DeepSeek-V4 点评

多层面技术提升训练规模，超长上下文进入普惠时代

优于大市

◆ 行业研究·行业快评 ◆ 计算机 ◆ 投资评级：优于大市（维持）

证券分析师： 熊莉 021-61761067 xiongli1@guosen.com.cn 执证编码：S0980519030002
联系人： 侯睿 hourui3@guosen.com.cn

事项：

2026 年 4 月 24 日，DeepSeek 最新模型 V4 预览版本正式上线并同步开源，包括两个 MoE 语言模型——DeepSeek-V4-Pro（总参数量 1.6 万亿，其中激活参数为 490 亿）和 DeepSeek-V4-Flash（总参数量 2840 亿，其中激活参数为 130 亿），两者均支持长达一百万 token 的上下文长度，DeepSeek-V4 系列在架构与优化方面进行了多项关键升级。

国信计算机观点：DeepSeek-V4 已经具备接近全球第一梯队的综合能力，同时通过极具竞争力的价格体系，打开了大规模企业级 AI Agent 落地的商业空间。其在长上下文训练中的优化为基础模型的进步提供了全新的方向，后续百万上下文有望成为前沿模型的标配。同时，DeepSeek-V4 在国产算力方面积极适配，有望推动整体国产算力需求增长。**风险提示：**下游需求不及预期、AI 应用落地不及预期、硬件技术落地进程不及预期、宏观经济波动等。

评论：

◆ **模型层：**

2026 年 4 月 24 日，DeepSeek 最新模型 V4 预览版本正式上线并同步开源，包括两个 MoE 语言模型——DeepSeek-V4-Pro（总参数量 1.6 万亿，其中激活参数为 490 亿）和 DeepSeek-V4-Flash（总参数量 2840 亿，其中激活参数为 130 亿），两者均支持长达一百万 token 的上下文长度，DeepSeek-V4 系列在架构与优化方面进行了多项关键升级：

1）混合注意力架构 CSA+HCA：不是继续沿用标准 dense attention，而是把注意力拆成两类，CSA 先把 KV 沿序列维压缩，再做稀疏选择；HCA 则用更激进的压缩，但保留 dense attention。两者交替使用，目标是同时兼顾局部依赖、全局检索能力和极端长序列下的成本控制。此设计不是单点优化，而是从 attention 结构层面重写了长上下文的成本函数，因此能把 1M context 真正做成系统级可运行方案。**在 100 万 token 场景下，V4-Pro 的单 token 推理 FLOPs 只有 DeepSeek-V3.2 的 27%，KV cache 只有 10%，V4-Flash 更低到 10% FLOPs 和 7% KV cache。**

2）mHC（Manifold-Constrained Hyper-Connections）：把残差连接从经验上有效变成数值上更稳定的可控结构。普通 Hyper-Connections 虽然能增强表达，但深层堆叠时容易数值不稳定；于是 V4 把残差映射矩阵约束到 doubly stochastic manifold 上，使其谱范数受限、残差传播变成 non-expansive，从而改善深层训练稳定性。

3）把 Muon optimizer 真正落到超大规模训练中：不是简单换了个优化器，而是把 Muon 作为大部分模块的主优化器，同时保留 AdamW 给 embedding、norm、head 等部分，再配合 hybrid Newton-Schulz orthogonalization 去提升收敛和稳定性。

4）FP4 量化训练（QAT）：DeepSeek 把 FP4 用在两个位置，一是 MoE expert weights，二是 CSA 里 indexer 的 QK 路径；同时还把 index scores 从 FP32 压到 BF16，使 top-k selector 达到 2×加速，同时保留 99.7% 的 KV 召回率。同时，FP4 到 FP8 的 dequantization 在其设定下可以无损地复用现有 FP8 训练框架，这使

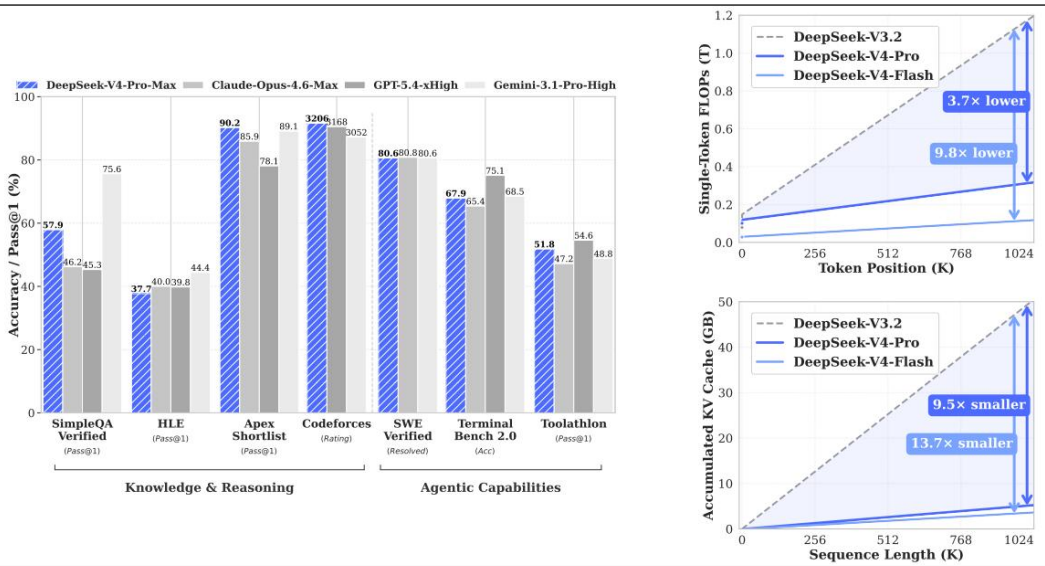
得低比特方案不只是理论节省显存，而是真正进入了可训练、可 rollout、可部署的主干流程。

