2025

INDUSTRY RESEARCH REPORT

DeepSeek技术全景解析

重塑全球AI生态的中国力量

编制：智研咨询

目·录

PART 01

DeepSeek企业背景

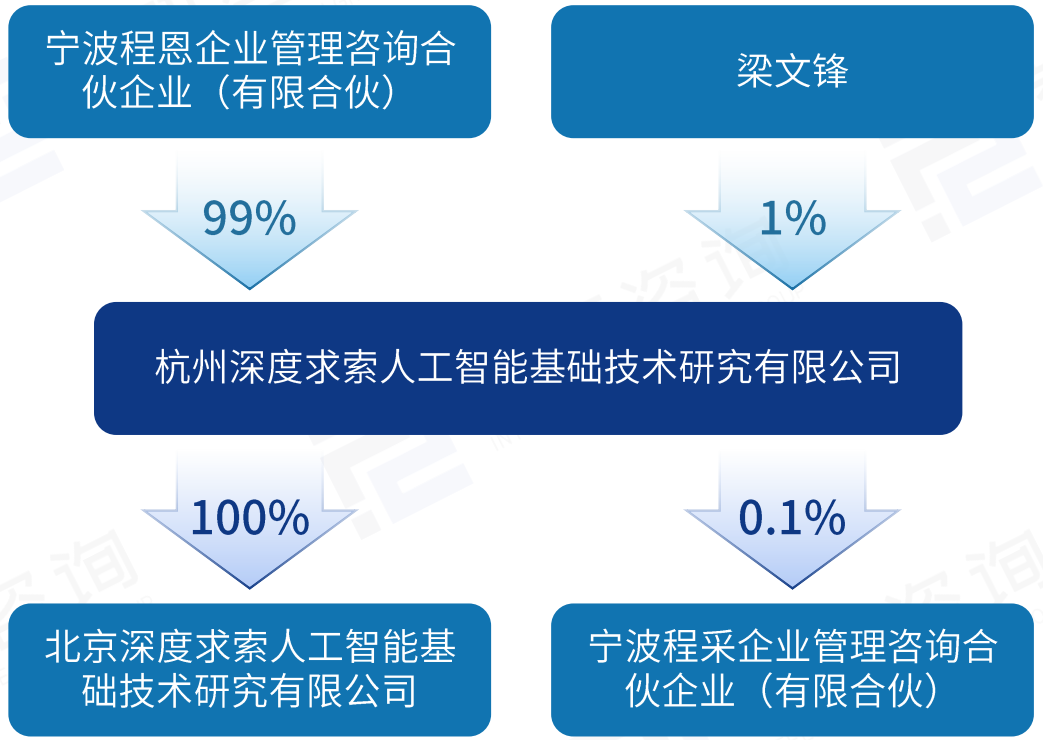
最全面的产业分析 ● 可预见的行业趋势

DeepSeek背靠资金实力雄厚的幻方量化

2025年1月，DeepSeek发布其最新开源模型DeepSeek R1，再度引发全球人工智能领域关注。DeepSeek，全称杭州深度求索人工智能基础技术研究有限公司，成立于2023年7月17日，一家创新型科技公司，专注于开发先进的大语言模型（LLM）和相关技术。DeepSeek背靠资金实力雄厚的幻方量化，DeepSeek创始人为梁文锋，梁文锋同时也是幻方量化的创始人，幻方量化是国内头部量化私募管理人，旗下有两家百亿量化私募，分别是2015年6月成立的浙江九章资产和2016年2月成立的宁波幻方量化。

DeepSeek公司简介

DeepSeek股权结构



资料来源：爱企查 www.chyxx.com

DeepSeek 大模型不断优化迭代

回顾其发展历史， 2024年1月，发布第一版大模型——DeepSeek LLM，这个版本使用传统的Transformer架构，但在训练方面，已经明显体现出DeepSeek团队通过不断优化训练策略，达到节约成本，提高效率的思想，这点也在后续的模型迭代中被发扬光大。2024年5月，DeepSeek-V2发布，从这一代开始，DeepSeek模型开始使用混合专家（MoE）架构，这是传统Transformer架构的一种改进和扩展，该架构使DeepSeek模型能以更低的计算成本进行更复杂的推理，极大提升了模型的性能。2024年12月，DeepSeek-V3上线并开源，V3版本对MoE架构进行了进一步优化，在维持低训练成本的同时，稳定性与多方面性能表现都达到了与领先闭源模型相当的水平。2025年1月，DeepSeek-R1正式发布，R1模型的推理能力得到极大加强，与OpenAI-o1模型不相上下，且推理过程完全透明，因此在全球范围备受关注。

DeepSeek发展历程

- 2023年7月：DeepSeek 成立
- 2024年1月：发布第一版大模型——DeepSeek LLM
- 2024年5月：宣布开源第二代MoE大模型DeepSeek V2
- 2024年9月：合并DeepSeek Coder V2和 DeepSeek V2 Chat 两个模型，升级推出全新的DeepSeek V2.5 新模型
- 2024年11月：推理模型DeepSeek R1 Lite预览版正式上线
- 2024年12月：宣布DeepSeek V3首个版本上线并同步开源模型权重
- 2025年1月：正式发布DeepSeek R1模型，并同步开源模型权重

