

分析师: 唐月
登记编码: S0730512030001
tangyue@ccnew.com 021-50586737

DeepSeek 近期成果分析及 V4 影响力预测

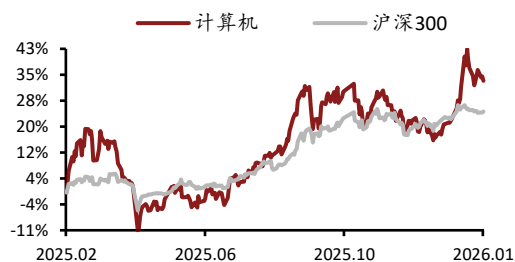
——计算机行业分析报告

证券研究报告-行业分析报告

强于大市(维持)

计算机相对沪深 300 指数表现

发布日期: 2026 年 01 月 29 日



资料来源: 中原证券研究所, 聚源

相关报告

《计算机行业月报: 手机端 AI 应用加速, DeepSeek 将加大预训练规模》 2025-12-08
《人工智能专题: DeepSeek 的稀疏注意力机制给 AI 产业释放更大的发展潜能》 2025-10-16
《人工智能专题: 后 R1 时代, DeepSeek 发展的三大阶段》 2025-10-14
《人工智能专题: 三大要素齐发力, AI 应用步入全面加速期》 2025-03-07

联系人: 李智

电话: 0371-65585629

地址: 郑州郑东新区商务外环路 10 号 18 楼

地址: 上海浦东新区世纪大道 1788 号 T1 座 22 楼

投资要点:

- The Information 报道, DeepSeek 将在 2026 年 2 月中旬推出新一代旗舰 AI 模型 DeepSeek V4, V4 编码能力超越 Claude 和 GPT 系列。我们认为 V4 对标预期中在 2025 年 5 月发布的 R2 模型。
- 2026 年 1 月 12 日, DeepSeek 论文聚焦分配的稀疏化方案, 引入了名为“Engram”的条件记忆模块, 明显改善了模型性能, 成为 MOE 的重要补充。同时通过对计算与内存的解耦, 缓解了当前 GPU 内存受限的困境, 有望大幅缓解国产 AI 芯片厂商 HBM 被卡脖子的境况。
- 2026 年 1 月 1 日, DeepSeek 论文提出了名为 mHC 的新网络架构, 解决信息的流动。mHC 架构是建立在此前字节发布的 HC 基础上, 重点改进了 ResNet 架构信息通道宽度受限、增加的冗余和内存占用的问题。在 MoE 模型上, mHC 使得模型训练的收敛速度提升了约 1.8 倍。
- DeepSeek 在模型 DeepSeek-OCR 和 DeepSeek-OCR2 中, 将视觉作为文本压缩媒介的新方法, 将文本以图片的方式进行输入, 可以极大减少输入所需要的 token 数量, 解决长文本输入问题。
- 2026 年 1 月 4 日, DeepSeek 更新了 R1 论文, 从 22 页增加到了 86 页, 让业界对 V4 的发布充满了更多的期待。根据论文的成本数据, R1 的总训练成本为 586 万美元, 远低于顶级模型训练动辄千万美元的门槛, 其中预训练和后训练分别占总成本的 95% 和 5%。
- 结合 DeepSeek 当前的研究成果, 我们给出 V4 潜在的创新方向的猜想和影响力预测:
 - (1) 模型成本的降低, 有望较大缓解地目前国内缺芯的状况。
 - (2) 继续开源路线, 同时模型能力超越闭源模型。有望深刻改变海外 AI 产业的发展格局, 利好 AI 应用的落地。
 - (3) 基于独立于 transformer 的全新架构。这意味着 V4 将带来里程碑意义的技术突破, 开启大模型发展的新范式, 帮助人类更快地通往 AGI。
 - (4) 与国产芯片进一步的深度融合, 可能部分或全部采用国产芯片进行训练, 利好国产算力的生态建设

风险提示: 国际局势的不确定性; 海外 AI 产业竞争格局变化带来市场调整风险。

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

内容目录

1. DeepSeek 最新进展.....	3
2. 稀疏化分配方案的引入, 将带来性能提升和计算与内存的解耦	5
3. 模型层间信息传输方式的底层架构创新	7
4. 长文本输入: 用图像承载文本信息, 实现高效压缩	8
5. R1 论文从 22 页更新至 86 页, 更加公开透明	8
6. V4 的潜在创新猜想和影响力预测	9
7. 风险提示	10