5）后训练专家独立训练+on-policy distillation 统一蒸馏：不是直接把一个通用模型拿去做混合 RL，而是先分别培养数学、代码、agent、instruction-following 等领域专家，再通过 on-policy distillation 把这些能力蒸馏回一个统一模型。设计的意义在于把专才能力最强和最终交付一个通用模型两个目标拆开做，兼顾 specialization 和 consolidation。

6）基础设施层面创新：MoE 中把通信、计算、访存做成单融合 kernel；更细粒度的 expert wave 调度来隐藏通信开销。这个 MoE 通信—计算融合方案不只理论可行，DeepSeek 在 NVIDIA GPUs 和 HUAWEI Ascend NPUs 平台上都对细粒度 EP 调度方案完成了验证，该方案在通用推理负载下可实现 1.50-1.73 倍的加速，在时延敏感型场景（如 RL 采样迭代、高速智能体服务）中，最高加速比可达 1.96 倍。

DeepSeek-V4 使用超 32 万亿 token 数据对模型进行预训练，并辅以完整的后训练流程，以释放并增强模型能力。其中，DeepSeek-V4-Pro-Max（DeepSeek-V4-Pro 的最高推理强度模式）在核心任务上重新定义了开源模型 SOTA，性能超越其前代模型。**DeepSeek-V4 系列在长上下文场景下具有极高的效率，在百万 token 的上下文设置中，DeepSeek-V4-Pro 的单 token 推理计算量（FLOPs）仅为 DeepSeek-V3.2 的 27%，KV 缓存仅为其 10%。这使得模型能够常规性支持百万 token 的上下文，从而让长时序任务更加可行。**

图1：DeepSeek-V4 通过更少计算量实现开源 SOTA



资料来源：《DeepSeek-V4:Towards Highly Efficient Million-Token Context Intelligence》，国信证券经济研究所整理

◆ 混合注意力架构 CSA+HCA：

在普通 Transformer 里，假设用户现在生成第 100 万个 token，理论上要去关注前面所有 token。因为每个 token 都要和前面大量 token 做匹配，序列越长，计算量和 KV cache 都会指数级增长，因此标准 attention 的二次复杂度是超长上下文和长推理过程的核心瓶颈。

CSA（Compressed Sparse Attention，压缩稀疏注意力）主要有以下效果：

1）Compressed（压缩 KV）：假设原来有 100 万个 token，每个 token 都有自己的 KV。CSA 不再保留 100 万个独立 KV，而是每隔一组 token 把它们压缩成一个“压缩 KV 条目”。CSA 会把每 m 个 token 的 KV cache 压缩成一个 entry，从而把序列长度压缩到原来的 1/m；

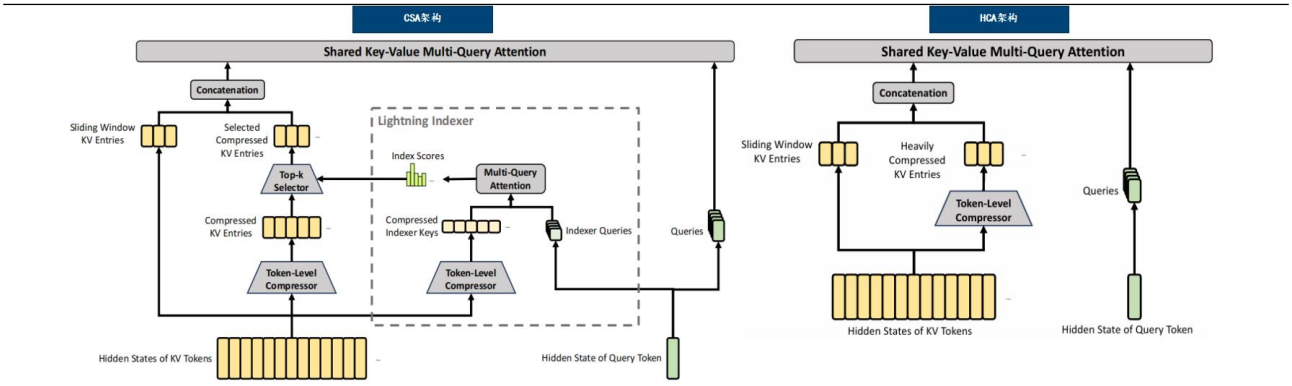
2）Sparse（稀疏选择）：压缩后当前 token 不是把所有摘要块都看一遍，而是通过一个轻量级 indexer，先判断哪些压缩块最相关，然后只选 top-k 个块进入真正的 attention。用 indexer 给压缩 KV 块打分，再用 top-k selector 选择一部分压缩 KV 进入后续核心 attention，即 Lightning Indexer for Sparse Selection。

但是单纯压缩会丢失细节，模型如果只看压缩摘要，可能看不到精细的局部关系。DeepSeek 额外加了一个 sliding window attention（滑动窗口注意力），它会让当前 token 仍然直接看最近的一小段未压缩 token。语言模型生成时，最近几个 token 往往特别重要，例如 The capital of France is ...，生成下一个词时最近的“France is”非常关键，不能只靠压缩摘要。

HCA（Heavily Compressed Attention，重度压缩注意力），会把更大范围的 token 压成一个 KV 块。HCA 使用更大的压缩率 m′，且 m′ 远大于 CSA 里的 m，即 m′ ≫m。HCA 把每 m′ 个 token 的 KV 压缩成一个 entry，但不使用稀疏 attention。

CSA 压缩更少，而且还会选 top-k 相关块，所以保留的信息更细一些，适合找具体内容，但它需要 indexer 和 top-k 选择，机制更复杂。HCA 极度省算力、省 KV cache，而且可以保留很长范围的全局信息，但压缩更多，细节丢得更多。因此 DeepSeek-V4 采用混合结构，有些层用 CSA 负责更细粒度地找相关内容、有些层用 HCA 负责低成本保留超长全局信息。HCA 用于看整体框架，CSA 用于查找相关框架内需要内容的摘要，有效降低了长上下文下的计算成本，提升了运行速度。