DeepSeek模型家族

deepseek

- DeepSeek R1
- DeepSeek V3
- DeepSeek Coder V2
- DeepSeek VL
- DeepSeek V2
- DeepSeek Coder
- DeepSeek Math
- DeepSeek LLM

资料来源：智研咨询整理 www.chyxx.com

PART 02

Deepseek模型家族

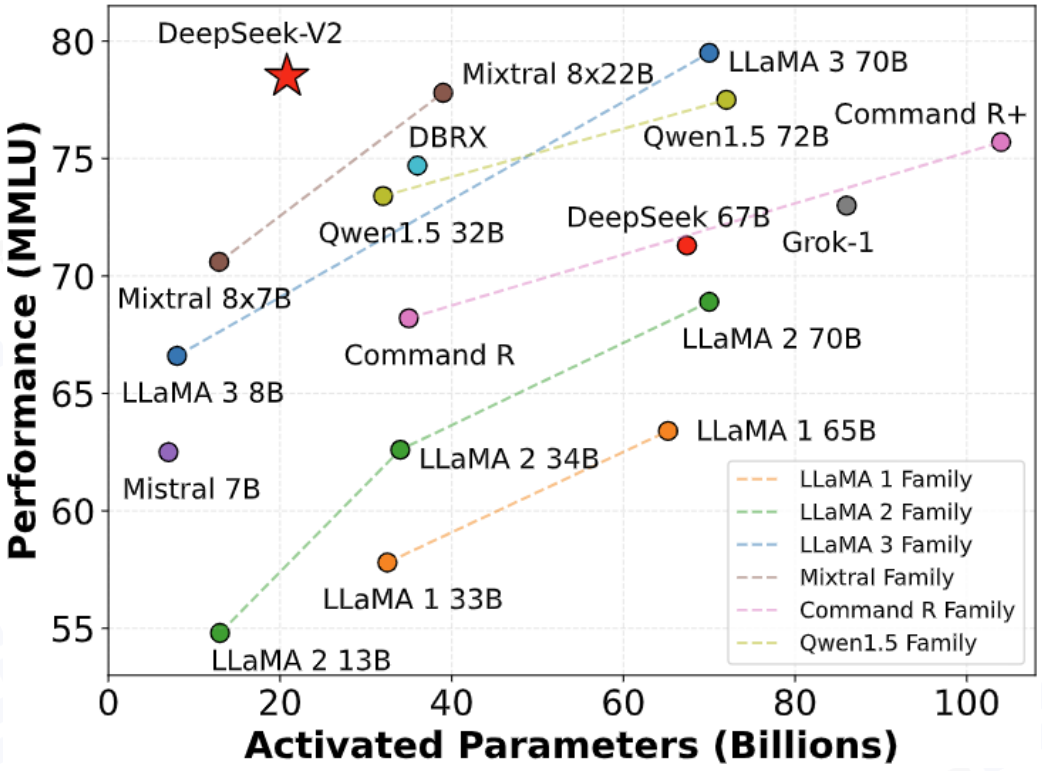
最全面的产业分析 ● 可预见的行业趋势

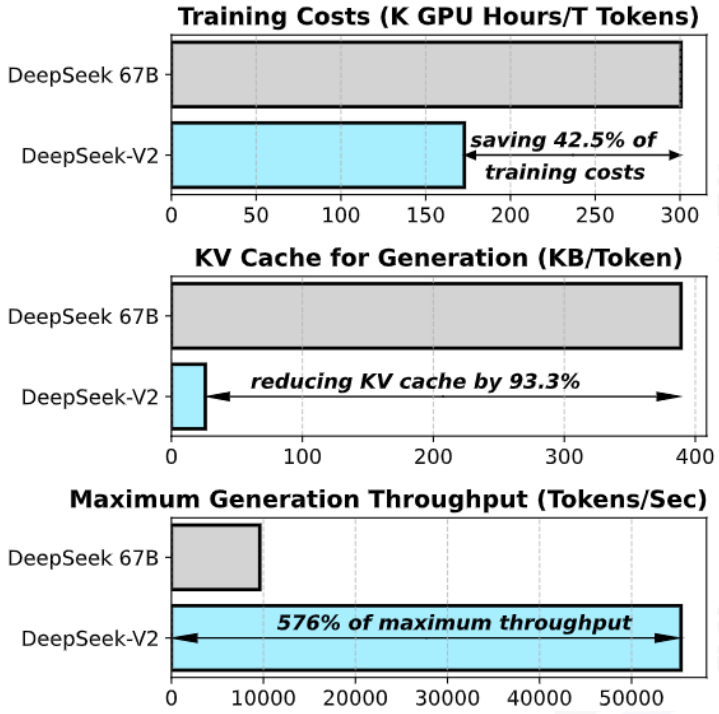
DeepSeek-V2模型性能进一步优化

从低成本的DeepSeek-V2，到超低价格的DeepSeek-V3，再到引起世界广泛关注的DeepSeek-R1，DeepSeek的成功主要依赖于DeepSeek自身深厚的技术积累和持续的技术创新突破。