图表目录

图 1: TileLang 简介	4
图 2: 模型 Scaling 的新范式	4
图 3: Engram 引入后的模型架构	5
图 4: 稀疏性分配与 Engram 的内存拓展对模型性能的影响	6
图 5: 密集型模型、MoE 模型、Engram 模型之间的预训练性能比较	6
图 6: mHC、HC、ResNet 架构对比图	7
图 7: DeepSeek-OCR 在压缩比和性能上的表现	8
图 8: DeepSeek-R1 的训练流程	9
图 9: 大模型智能水平随时间增长的情况	10
表 1: DeepSeek 的主要模型发布情况	3

1. DeepSeek 最新进展

根据 The Information 报道, DeepSeek 将在 2026 年 2 月中旬推出新一代旗舰 AI 模型 DeepSeek V4, V4 编码能力超越 Claude 和 GPT 系列。

继 2025 年春节期间推出 DeepSeek-R1 以后, 预期在 2025 年 5 月推出的 R2 一直是市场关注的焦点, 但是 R1 在 5 月仅进行一次小版本更新看齐头部模型性能。接着在后续 V3.1 和 V3.2 的更新中, DeepSeek 不再采取基础模型加推理模型这种模式 (V3 和 R1), 而是直接推出两者结合的混合模型, 因而 V4 就已经对标预期中的 R2 模型。

表 1: DeepSeek 的主要模型发布情况

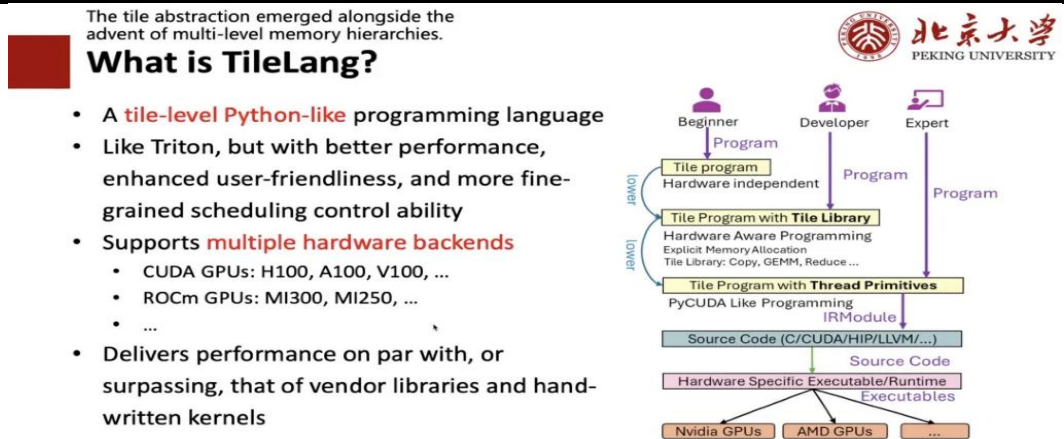
时间	模型	架构	参数	模型类型	备注
2023. 11. 29	DeepSeek LLM 67B		670 亿	语言模型	开源, 对标 LLaMA2 70B
2024. 1. 11	DeepSeek-MoE	MoE	1450 亿		开源, 国内首个 MoE 大模型, 有 2B、16B、145B 三个尺寸
2024. 5. 6	DeepSeek-V2	MoE	2360 亿	语言模型	开源, 性能比肩 GPT-4 Turbo, 价格为其 1/70
2024. 9. 5	DeepSeek V2. 5	MoE		语言模型	开源, 由 DeepSeek Coder V2 和 DeepSeek V2 Chat 两个模型合并升级而来
2024. 11. 20	DeepSeek-R1-Lite	MoE		推理模型	发布了对标 o1-preview 的预览版, 并开放思维链输出功能
2024. 12. 26	DeepSeek-V3	MoE	6710 亿	语言模型	开源, 性能比肩 GPT-4o
2025. 1. 20	DeepSeek-R1	MoE	6710 亿	推理模型	开源, 性能比肩 OpenAI o1 正式版, 价格约为其 1/30, 并开放思维链输出功能, 64K 上下文
2025. 3. 24	DeepSeek-V3-0324	MoE	6600 亿	语言模型	开源, 性能超过 GPT-4. 5
2025. 5. 28	DeepSeek-R1-0528	MoE	6850 亿	推理模型	开源, 上下文 128K, 以 DeepSeek V3 Base 为基座, 性能比肩 o3 与 Gemini-2. 5-Pro
2025. 8. 21	DeepSeek-V3. 1	MoE	6850 亿	混合模型	开源, 上下文 128K, 更高思考效率, 更强 Agent 能力
2025. 9. 22	DeepSeek-V3. 1-Terminus	MoE	6850 亿	混合模型	提升了 DeepSeek-V3. 1 的语言一致性和 Agent 能力
2025. 9. 29	DeepSeek-V3. 2-Exp	MoE	6850 亿	混合模型	开源, 实验性版本, 引入 DeepSeek Sparse Attention 稀疏注意力机制, API 大幅降价 (超 50%), 发布当日寒武纪和华为昇腾实现 day 0 适配和优化
2025. 12. 1	DeepSeek-V3. 2	MoE	6850 亿	混合模型	上下文 128K, 输出长度大幅降低
2025. 12. 1	DeepSeek-V3. 2-Special			混合模型	是 DeepSeek-V3. 2 的长思考增强版, 追求极致推理能力, 上下文 128K。