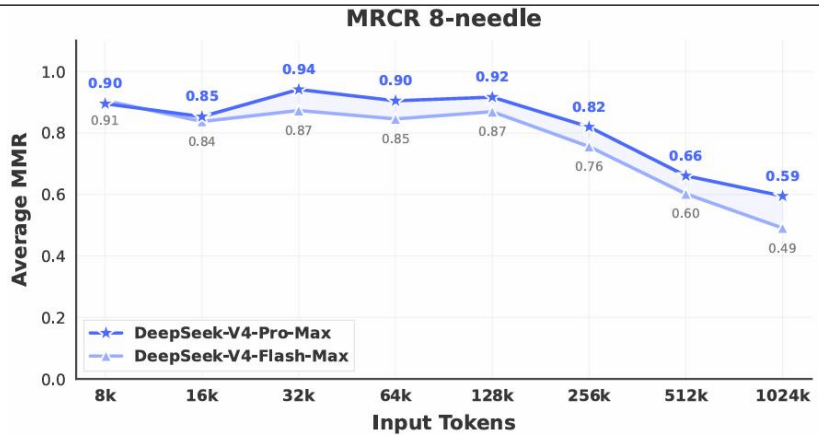
图2：CSA 以及 HCA 架构



资料来源：《DeepSeek-V4:Towards Highly Efficient Million-Token Context Intelligence》，国信证券经济研究所整理

通过压缩的架构，模型不再受困于计算资源的限制，在长上下文任务方面展现出极强的可用性。在 Codeforces 这种编程竞赛中，V4-Pro-Max 以 3206 分的 Rating 首平了 OpenAI 的 GPT-5.4 等闭源顶流。在用于衡量模型的上下文内检索能力的 MRCR 任务中，DeepSeek-V4-Pro 表现优于 Gemini-3.1-Pro，但仍落后于 Claude Opus 4.6。在 128K 上下文窗口以内，模型的检索表现保持高度稳定。虽然超过 128K 后性能下降开始变得明显，但在 100 万 token 场景下，相较于闭源和开源竞品，DeepSeek-V4-Pro 的检索能力仍然相当强。

图3：DeepSeek-V4 长上下文任务的表现



资料来源：《DeepSeek-V4:Towards Highly Efficient Million-Token Context Intelligence》，国信证券经济研究所整理

◆ **新型残差连接架构 mHC：**

大模型层数很深时，每一层都在不断改写信息，如果层与层之间的信息传递太随意，信号可能越传越乱、越传越多，训练就不稳定。Transformer 里每一层不是直接把输入扔掉，然后只保留这一层算出来的新结果。而通常是下一层输入 = 原来的输入 + 当前层算出来的新信息（x_{l+1} = x_l + F_l(x_l)），如果模型很深，信息一层一层传下去容易丢失。Residual connection 给信息提供了传递路径，即使一层没计算好，原始信息也还能传到后面。

普通 residual 默认每一层的信息传递方式都为保留旧信息 + 加入新信息，但是超大模型有时需要多保留一点旧信息 + 少加一点新信息、不同信息内容重新混合或让前面某些层的信息绕过中间层传得更远，此时普通 residual 只有一条信息流（x_l）的模式难以满足模型需求。

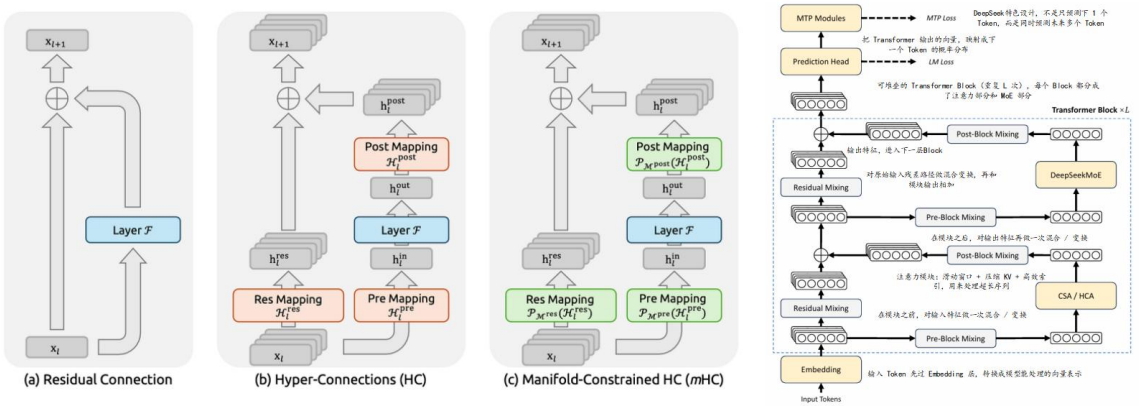
Hyper-Connections 会把 residual stream 扩展成多条信息流，HC 会把 residual stream 的宽度从 d 扩展到 n_hc × d，不是改变 Transformer 内部 attention/MoE 的 hidden size，而是在层与层之间额外维护多个 residual 状态，模型不再只有一条信息传递路径，而是可以在多个 residual 通道之间动态混合。但普通 HC 容易不稳定，如果有多条信息流，每一层还可以自由混合这些信息流，层数一深，信号就可能不断放大。这样就会出现数值爆炸，训练不稳定，影响 HC 的规模化使用。

mHC 全称 Manifold-Constrained Hyper-Connections（流形约束的超连接），HC 的核心是多条 residual 信息流之间可以混合。这个混合靠一个矩阵完成，论文里叫 B_l，在 mHC 中 B_l 被约束成一种特殊矩阵 doubly stochastic matrix（双随机矩阵）。双随机矩阵指所有元素都不能是负数、每一行加起来等于 1、每一列加起来也等于 1 的矩阵，只在重新分配信息，不会凭空把信息无限放大。

mHC 被放在 Transformer block 之间，用来加强传统 residual connections。整个 block 仍然包括 CSA/HCA attention 和 DeepSeekMoE，但层与层之间的信息传递方式被 mHC 改造了，每一层 residual 信息流的混合方式可以根据当前输入稍微调整，有三个主要映射：A_l 读信息（决定从多条 residual 流中取哪些信息送进当前层）、B_l 混合旧信息（决定 residual 流之间怎么混合）、C_l 写入新信息（决定当前层输出的新信息怎么写回 residual 流）。