DeepSeek-V2采用的是MoE架构，全参数量为236B，激活参数量是21B。其采用了两大创新技术：DeepSeekMoE架构和多头潜在注意力（MLA），使得DeepSeek-V2的训练成本大为降低并且提升推理速度。MLA通过将Key-Value缓存压缩为潜在向量来提高推理效率，从而提高吞吐量。DeepSeek MoE架构允许通过稀疏计算进行有效的推理。相比DeepSeek LLM 67B（Dense），DeepSeek- V2的性能更强，同时节省了42.5%的训练成本，减少了93.3%的KV缓存，最大生成吞吐量提高到5.76倍。

DeepSeek- V2性能



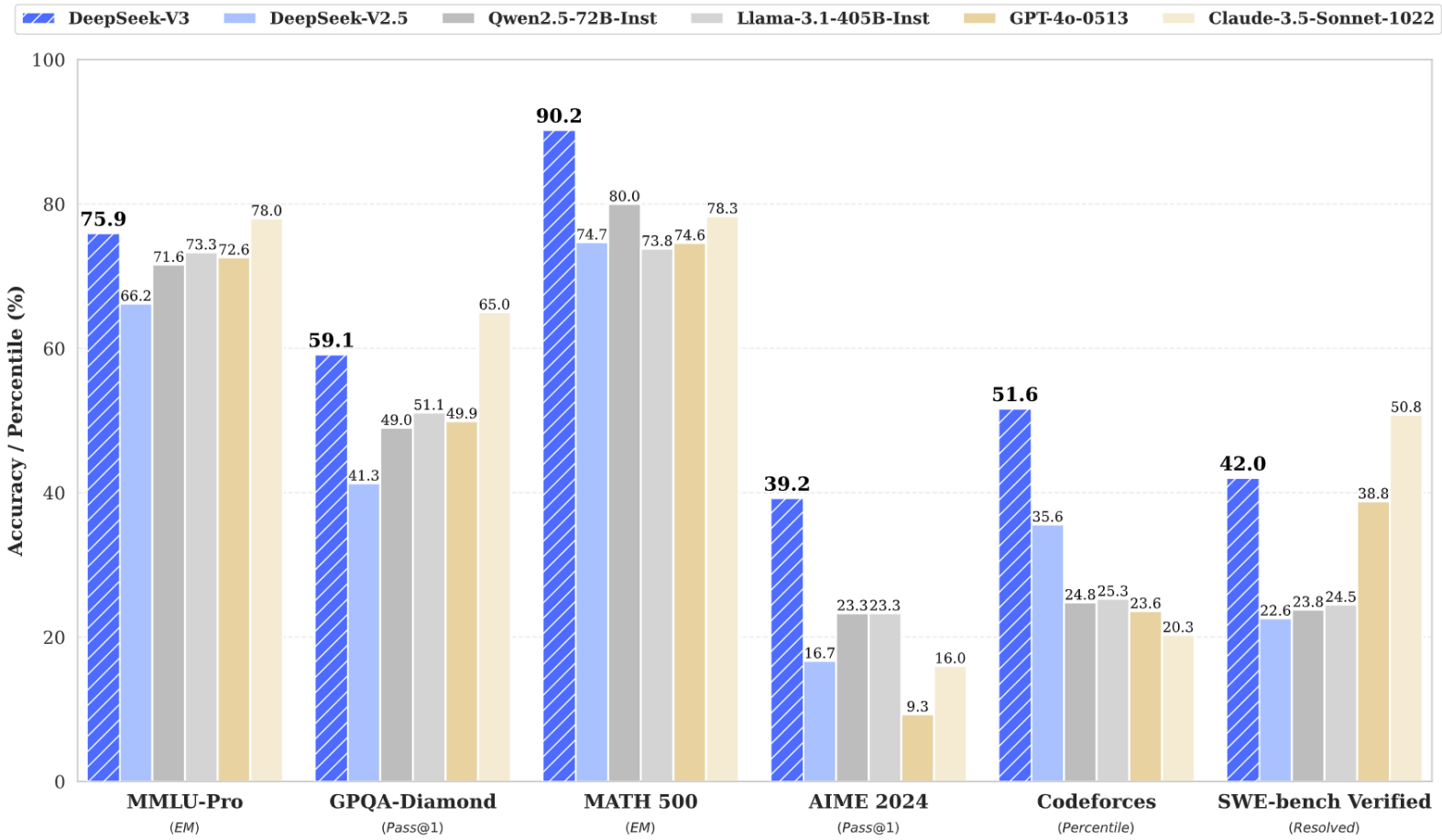


资料来源：DeepSeek、智研咨询整理

DeepSeek-V3模型性能大幅提升

DeepSeek-V3是一个强大的专家混合（MoE）语言模型，具有671B个总参数，激活参数量为37B。相较历史模型，DeepSeek-V3 在推理速度上有了大幅提升。此外在目前大模型主流榜单中，DeepSeek-V3 在开源模型中位列榜首，与世界上最先进的闭源模型不分伯仲。