资料来源: DeepSeek, 中原证券研究所

此前我们在专题报告中分析了 DeepSeek 在 R1 之后进行的重要发展动向, 包括:

(1) 适配国产芯片: 2025 年 8 月 DeepSeek 发布了 V3.1, 其采用 UE8M0 FP8 缩放格式训练, 面向即将发布的下一代国产芯片而设计。9 月 DeepSeek 发布 V3.2-Exp, 发布当天华为昇腾和寒武纪宣布完成对其的零日适配, V3.2-Exp 还同时开源 TileLang 和 CUDA 两个版本的算子, 极大地改善了 CUDA 带来的从英伟达 GPU 切换到国产芯片的生态壁垒难题 (详见《人工智能专题: 后 R1 时代, DeepSeek 发展的三大阶段》)。

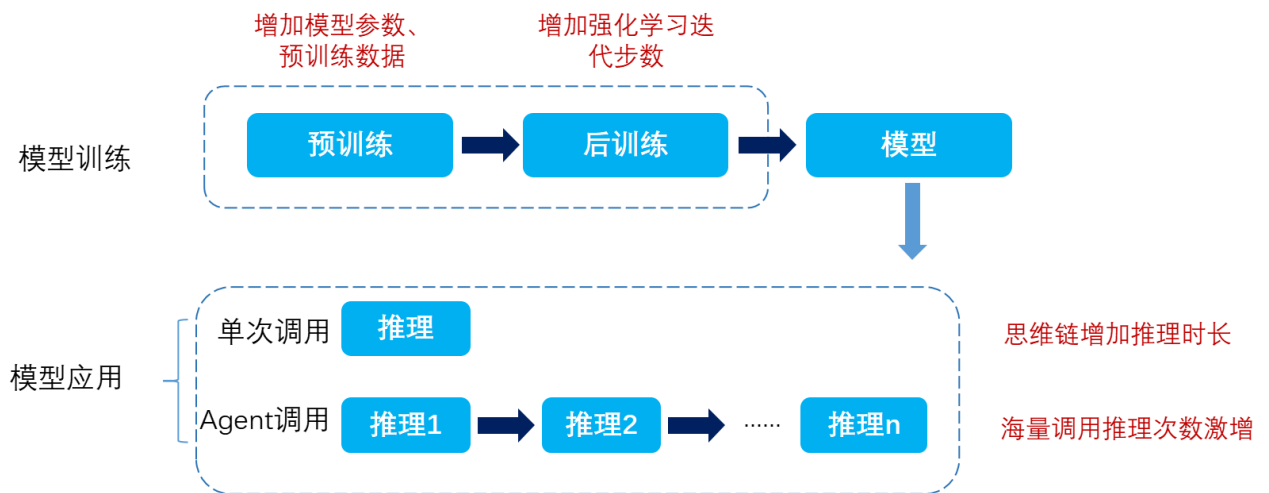
图 1: TileLang 简介



资料来源: 北京大学, 中原证券研究所

(2) 在新的注意力机制上的研究: 包括了在 2025 年 2 月在论文《Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention》中提出了原生稀疏注意力 (NSA), 在 2025 年 9 月发布的 V3.2-Exp 中引入了 **DeepSeek 稀疏注意力机制 (DSA)**, 从而将稀疏注意力从推理阶段进一步拓展到了预训练阶段, 为模型计算效率的提升、模型上下文的拓展、后训练潜能的释放留下了更多的发展空间 (详见《人工智能专题: DeepSeek 的稀疏注意力机制给 AI 产业释放更大的发展潜力》)。