普通 HC 的公式是：X_{l+1} = B_l X_l + C_l F_l(A_l X_l)：下一层的多路 residual 状态 = 对旧 residual 状态做稳定混合 + 把当前 Transformer 层算出的新信息写进去。mHC 是多份旧信息经过稳定混合 + 新信息按规则写回多个 residual 通道，它维护了多条 residual 通路，不同层的信息可以更灵活地组合，模型能力更强，同时 residual mixing matrix 被约束成双随机矩阵，不容易把信号无限放大，模型训练更稳定。**对于 DeepSeek-V4 这种超大 MoE 模型来说，模型规模很大、上下文很长、训练 token 很多，任何数值不稳定都会被放大，所以 mHC 的作用是给整个深层模型提供更稳定的信息传递结构。**

图4：DeepSeek-V4 长上下文任务的表现



资料来源：《DeepSeek-V4:Towards Highly Efficient Million-Token Context Intelligence》，国信证券经济研究所整理

◆ **Muon optimizer：**

Muon optimizer 的作用是让模型每一步更新参数时，不只是沿着梯度方向乱冲，而是先把更新方向整理得更规整、更稳定，再去更新参数，大模型训练收敛更快。大模型训练时，本质上是在不断调整参数，在模型回答错误时，训练系统会算出一个 loss，也就是错误程度，进而对模型参数进行调整，怎么调参数的规则即 optimizer（优化器）。

AdamW 为训练常用的优化器，不只看当前梯度，还会参考过去的梯度趋势，并且给不同参数自动调整步长。但对于超大模型，尤其是矩阵参数很多的模型，AdamW 主要还是按每个参数来处理更新，而 Muon 更关注整个矩阵作为一个整体，更新方向是不是规整。

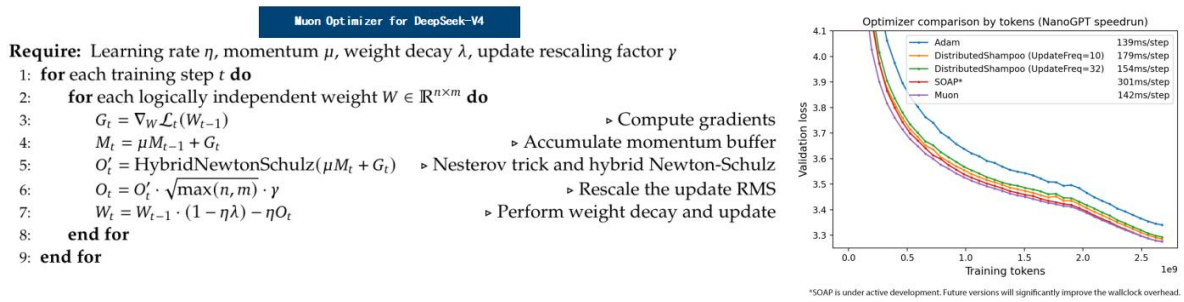
大模型里的很多参数都是矩阵，比如 MoE 里的专家权重矩阵。训练时，梯度也通常是一个矩阵，如果直接按梯度去更新，可能某些方向特别强，某些方向特别弱。Muon 在真正更新参数前，先对这个更新矩阵做一次正交化/规整化，即 orthogonalization，将模型整理成更均衡、更稳定的更新方向。

Muon 主要分为以下步骤：1）计算梯度（G_t = 当前这一步的梯度）；2）累积 momentum（M_t = μ M_{t-1} + G_t），如果过去几步都在往同一个方向走，说明这个方向具有效果，将保持惯性；3）用 Nesterov trick 加上当前梯度趋势（μM_t + G_t），不只看现在的位置，还预判一下按当前惯性继续走会怎么样；4）做 Hybrid Newton-Schulz（O'_t = HybridNewtonSchulz(...)），把更新矩阵做近似正交化，让它更规整；5）重新缩放更新大小（O_t = O'_t × sqrt(max(n,m)) × γ），正交化之后，更新矩阵的大小可能被改变，所以要把它缩放到合适的尺度；6）真正更新权重（W_t = W_{t-1} × (1 − η λ) − ηO_t），包含防止参数越来越大（权重衰减）、沿着优化后的方向推进（按学习率更新）。

Muon 里的正交化不是直接做 SVD（奇异值分解，把一个矩阵强行变成正交矩阵、做正交归一化），因为大模型 Transformer 的矩阵维度大，每一步训练迭代都对海量权重做 SVD，训练计算量指数级增长，显存、算力崩溃，导致无法训练。DeepSeek 用的是 Newton-Schulz iterations，近似把矩阵变成更正交的形式，用多轮快速计算，快速把奇异值（矩阵在各个方向上的缩放强度）推近 1。

DeepSeek-V4 对大部分模块使用 Muon，但对 embedding、prediction head、mHC 的 static biases 和 gating factors、RMSNorm 权重仍然使用 AdamW。因为 Muon 更适合处理大块的矩阵权重，例如 MoE 专家权重矩阵，这些参数更适合做正交化更新。但有些参数比较敏感，这些参数如果用 Muon 会导致不稳定，因此 DeepSeek-V4 采用了混合优化策略。V4-Pro 有 1.6T 参数、每次激活 49B 参数；V4-Flash 有 284B 参数、每次激活 13B 参数，且训练 token 超 32T，还要逐步扩展到 1M 上下文，这种训练存在显著稳定性挑战，而 Muon 帮助大模型收敛可能更快、训练过程更稳定，从而解决 trillion-parameter MoE 训练存在的挑战。