DeepSeek- v3性能



- DeepSeek-V3 遵循 DeepSeek-V2 的设计，采用多头潜在注意力（MLA）和DeepSeekMoE架构。
- 采用了无辅助损失的负载均衡策略，最大限度地减少了由于鼓励负载平衡而引起的性能下降。
- 引入一个多token预测（MTP）目标，证明它有利于模型的性能，也可用于推理加速的推测解码。

资料来源：DeepSeek、智研咨询整理

DeepSeek-V3模型训练成本大幅降低

根据DeepSeek团队在论文中强调，通过优化算法、框架和硬件的协同设计实现的。在预训练阶段，每万亿个token上训练DeepSeek-V3只需要180 KH800 GPU小时，也就是说，在其拥有2048个H800GPU的集群上只需要3.7天。因此，公司的预训练阶段在不到两个月的时间内完成，花费了2664K GPU小时。加上上下文长度扩展的119K GPU小时和后训练的5K GPU小时，DeepSeek-V3完整训练仅花费278.8万GPU小时。

假设H800GPU的租赁价格为每小时2美元，则代表着其总训练成本仅为557.6万美元。相比同等规模的模型（如GPT-4、GPT-4o、Llama 3.1），训练成本大幅降低。但DeepSeek团队还特意强调，上述成本仅包括DeepSeek-V3的官方训练，不包括与架构、算法或数据的先前研究和消融实验相关的成本。

DeepSeek-V3的训练成本（假设H800的租赁价格为2美元/GPU小时）

训练成本	预训练	上下文扩展	后训练	总计
H800 GPU小时（小时）	2664K	119K	5K	2788K
美元	$5.328M	$0.238M	$0.01M	$5.576M

DeepSeek-V3节省训练成本的方法

模型结构 Architecture	模型训练方式 Pre-Train	针对性GPU优化
DeepSeek MoE+MLA	Dual Pipe	低精度FP8训练
无需辅助损失的负载均衡	All To ALL通信内核 IB+NVLink	PTX语言
多token预测（MTP）	无张量并行TP	带宽限制

资料来源：DeepSeek、智研咨询整理 www.chyxx.com

核心技术——无需辅助损失的负载均衡

DeepSeek-V3 采用了一种无需辅助损失的负载均衡策略，旨在最大限度地减少因负载均衡优化而对模型性能造成的不利影响。MoE 模型容易出现“专家负载不均衡”（有的专家忙，有的专家闲），传统的解决方法是加一个辅助损失，但这可能会损害模型性能。DeepSeek-V3引入了一种新方法，通过动态调整每个专家的“偏置项” ，来平衡负载。这种方法不依赖辅助损失，减少了对性能的负面影响。此外，为了防止在单个序列内出现极端不平衡情况，也引入了一种补充的序列级平衡损失，但影响很小。

$$g'_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} + b_i \in \text{Topk}(\{s_{j,t} + b_j | 1 \leqslant j \leqslant N_r\}, K_r), \\ 0, & \text{otherwise.} \end{cases}$$

无需辅助损失的负载均衡：

具体而言，为每个专家引入一个偏置项 b_i，并将其添加到对应的亲和度得分 $s_{i,t}$，以确定 Top-K 路由。

$$\mathcal{L}_{\text{Bal}} = \alpha \sum_{i=1}^{N_r} f_i P_i,$$

$$f_i = \frac{N_r}{K_r T} \sum_{t=1}^{T} \mathbb{1}\left(s_{i,t} \in \text{Topk}(\{s_{j,t} | 1 \leqslant j \leqslant N_r\}, K_r)\right),$$

