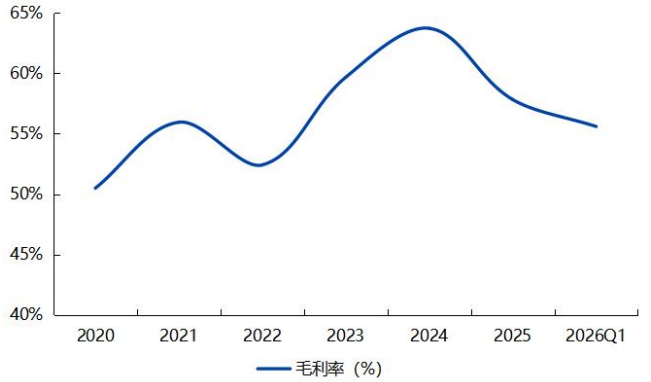
2025 年公司实现营业收入 143.7 亿元，同比+56.92%。2025 年公司毛利率为 57.83%。归母净利润 25.45 亿元。2025 年公司综合毛利率为 57.83%，同比-5.89 PCT，主要受上游晶圆代工及封装测试成本上涨、以及高增长的 DCU 业务占比提升影响。但公司持续深化与头部大模型厂商的全栈式生态绑定，通过“芯片+模型”联合创新加速 DCU 业务商业化落地。

图表 21：2020-2026 Q1 海光信息营业收入及同比



资料来源：公司公告，爱建证券研究所

图表 22：2020-2026 Q1 海光信息毛利率情况



资料来源：公司公告，爱建证券研究所

2026 年 4 月 24 日，海光信息宣布其海光 DCU 已同步完成对 DeepSeek-V4 的“Day 0”极速适配，实现了“模型发布—芯片适配—产业落地”的高效闭环，为全球开发者与企业客户提供即取即用的部署方案。

本次适配过程中，公司依托自研的 DTK 异构计算平台与 DAS 人工智能基础软件系统，海光 DCU 对 DeepSeek-V4 模型实现了全栈深度调优，再次达成业界领先的计算效率。其中，DTK 凭借完整成熟的计算库全面覆盖训练、推理、AI for Science 等全场景，为模型提供坚实的软件生态支撑；DAS 则集成了超 2000 个算子，支持 PyTorch、TensorFlow、vLLM、SGLang 等超过 100 个主流 AI 框架组件，通过算子调优、编译优化、通算融合等多重技术手段极致释放 DCU 算力，大幅提升模型的微调与推理性能。

3. 风险提示

1）技术迭代不及预期风险：大模型算法架构迭代、长上下文推理优化进度放缓，同时国产 AI 芯片与大模型适配兼容、生态完善进程慢于预期，制约技术落地与产业规模化发展。

2）商业化落地放缓风险：大模型行业应用场景渗透不及预期，商业化落地节奏放缓。

3）行业竞争加剧风险：全球大模型厂商密集发布新品，行业竞争加剧或引发 API 及服务价格下行，压缩盈利空间。

爱建证券
AIJIAN SECURITIES

爱建证券有限责任公司

上海市浦东新区前滩大道 199 弄 5 号
电话：021-32229888　　传真：021-68728700　　服务热线：956021
邮政编码：200124　　邮箱：ajzq@ajzq.com　　网址：http://www.ajzq.com

评级说明

投资建议的评级标准

报告中投资建议所涉及的评级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 个月内的相对市场表现，也即：以报告发布日后的 6 个月内的公司股价（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。其中：A 股市场：沪深 300 指数（000300.SH）；新三板市场：三板成指（899001.CSI）（针对协议转让标的）或三板做市指数（899002.CSI）（针对做市转让标的）；北交所市场：北证 50 指数（899050. BJ）；香港市场：恒生指数（HIS.HI）；美国市场：标普 500 指数（SPX.GI）或纳斯达克指数（IXIC.GI）。

股票评级

买入	相对同期相关证券市场代表性指数涨幅大于 15%
增持	相对同期相关证券市场代表性指数涨幅在 5%～15%之间
持有	相对同期相关证券市场代表性指数涨幅在-5%～5%之间
卖出	相对同期相关证券市场代表性指数涨幅小于-5%

行业评级

强于大市	相对表现优于同期相关证券市场代表性指数
中性	相对表现与同期相关证券市场代表性指数持平
弱于大市	相对表现弱于同期相关证券市场代表性指数

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告采用信息和数据来自公开、合规渠道，所表述的观点均准确地反映了我们对标的证券和发行人的独立看法。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法可能存在局限性，请谨慎参考。

法律主体声明

本报告由爱建证券有限责任公司（以下统称为“爱建证券”）证券研究所制作，爱建证券具备中国证监会批复的证券投资咨询业务资格，接受中国证监会监管。

本报告是机密的，仅供我们的签约客户使用，爱建证券不因收件人收到本报告而视其为爱建证券的签约客户。本报告中的信息均来源于我们认为可靠的已公开资料，但爱建证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供签约客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见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。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，爱建证券及其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测后续可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，爱建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。

版权声明