图5：Muon 以更低成本体现出更好训练效果



资料来源：《DeepSeek-V4:Towards Highly Efficient Million-Token Context Intelligence》，国信证券经济研究所整理

◆ **FP4 量化感知训练（QAT）：**

DeepSeek-V4 不是等模型训练完以后再临时压缩成 FP4，而是在后训练阶段就让模型提前适应 FP4 这种低精度。这样模型部署时真的用 FP4，也不容易掉性能，同时还能省显存、提速度。FP4 的核心优势在于一个数字只用 4 bit 表示，显存占用更小、数据搬运更少、计算更快，但损失计算精度。量化是把高精度数字变成低精度数字，通常用于加快推理，但是量化同样会带来计算误差，导致模型在部署时性能下降。

QAT 全称是 Quantization-Aware Training（量化感知训练），旨在训练的时候就模拟量化，让模型提前适应低精度误差。普通训练后量化是先用高精度训练好模型，训练完后压缩成低精度，会导致较大精度损失。QAT 是训练/后训练过程中就加入 FP4 量化模拟，使模型适应误差，部署时性能更稳。DeepSeek-V4 在 post-training 阶段引入 QAT，让模型适应量化带来的精度下降，从而在部署时实现推理加速和显存节省。

DeepSeek-V4 在 MoE 专家权重和 CSA 里面负责检索相关 KV 块的 Query-Key 路径中使用了 FP4，目的在于，DeepSeek-V4 是 MoE 模型，每次处理 token 时，不是所有专家都启动，而是只激活一部分专家。虽然每次只激活一部分专家，但所有专家的权重都要存在显存或存储系统里，每个专家都有大量参数，因此非常占用空间。MoE expert weights 是 GPU memory occupancy 的主要来源之一，所以非常适合用 FP4 压缩，显存占用显著下降。

CSA 会先把长上下文压缩成很多 KV 块，然后用一个 indexer 找当前 token 最应该看哪些压缩块，依照当前 query 和很多 compressed key 做匹配，给每个压缩块打分选 top-k 个最相关的块，这个匹配过程就是 QK 路径。在 100 万 token 上下文里，即使 KV 已经被压缩，候选块仍然很多。Indexer 要频繁做 QK 分数计算，所以 DeepSeek-V4 把 CSA indexer 的 QK activations 用 FP4 来缓存、加载、相乘，加速长上下文场景下的 attention score computation。

Indexer 算完 QK 后,会得到很多分数,DeepSeek 把这些 index scores 从 FP32 降到 BF16,使 top-k selector 达到 2×speedup，同时还能保留 99.7%的 KV entries recall，即基本不影响计算结果。对 MoE expert weights 训练时 optimizer 维护的是 FP32 master weights，即模型真正的高精度母版权重还是 FP32。之后 FP32 master weights 量化成 FP4，再反量化成 FP8，用 FP8 做计算。因为当前训练框架和硬件生态对 FP8 支持更成熟。DeepSeek-V4 已经有现成的 FP8 training framework。如果 FP4 可以无损转回 FP8 参与计算，就能复用原来的 FP8 框架，不需要重写整套训练系统。

训练时有 forward 和 backward，forward 为模型往前算，得到输出和 loss，backward 根据 loss 反向传播，更新参数。DeepSeek 在 backward pass 里，梯度相对于 forward pass 中同样的 FP8 weights 计算的，然后直接传回 FP32 master weights,即 FP32 master weights 模拟量化成 FP4,之后反量化成 FP8 做 forward，backward 算梯度直接更新 FP32 master weights，既让模型感受到 FP4 的低精度误差，又保留 FP32 母版权重用于稳定更新。

在 inference 和 RL rollout 阶段，因为不涉及 backward pass，所以直接使用真实 FP4 quantized weights。即训练时因为要反向传播，所以还要保留 FP32 master weights，但推理时不需要更新参数，所以可以直接使用 FP4 权重省显存、提速度。

◆ **后训练专家独立训练+on-policy distillation 统一蒸馏：**

DeepSeek-V4 不是直接把一个模型同时往数学、代码、Agent、写作、指令跟随等所有方向硬训，而是先分别训练出一批专家模型，然后再训练一个统一学生模型，把专家模型的能力集合到一个模型里。大模型训练通常先进行 pre-training（预训练），模型读大量文本、代码、论文、网页，学习语言、知识、推理的基础能力，此时模型无法稳定回答问题、不具备强工具使用能力等。此时进入 post-training（后训练），后训练的目标是把基础模型变成真正可用的助手模型，提升模型的回答问题、遵守指令、Agent 等能力。DeepSeek-V4 相比 DeepSeek-V3.2 把原来后训练的 mixed RL 阶段替换成了 On-Policy Distillation，简称 OPD。

传统做法是拿一个基础模型，然后把所有任务混在一起训练，训练数据里同时放数学、编程、Agent 工具调用等任务，然后对同一个模型做监督微调（SFT）或强化学习（RL）。这种模式不同能力之间难以达到平衡，比如数学 RL 希望模型多思考、写长推理链，而普通聊天希望模型自然、简洁，所有目标统一在模型身上会导致某些领域互相干扰。

DeepSeek-V4 首先会针对每个目标领域，独立训练一个 expert model，这里的 expert model 不是 MoE 结构里的专家，而是完整模型层面的专家，一个专门被训练到擅长某个领域的模型。在基础模型上用该领域高质量数据做 SFT，再用该领域 reward 做 RL，得到一个领域专家模型，在 RL 阶段 DeepSeek-V4 使用的是 GRPO（轻量化 RL 算法），用特定领域的 reward model 或成功标准来优化模型，使每个专家在自己的领域更强。

在后续阶段，DeepSeek-V4 把多个专家模型的能力合并回一个统一模型，这个过程就是 distillation（蒸馏），用 on-policy distillation 训练一个 unified model，这个统一模型作为 student，向多个 teacher models 学习，并优化 reverse KL loss。普通蒸馏是拿一批固定问题让 teacher models 回答，student 模仿答案，但 on-policy distillation 训练样本不是完全固定由老师生成，而是让学生模型自己在当前状态下生成，然后老师模型在学生模型当前生成轨迹上给 logits 分布，学生模型以此调整自己的概率分布。这样训练后的模型不仅兼顾了通用性，也在各个领域具备顶尖的性能。