$$s'_{i,t} = \frac{s_{i,t}}{\sum_{j=1}^{N_r} s_{j,t}},$$

$$P_i = \frac{1}{T} \sum_{t=1}^{T} s'_{i,t},$$

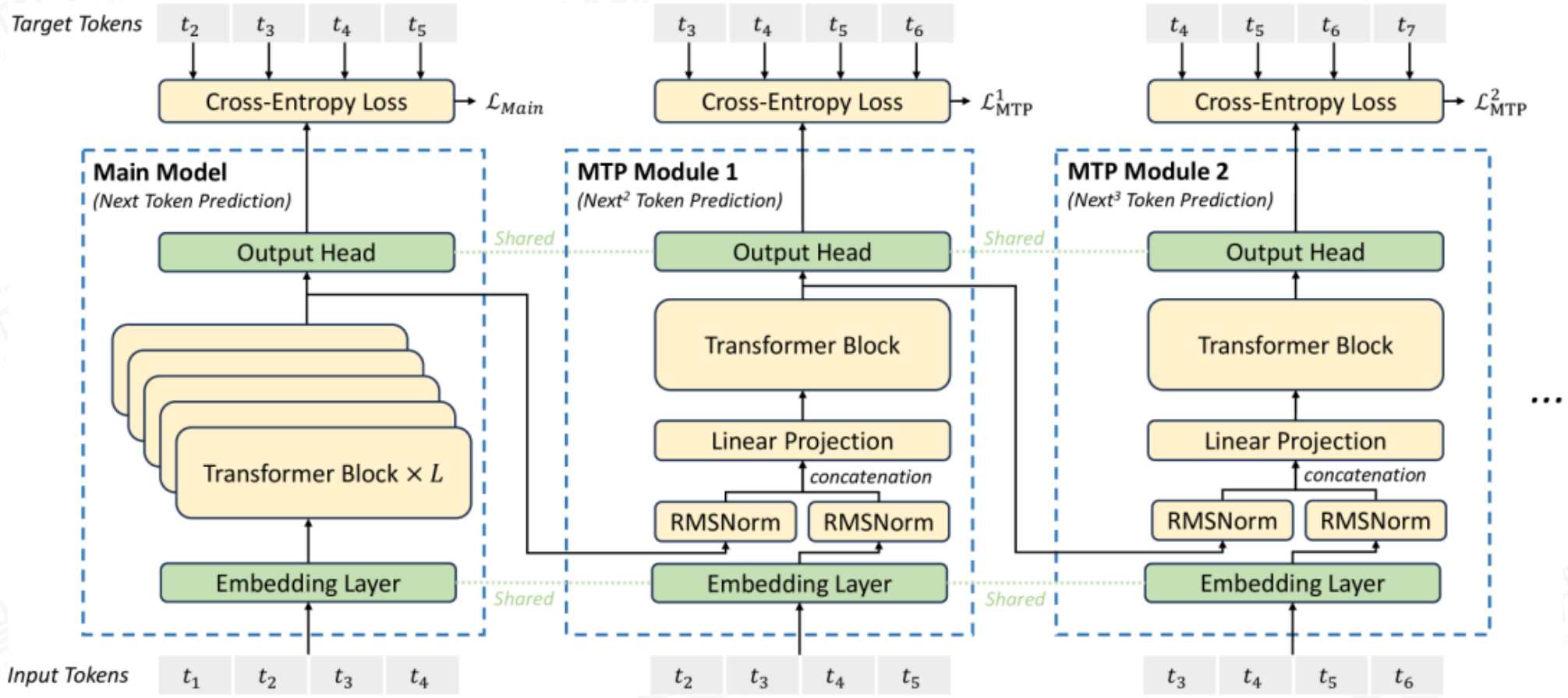
补充的序列级辅助损失：

其中，平衡因子α是一个超参数，对于 DeepSeek-V3 被设置为极小的值；1(.) 表示指示函数；*T*表示序列中的令牌数量。序列级平衡损失鼓励在每个序列内实现专家负载的平衡。

核心技术——多token预测（MTP）

传统语言模型通常只预测下一个token，而DeepSeek-V3在训练中采用 MTP目标，在每个位置预测多个未来token。这种方式增加训练信号密度，提高数据效率，使模型更好规划表示，准确预测未来token。具体通过多层次模块预测多个附加token，各模块共享嵌入层和输出头，保持预测因果链，提高推理生成速度，提升模型整体性能。