图 2: 模型 Scaling 的新范式



资料来源: 中原证券研究所

(3) 大模型发展路径的探索: 2025 年 12 月 1 日, DeepSeek 发布了 V3.2 和 V3.2-Speciale, 同时进行了更多关于模型发展路径及 Scaling 的有效性的探索。V3.2 并没有针对测试集的工具进行特殊训练, 是真实获得的泛化能力, 证明了强化学习有效性。同时在模型训练方面, V3.2 验证了扩大后训练强化学习带来的模型能力提升路径 (DeepSeek-V3.2 把相当于预训练成本 10% 以上的算力投入在了后训练的强化学习中), 同时 DeepSeek 也计划加大预训练规模来实现模型能力更大的提升。DeepSeek 通过在强化学习中使用合成数据, 带来 DeepSeek-V3.2 在 Tau2Bench、MCP-Mark 和 MCP-Universe 基准测试中的性能显著提升, 取得了在真实环境中进行强化学习所无法获得的能力, 证实了合成数据的发展潜力。(详见《计算机行业月报:

手机端 AI 应用加速, DeepSeek 将加大预训练规模》)

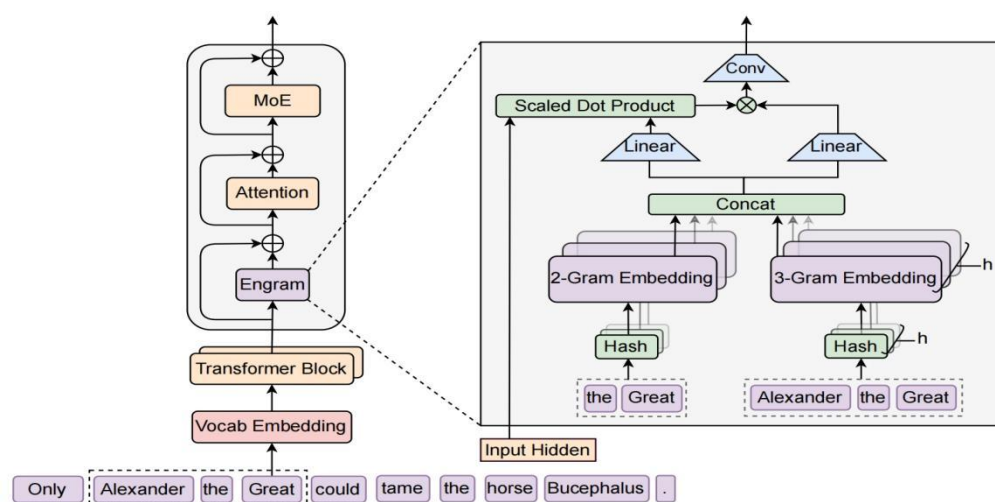
下面我们将就 DeepSeek 后续发布的 5 篇论文展开技术走向分析, 并对 V4 的潜在影响力进行预测。

2. 稀疏化分配方案的引入, 将带来性能提升和计算与内存的解耦

2026 年 1 月 12 日, DeepSeek 发布新论文《Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models》, 聚焦分配的稀疏化方案。

当前主流的大语言模型在面对知识类的问题时, 仍然需要动用宝贵的推理资源通过计算来模拟检索过程, 造成了大量的资源浪费。考虑到大模型的稀疏性不应该只体现在计算上, 还应该体现在分配上, DeepSeek 本次提出了“条件记忆”的新机制, 并引入了名为“Engram”的条件记忆模块, 依靠稀疏查找来检索固定知识的静态嵌入。

图 3: Engram 引入后的模型架构



资料来源: DeepSeek, 中原证券研究所

由于 Engram 是对文本进行哈希映射, 映射对象为规模可扩展的静态记忆表, 进行常数时间复杂度的知识检索, 查找复杂度与模型规模无关, 因而达到了稀疏的效果。实验表明, 当 20%-25% 的稀疏参数预算分配给 Engram 时 (剩余部分留给 MoE), 模型整体性能达到最佳。