图6：DeepSeek-V4-Pro-Max 与闭源、开源模型的综合对比

	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

	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

资料来源：《DeepSeek-V4:Towards Highly Efficient Million-Token Context Intelligence》，国信证券经济研究所整理

◆ 系统与基础设施层面的创新：

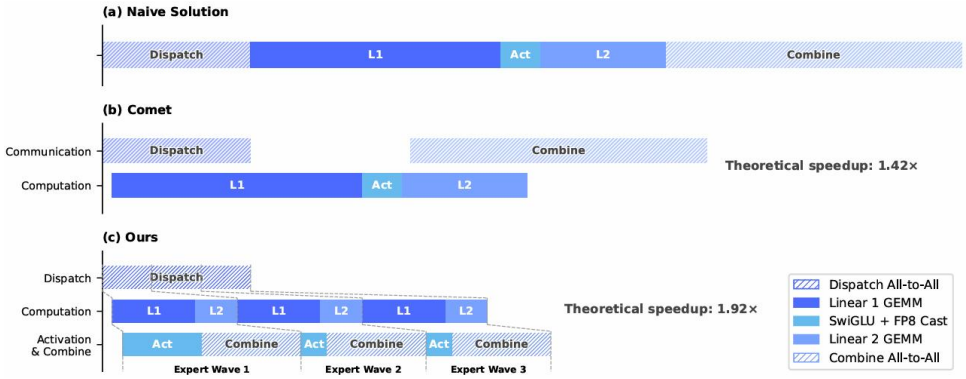
DeepSeek-V4 不只提出了一个百万 token 模型结构，同时重做了底层算子、通信方式、训练框架、推理缓存、低精度计算和 Agent/RL 服务系统。确保论文里的 1M token、MoE、高速推理、长上下文 RL 不止局限于理论，而是能够真正落地。

MoE 模型里有很多专家，每个 token 进来后 router 会决定它该交给哪些专家处理。如果专家分布在不同 GPU 上，系统就要先把 token 发送到对应 GPU（Dispatch），专家算完后把结果发回来进行 Combine。普通做法中是一步一步做，Dispatch（把 token 发到对应专家所在设备）、Linear-1（专家第一段矩阵计算）、Activation（激活函数，比如 SwiGLU）、Linear-2（专家第二段矩阵计算）、Combine（专家结果汇总）。此时，通信的时候计算单元空闲，计算的时候网络闲置，GPU/NPU 利用率不高。DeepSeek-V4 把这些过程融合成一个更大的 pipeline kernel，让通信、计算、访存尽量同时发生。

DeepSeek-V4 优化的核心在于 expert wave，如果 MoE 层有很多专家，普通做法是等所有专家的数据都通信完，再统一开始计算。DeepSeek-V4 把专家分成多个 waves，即专家分组。在计算过程中，只要 Wave 1

的通信完成，就立刻开始计算 Wave 1，此时 Wave 2 的数据可以继续传输，Wave 0 的结果可以继续发回。在稳定状态下，当前 wave 的计算、下一个 wave 的 token 传输、已完成专家的结果发送可以同时进行。**这个 MoE 通信—计算融合方案不只理论可行，DeepSeek 在 NVIDIA GPUs 和 HUAWEI Ascend NPUs 平台上都对细粒度 EP 调度方案完成了验证，该方案在通用推理负载下可实现 1.50-1.73 倍的加速，在时延敏感型场景（如 RL 采样迭代、高速智能体服务）中，最高加速比可达 1.96 倍。**

图7：MegaMoE 与其他 MoE 优化的比较

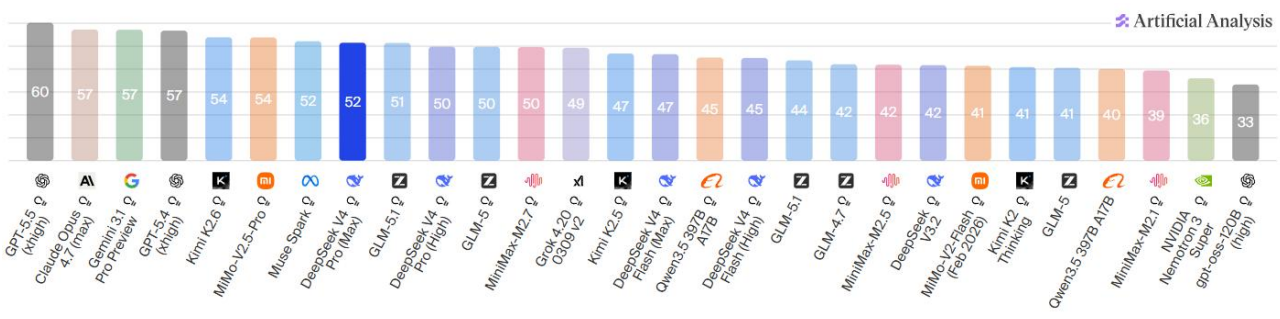


资料来源：《DeepSeek-V4:Towards Highly Efficient Million-Token Context Intelligence》，国信证券经济研究所整理

◆ 模型表现与国产算力适配：

DeepSeek-V4 性能表现大幅提升，在智能测评表现中名列前茅。Artificial Analysis 智能指数纳入了对模型能力的 10 项主流评测，包括 GDPval-AA（工作任务评测）、Terminal-Bench Hard（终端操作智能体评测）、SciCode（科学计算代码评测）、AA-LCR（长上下文推理评测）、Humanity's Last Exam（前沿学术能力测评）等。DeepSeek-V4 Pro Max 在 Artificial Analysis Intelligence Index 中得分 52 分，已经进入全球最强模型阵营，测评表现超越 GLM、MiniMax 等模型，在开源体系中具备较强竞争力，但距离榜首 GPT-5.5 的 60 分相比仍有明显差距。