MTP实现的示意图

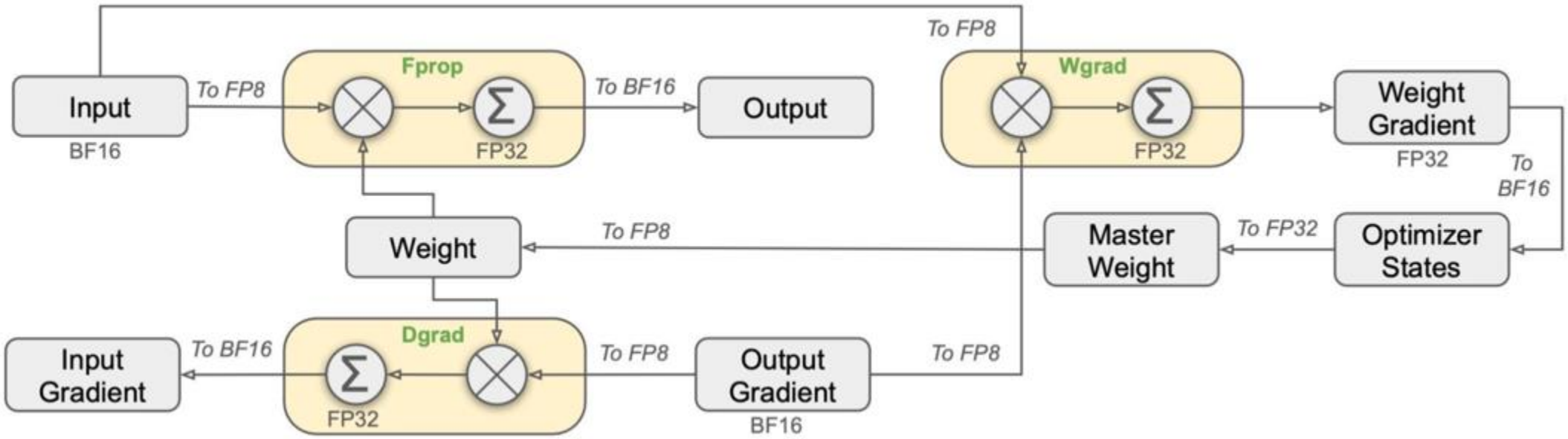


核心技术——FP8混合精度训练

通常的大模型训练会采用BF16或FP32/TF32精度作为数据计算和存储的格式，来确保较高的训练精度。相比之下，FP8占用的数据位宽仅为FP32的1/4，FP16的1/2，可以提升计算速度，降低对存储的消耗。微软2023年的论文《FP8-LM: Training FP8 Large Language Models》就提出了一种用于LLM训练的极度优化的FP8混合精度框架。其核心思想是计算、储存和通信（包括正向和反向传播）全部使用低精度FP8，从而大大降低系统工作负载。然而，使用FP8格式训练LLM存在数据下溢出或上溢出等挑战以及FP8数据格式较低精度所导致训练失败等问题。

DeepSeek团队在训练DeepSeek-V3时，采用的是混合精度框架，大部分密集计算操作都以FP8格式进行，而少数关键操作则策略性地保留其原始数据格式，以平衡训练效率和数值稳定性。通过使用FP8格式，DeepSeek能够在有限的计算资源下，实现更高的计算效率。例如，在处理大规模数据集时，FP8格式可以显著减少显存的占用，从而提高模型的训练速度。

DeepSeek-V3 混合精度框架示意图



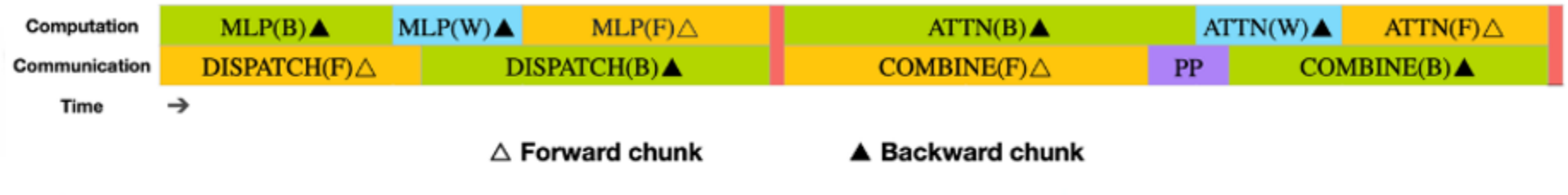
核心技术——Dual Pipe算法

在应用分布式并行策略时，无论是数据并行策略下的梯度聚合步骤，还是模型并行下各模型组件之间的通信，都会带来大量的跨设备数据传输需求。若不同阶段的计算耗时差别较大，则会出现计算设备的空闲，即为“气泡（bubble）”。为解决这一问题，流水线并行（pipeline parallel，PP）策略应运而生。其通过将一个较大数据批次分解为多个微批次（micro batch），使得每次计算的总耗时减少，从而减少了计算设备所处于的计算和等待两种状态在时间轴上的颗粒度，进而使得每个bubble被缩小。

在这一背景下，DeepSeek团队在传统PP策略的基础上创新性地提出并应用了Dual Pipe技术。与传统PP策略相比，Dual Pipe技术最明显的革新在于其有效地融合了前向和后向计算加速通信。此外， DeepSeek团队还通过调节GPU中流式多处理器（SM）的调度来实现对其在计算和通信之间进行精细化分配，进而进一步加速了通信过程。

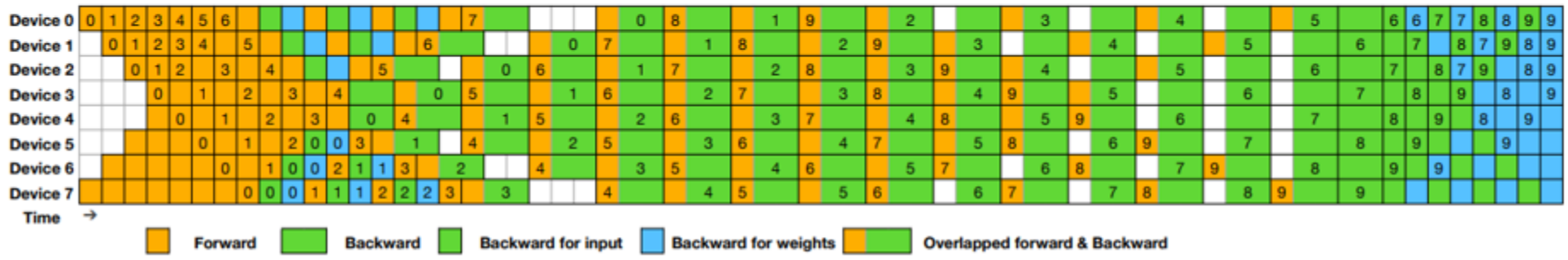
Dual Pipe算法示意图

Chunk拆分



DualPipe

8 个 PP rank 和 20 个 micro-batch 的 DualPipe 调度示例



DeepSeek-R1性能对标OpenAI o1正式版

DeepSeek-R1基于DeepSeek-V3训练优化得到，增强了复杂逻辑推理能力，全参数量是671B，激活参数37B。在数学、代码、自然语言推理等任务上，性能比肩 OpenAI o1 正式版，并且开源模型权重，引发了全球的广泛关注。