图 4：稀疏性分配与 Engram 的内存拓展对模型性能的影响

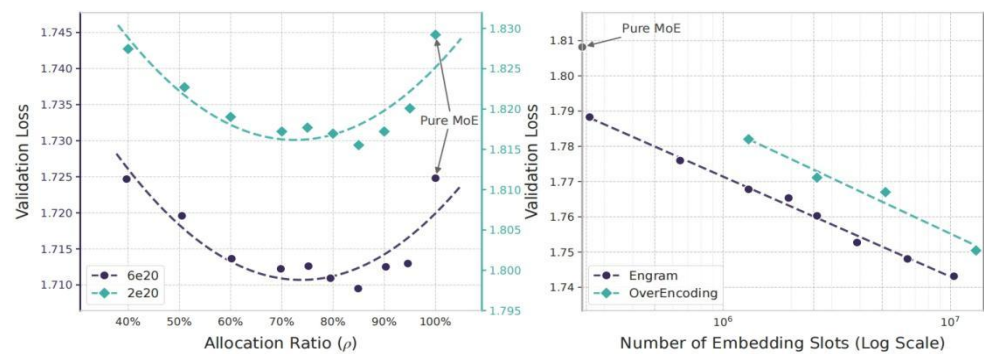


Figure 3 | **Sparsity allocation and Engram scaling.** Left: Validation loss across allocation ratios ρ . Two compute budgets are shown (2e20 and 6e20 FLOPs). Both regimes exhibit a U-shape, with hybrid allocation surpassing Pure MoE. Right: Scaling behavior in the infinite-memory regime. Validation loss exhibits a log-linear trend with respect to the number of embeddings.

资料来源：DeepSeek，中原证券研究所

DeepSeek 的实验数据研究表明，Engram 的引入可以实现更高的效率，成为 MoE 的理想补充。除了知识密集型任务，Engram 的引入还可以使得通用推理、代码、数学问题上获得显著的改进。DeepSeek 认为这主要源于 Engram 减轻了主干网络在早期重构静态知识的负担，从而增加了可用于复杂推理的有效深度。同时，Engram 架构还在长文本处理方面展现出了结构性优势，在处理需要极高精度的长文本任务时，具备更强的优越性。

图 5：密集型模型、MoE 模型、Engram 模型之间的预训练性能比较

	Benchmark (Metric)	# Shots	Dense-4B	MoE-27B	Engram-27B	Engram-40B
	# Total Params		4.1B	26.7B	26.7B	39.5B
	# Activated (w/o token embed)		3.8B	3.8B	3.8B	3.8B
	# Trained Tokens		262B	262B	262B	262B
	# Experts (shared + routed, top-k)		-	2 + 72 (top-6)	2 + 55 (top-6)	2 + 55 (top-6)
	# Engram Params		-	-	5.7B	18.5B
Language Modeling	Pile (loss)	-	2.091	1.960	1.950	1.942
	Validation Set (loss)	-	1.768	1.634	1.622	1.610
Knowledge & Reasoning	MMLU (Acc.)	5-shot	48.6	57.4	60.4	60.6
	MMLU-Redux (Acc.)	5-shot	50.7	60.6	64.0	64.5
	MMLU-Pro (Acc.)	5-shot	21.1	28.3	30.1	31.3
	CMMLU (Acc.)	5-shot	47.9	57.9	61.9	63.4
	C-Eval (Acc.)	5-shot	46.9	58.0	62.7	63.3
	AGIEval (Acc.)	0-shot	29.1	38.6	41.8	45.9
	ARC-Easy (Acc.)	25-shot	76.8	86.5	89.0	90.1
	ARC-Challenge (Acc.)	25-shot	59.3	70.1	73.8	76.4
	TriviaQA (EM)	5-shot	33.0	48.8	50.7	51.8
	TriviaQA-ZH (EM)	5-shot	62.8	74.8	76.3	77.9
	PopQA (EM)	15-shot	15.1	19.2	19.4	21.2
	CCPM (Acc.)	0-shot	72.2	79.6	87.1	87.7
	BBH (EM)	3-shot	42.8	50.9	55.9	57.5
	HellaSwag (Acc.)	0-shot	64.3	71.8	72.7	73.1
	PIQA (Acc.)	0-shot	63.8	71.9	73.5	76.5
	WinoGrande (Acc.)	5-shot	64.0	67.6	67.8	68.1
Reading Comprehension	DROP (F1)	1-shot	41.6	55.7	59.0	60.7
	RACE-Middle (Acc.)	5-shot	72.4	80.9	82.8	83.3
	RACE-High (Acc.)	5-shot	66.0	75.4	78.2	79.2
	C3 (Acc.)	0-shot	57.7	60.1	63.6	61.8
Code & Math	HumanEval (Pass@1)	0-shot	26.8	37.8	40.8	38.4
	MBPP (Pass@1)	3-shot	35.4	46.6	48.2	46.2
	CruxEval-i (EM)	0-shot	27.6	30.7	32.2	36.2
	CruxEval-o (EM)	0-shot	28.7	34.1	35.0	35.3
	GSM8K (EM)	8-shot	35.5	58.4	60.6	62.6
	MGSM (EM)	8-shot	27.0	46.8	49.4	52.4
	MATH (EM)	4-shot	15.2	28.3	30.7	30.6