图8：DeepSeek-V4 智能测评结果

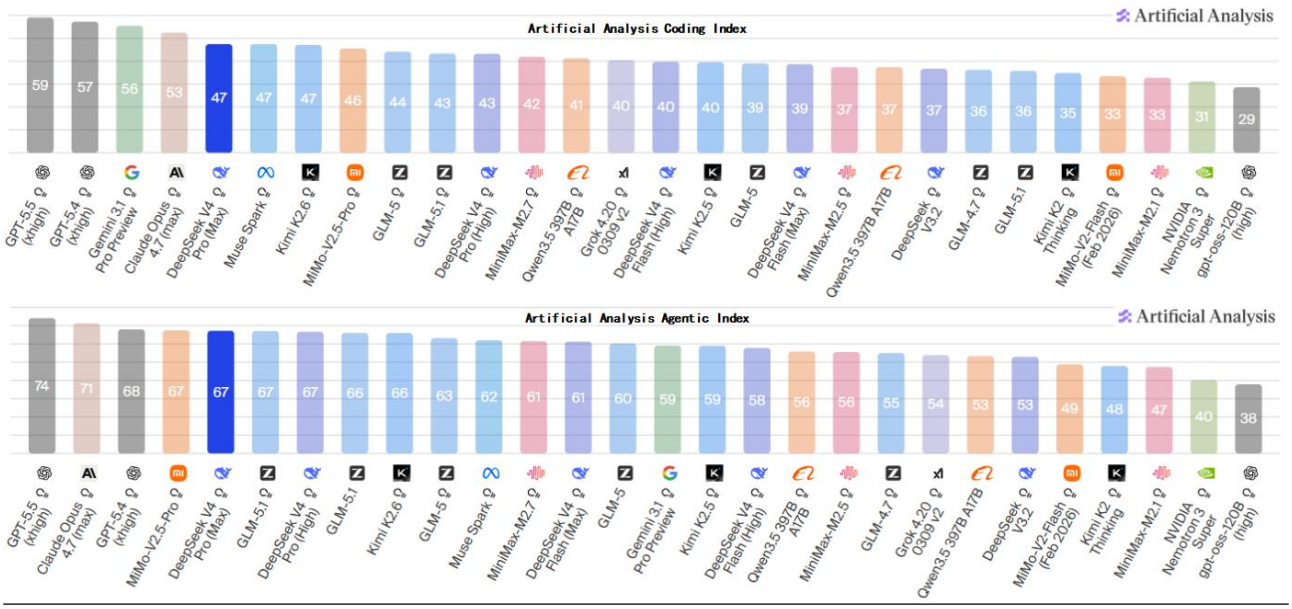


资料来源：Artificial Analysis，国信证券经济研究所整理

在当前主流应用场景测评中，DeepSeek-V4 达到全球顶尖表现。在 Artificial Analysis Coding Index

中，DeepSeek-V4 Pro Max 得分为 47，这个指数由 Terminal-Bench Hard 和 SciCode 构成，不仅考察传统代码生成，还考察终端任务执行、科学计算代码和复杂工程问题。DeepSeek-V4 在这些任务中取得开源 SOTA 表现，仅次于 Claude 最新模型。DeepSeek-V4 在 Artificial Analysis Agentic Index 中表现更突出，得分为 67，仅次于 GPT-5.5 Thinking 74、Claude Opus 4.7 Max 71、GPT-5.4 Thinking 68，其在智能体任务表现上处于全球第一梯队，尤其是在真实工作任务和多轮交互任务中表现较强。

图9：DeepSeek-V4 编程、智能体测评结果



资料来源：Artificial Analysis，国信证券经济研究所整理

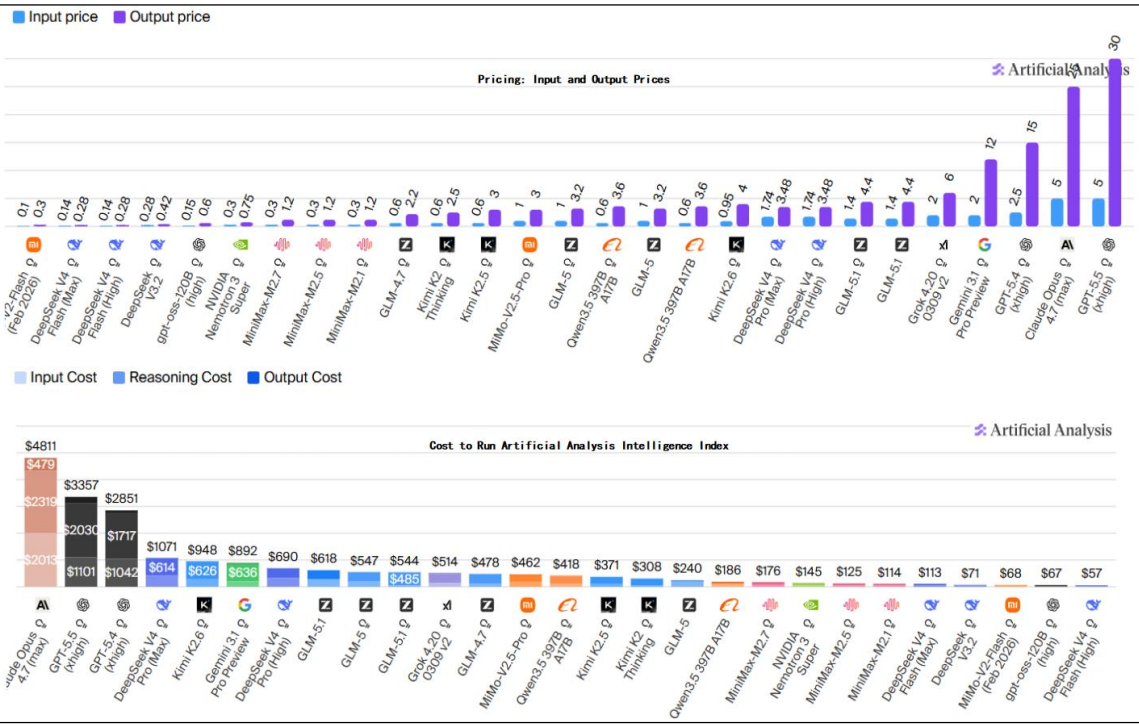
推理算力需求明显提升，有望刺激国产算力用量。DeepSeek-V4 的思维链明显拉长，整体消耗 token 量增多。在运行 Artificial Analysis Intelligence Index 测评中，DeepSeek-V4 Flash Max 消耗约 240M tokens，DeepSeek-V4 Pro Max 消耗约 190M tokens，均较之前版本有大幅提升。**说明 DeepSeek-V4 在高推理模式下属于明显的高计算投入型模型，有望大幅提升推理部署的算力资源用量。DeepSeek 在官方文档中表示受限于高端算力，当前 Pro 服务吞吐量有限，预计下半年 950 超节点批量上市后，Pro 的价格会大幅下调，有望从推理侧刺激国产算力在高端模型中的使用。**