DeepSeek-R1评估结果

Category	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	91.8	90.8
	MMLU-Redux (EM)	88.9	88	89.1	86.7	-	92.9
	MMLU-Pro (EM)	78	72.6	75.9	80.3	-	84
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	92.2
	IF-Eval (Prompt Strict)	86.5	84.3	86.1	84.8	-	83.3
	GPQA-Diamond (Pass@1)	65	49.9	59.1	60	75.7	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7	47	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	82.5
	AlpacaEval2.0 (LC-winrate)	52	51.1	70	57.8	-	87.6
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92	-	92.3
Code	LiveCodeBench (Pass@1-COT)	33.8	34.2	-	53.8	63.4	65.9
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	96.6	96.3
	Codeforces (Rating)	717	759	1134	1820	2061	2029
	SWE Verified (Resolved)	50.8	38.8	42	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16	49.6	32.9	61.7	53.3
Math	AIME 2024 (Pass@1)	16	9.3	39.2	63.6	79.2	79.8
	MATH-500 (Pass@1)	78.3	74.6	90.2	90	96.4	97.3
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	78.8
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	92.8
	C-Eval (EM)	76.7	76	86.5	68.9	-	91.8
	C-SimpleQA (Correct)	55.4	58.7	68	40.3	-	63.7

资料来源：DeepSeek、智研咨询整理 www.chyxx.com

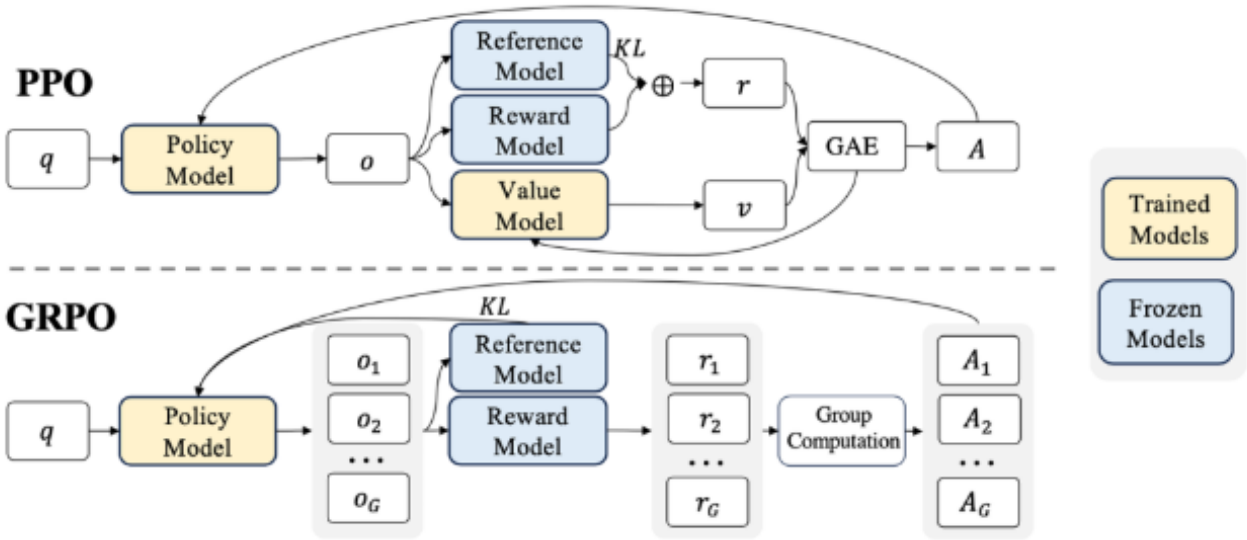
核心技术——纯强化学习训练

DeepSeek-R1具备以下亮点：

（1）纯强化学习训练：基于DeepSeeK-V3应用大规模强化学习，直接将RL应用于基础模型而不依赖监督微调（SFT）作为初始步骤，这种方法允许模型探索解决复杂问题的思维链（CoT），由此开发出DeepSeek-R1-Zero。DeepSeek-R1-Zero是第一个纯强化学习训练得到的LLM，并且展示了自我验证、反思和生成长CoTs等功能，标志研究界的一个重要里程碑。