资料来源：DeepSeek，中原证券研究所

除此以外, Engram 引入后, 将模型参数表卸载到主机内存并不会带来显著效率损失, 因而很大程度上缓解了当前 GPU 内存受限的困境, 为模型参数的进一步提升打开了通道。这意味着 Engram 实现了计算与内存的解耦, 为挂载 TB 级别的超大规模记忆库提供了可行路径, 有望大幅缓解国产 AI 芯片厂商 HBM 被卡脖子的境况。

3. 模型层间信息传输方式的底层架构创新

2026 年 1 月 1 日, DeepSeek 发布新论文《mHC: Manifold-Constrained Hyper-Connections》, 并提出了名为 mHC 的新网络架构, 解决信息的流动。

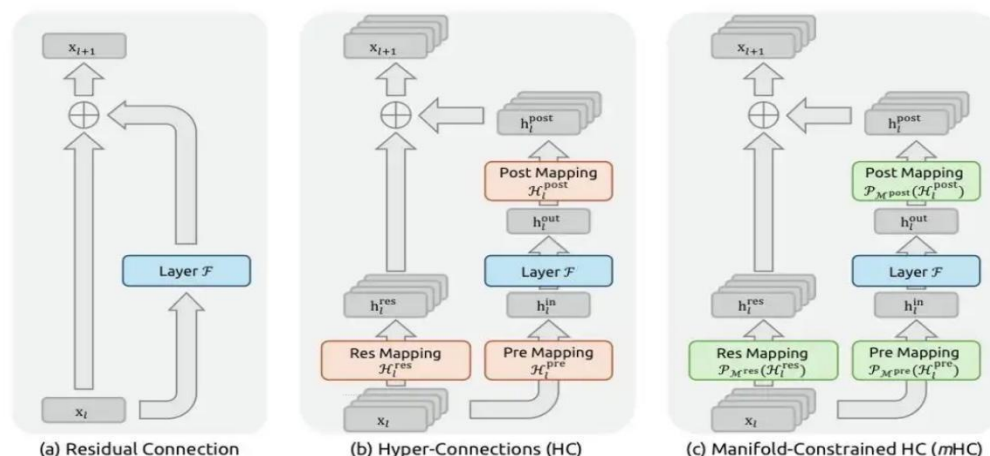
mHC 架构是建立在此前字节发布的 Hyper-Connections (HC) 基础上, 重点改进了 ResNet 架构。

2015 年微软亚洲研究院的何恺明等人提出 ResNet (残差网络) 架构, 让信息可以跳过某些层直接传递, 解决了此前神经网络越深、训练越困难的问题, 让上百层甚至上千层网络成为可能, 因而成为了 2017 年问世的 Transformer 的底层组件, 也是目前 AI 大模型所使用的主流技术架构。ResNet 虽然稳定, 但是信息通道宽度受限, 增加的计算冗余和内存占用, 随着模型参数的增加, 模型层数的增长, 带来的浪费就更加明显。

2024 年 9 月, 字节跳动发表了 Hyper-Connections (超连接), 提出将单一残差流拓展为多流并行架构, 但是 HC 在提升模型性能的同时, 信号被持续地放大, 将模型训练变得不稳定。随着模型的增大、层数的增多, HC 所展现出来的问题越严重, 因而只能用于小模型和理论推演中, 在大模型训练中无法有效应用。

为了解决 HC 的稳定性和可扩展问题, DeepSeek 推出了 mHC。mHC 引入了一种类似“加权平均”的思路, 由于凸组合的结果不会超过输入的最大值, 保证信号不会被无限放大。试验结果表明在 MoE 模型上, mHC 使得模型训练的收敛速度提升了约 1.8 倍。