图10：DeepSeek-V4 编程、智能体测评结果

资料来源：Artificial Analysis，国信证券经济研究所整理

同时，DeepSeek-V4 仍具备极高的性价比，从 API 输入/输出单价来看，DeepSeek-V4 Flash Max 以及 DeepSeek-V4 Flash High 每百万 tokens 调用价格约为输入 0.14 美元,输出 0.28 美元,仅为 DeepSeek-V3.2 的约 50%。与海外旗舰模型相比，DeepSeek-V4 系列的价格优势非常明显，调用价格仅为 OpenAI 和 Claude 旗舰模型的约 1/18 至 1/100，其中输入价格约为 GPT-5.5 Thinking 和 Claude Opus 4.7 Max 的 1/36，输出价格约为 GPT-5.5 Thinking 的 1/107、Claude Opus 4.7 Max 的 1/89。这一价格结构显著降低了高频调用和大规模部署的边际成本，使 DeepSeek-V4 系列在企业客服、批量内容生成、智能体和流程自动化等成本敏感场景中具备较强商业竞争力。在运行整套 Intelligence Index 后，DeepSeek-V4 Pro Max 的运行成本约为 1071 美元，显著低于 Claude Opus 4.7 Max 的 4811 美元、GPT-5.5 的 3357 美元和 GPT-5.4 的 2851 美元，模型在保持较高表现的同时做了极强的成本控制。

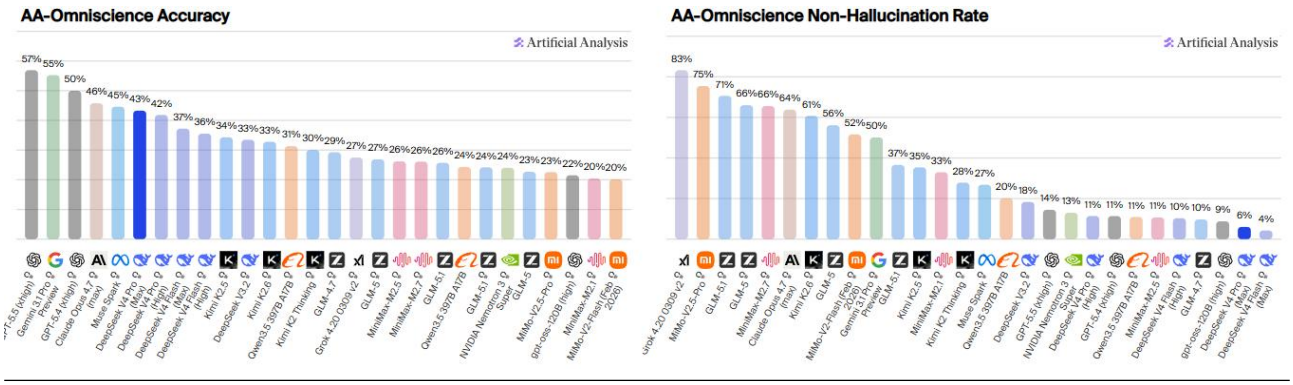
图11：DeepSeek-V4 成本表现



资料来源：Artificial Analysis，国信证券经济研究所整理

幻觉仍是当前模型存在的一大问题，在 AA-Omniscience 的两个拆分指标 Accuracy（准确率）、Non-Hallucination Rate（非幻觉率）测评中，DeepSeek-V4 表现与主流模型相差较大。准确率方面 DeepSeek-V4 系列整体处于顶尖水平，DeepSeek-V4 Pro Max 得分为 43%，次于 GPT-5.5 Thinking 57%、Gemini 3.1 Pro 55%、GPT-5.4 Thinking 50%。在非幻觉率方面，DeepSeek-V4 系列的非幻觉率偏低，DeepSeek-V4 Pro Max 的非幻觉率仅为 6%，较 Grok 83%、MiMo-V2.5-Pro 75%等有明显差距，这意味着在 AA-Omniscience 这类评测中，DeepSeek-V4 尤其是 Max 模式更倾向于尝试作答，而不是在不确定时拒绝回答。因此，DeepSeek-V4 在很多问题上可以给出正确答案，但面对知识边界之外或高度不确定的问题时，可能更容易生成看似合理但实际错误的答案。对于通用聊天、创意写作或轻量办公，这个问题可能不明显，但对于金融、法律、医疗、科研、投研等高准确性场景，就需要额外的检索增强、引用校验或人工复核机制。

图12：DeepSeek-V4 在 AA-Omniscience 测评中的表现



资料来源：Artificial Analysis，国信证券经济研究所整理

◆ 风险提示：

下游需求不及预期、AI 应用落地不及预期、硬件技术落地进程不及预期、宏观经济波动等。

相关研究报告：

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

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