图 6: mHC、HC、ResNet 架构对比图

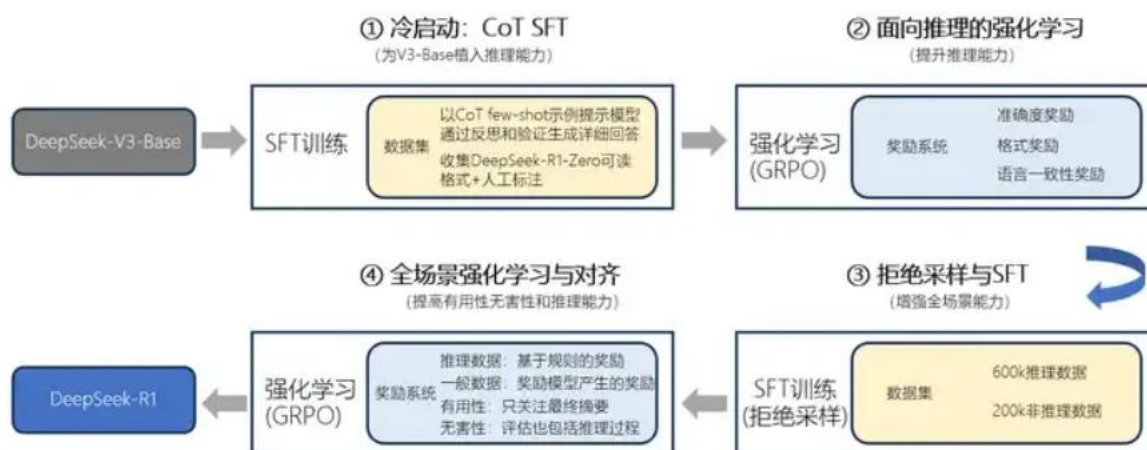


资料来源: DeepSeek, 中原证券研究所

消融实验、中间检查点、失败的尝试、R1 模型的不足等等。DeepSeek 更加公开透明，必然来自于对当前研究成果的自信，因而也让业界对 V4 的发布充满了更多的期待。

继此前披露了 DeepSeek-V3 的训练成本以后，本次论文进一步披露了 DeepSeek-R1 进行进一步训练的成本为 29.4 万美元。这意味着 R1 的总训练成本为 586 万美元，远低于顶级模型训练动辄千万美元的门槛，其中预训练和后训练分别占总成本的 95%和 5%。

图 8：DeepSeek-R1 的训练流程



资料来源：DeepSeek，中存算，中原证券研究所

6. V4 的潜在创新猜想和影响力预测

结合 DeepSeek 当前的研究成果，我们给出 V4 潜在的创新方向的猜想和影响力预测：

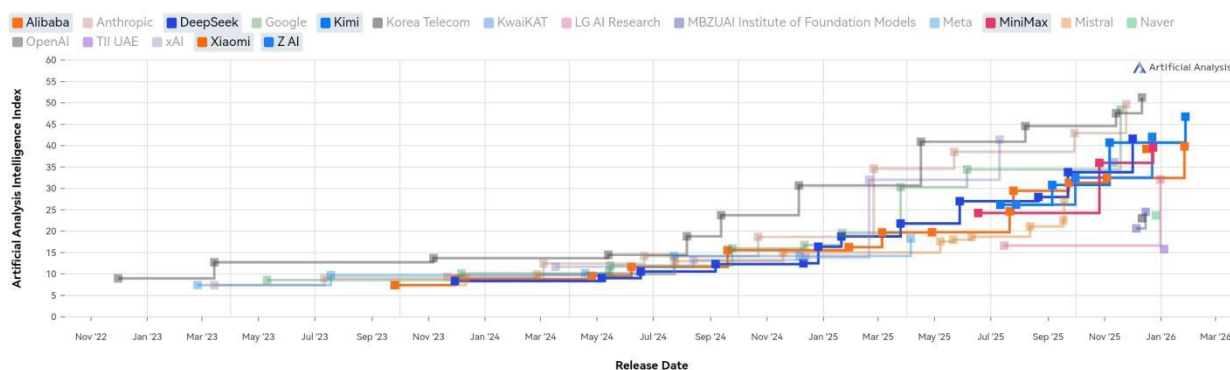
（1）模型成本的降低。

DeepSeek 在 2026 年新发布的论文中公开的 Engram 架构和 mHC 都有望较大改善模型的成本，因而对于新模型来说，成本有望大幅降低，有望较大地缓解目前国内缺芯的状况。

（2）继续开源路线，同时模型能力超越闭源模型。

以 DeepSeek 目前的行事风格来看，其仍然是开源路线最坚定的拥护者，有望继续开源路线。目前，国产厂商主要坚持开源路线，且模型性能较海外还有一定落后，但是成本控制上优势较为明显。我们认为 V4 如果能够实现能力较大超越，将对海外闭源模型厂商，特别是以 OpenAI、Anthropic 为代表的仅专注于大模型赛道的厂商形成较大的盈利冲击，并有望深刻改变海外 AI 产业的发展格局。

图 9：大模型智能水平随时间增长的情况



资料来源：Artificial Analysis，中原证券研究所（此处没有比较 DeepSeek-V3.2-Special 的模型能力）

V4 如果继续采用开源，其模型能力的提升将给下游 AI 应用厂商带来更大的助力，利好 AI 应用的落地。

（3）基于独立于 transformer 的全新架构。这意味着 V4 将带来里程碑意义的技术突破，开启大模型发展的新范式，帮助人类更快地通往 AGI。

2026 年 1 月 20 日，DeepSeek 在更新的 FlashMLA 代码库中，意外曝光了名为 Model 1 的新模型，还拥有了与 DeepSeek-V3.2 并列的独立文件。从 DeepSeek 的命名体系来看，Model 1 属于全新的模型体系，这或许意味着市场预期中的 V4 将采用全新的技术路径或基础架构。

（4）与国产芯片进一步的深度融合，可能部分或全部采用国产芯片进行训练。

考虑到 2025 年 DeepSeek 也已经实现与国产芯片的协同优化，2026 年 DeepSeek 也有望在国产适配方面获得更多的进展，利好国产算力的生态建设。

7. 风险提示

国际局势的不确定性；海外 AI 产业竞争格局变化带来市场调整风险。

行业投资评级

强于大市：未来 6 个月内行业指数相对沪深 300 涨幅 10% 以上；

同步大市：未来 6 个月内行业指数相对沪深 300 涨幅-10% 至 10% 之间；

弱于大市：未来 6 个月内行业指数相对沪深 300 跌幅 10% 以上。

公司投资评级

买入：未来 6 个月内公司相对沪深 300 涨幅 15% 以上；

增持：未来 6 个月内公司相对沪深 300 涨幅 5% 至 15%；

谨慎增持：未来 6 个月内公司相对沪深 300 涨幅-10% 至 5%；

减持：未来 6 个月内公司相对沪深 300 涨幅-15% 至 -10%；

卖出：未来 6 个月内公司相对沪深 300 跌幅 15% 以上。

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分析师执业资格，本人任职符合监管机构相关合规要求。本人基于认真审慎的职业态度、专业严谨的研究方法与分析逻辑，独立、客观的制作本报告。本报告准确的反映了本人的研究观点，本人对报告内容和观点负责，保证报告信息来源合法合规。

重要声明

中原证券股份有限公司具备证券投资咨询业务资格。本报告由中原证券股份有限公司（以下简称“本公司”）制作并仅向本公司客户发布，本公司不会因任何机构或个人接收到本报告而视其为本公司的当然客户。

本报告中的信息均来源于已公开的资料，本公司对这些信息的准确性及完整性不作任何保证，也不保证所含的信息不会发生任何变更。本报告中的推测、预测、评估、建议均为报告发布日的判断，本报告中的证券或投资标的价格、价值及投资带来的收益可能会波动，过往的业绩表现也不应当作为未来证券或投资标的表现的依据和担保。报告中的信息或所表达的意见并不构成所述证券买卖的出价或征价。本报告所含观点和建议并未考虑投资者的具体投资目标、财务状况以及特殊需求，任何时候不应视为对特定投资者关于特定证券或投资标的的推荐。

本报告具有专业性，仅供专业投资者和合格投资者参考。根据《证券期货投资者适当性管理办法》相关规定，本报告作为资讯类服务属于低风险（R1）等级，普通投资者应在投资顾问指导下谨慎使用。

本报告版权归本公司所有，未经本公司书面授权，任何机构、个人不得刊载、转发本报告或本报告任何部分，不得以任何侵犯本公司版权的其他方式使用。未经授权的刊载、转发，本公司不承担任何刊载、转发责任。获得本公司书面授权的刊载、转发、引用，须在本公司允许的范围内使用，并注明报告出处、发布人、发布日期，提示使用本报告的风险。

若本公司客户（以下简称“该客户”）向第三方发送本报告，则由该客户独自为其发送行为负责，提醒通过该种途径获得本报告的投资者注意，本公司不对通过该种途径获得本报告所引起的任何损失承担任何责任。

特别声明

在合法合规的前提下，本公司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、财务顾问等各种服务。本公司资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告意见或者建议不一致的投资决策。投资者应当考虑到潜在的利益冲突，勿将本报告作为投资或者其他决定的唯一信赖依据。