在大语言模型（LLM）的微调过程中，强化学习（RL）扮演着至关重要的角色。传统的近端策略优化（PPO）算法虽然被广泛应用于LLM的微调，但其在处理大规模模型时面临着巨大的计算和存储负担。PPO算法需要维护一个与策略模型大小相当的价值网络来估计优势函数，这在大模型场景下会导致显著的内存占用和计算代价。此外，PPO算法在更新策略时可能会导致策略分布发生剧烈变化，从而影响训练的稳定性。为了解决这些问题，DeepSeek提出了一种新的强化学习算法——组相对策略优化（GRPO），旨在减少对价值网络的依赖，同时保持策略更新的稳定性和高效性。

算法结构对比



GRPO方法的优势在于：

（1）减少计算负担：通过避免维护一个与策略模型大小相当的价值网络，GRPO显著降低了训练过程中的内存占用和计算代价。

（2）提高训练稳定性：GRPO通过组内比较来估计优势函数，减少了策略更新的方差，从而确保了更稳定的学习过程。

（3）增强策略更新的可控性：GRPO引入了KL散度约束，防止策略更新过于剧烈，从而保持了策略分布的稳定性。

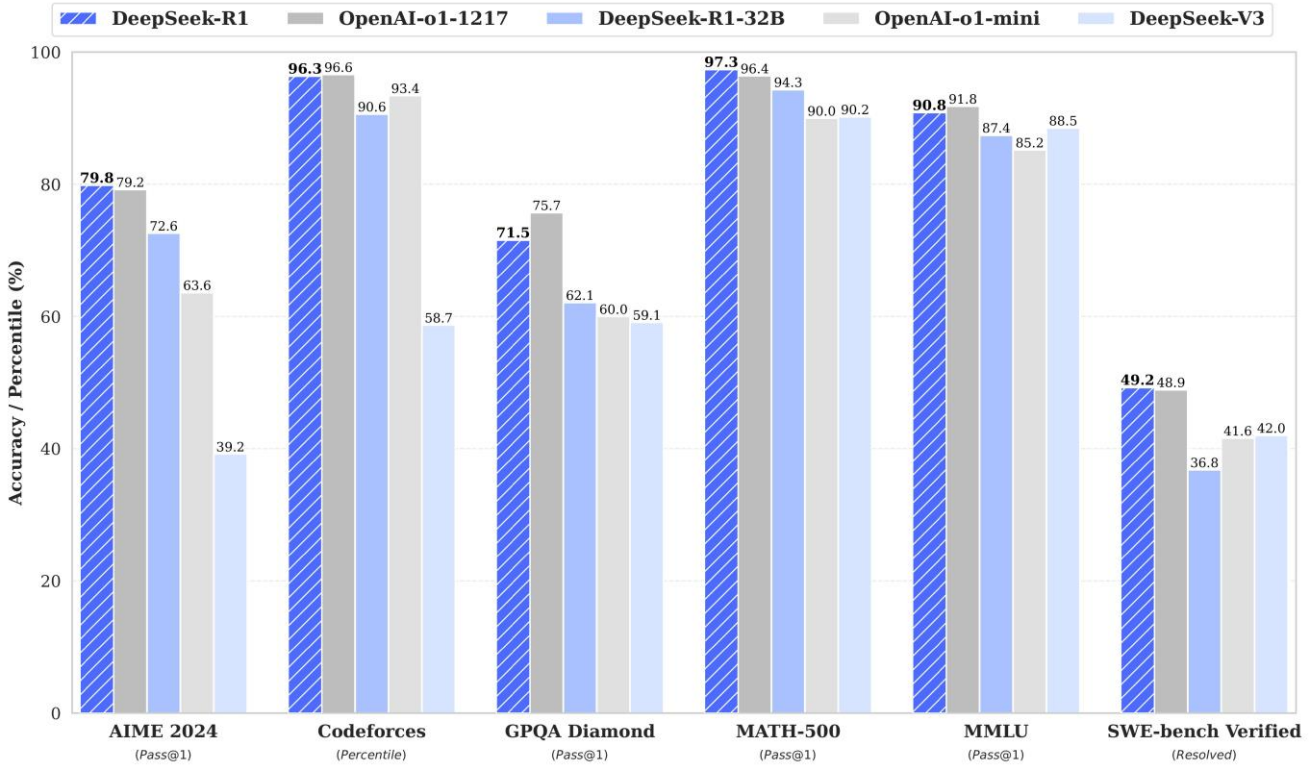
核心技术——冷启动数据&多阶段训练策略

（2）冷启动数据&多阶段训练策略：DeepSeek-R1 是为解决 DeepSeek-R1-Zero 存在的问题并进一步提升推理性能而开发的模型，它在训练过程中融入了冷启动数据和多阶段训练策略。

冷启动数据：收集少量高质量长链推理数据，通过SFT初始化模型，提升可读性和性能。

多阶段训练：第一阶段 RL 专注于数学、编程等明确答案的任务。第二阶段结合拒绝采样生成 SFT 数据，增强通用能力（写作、问答等）。最终RL对齐人类偏好（如无害性、有用性）。

DeepSeek-R1的基准性能



DeepSeek-R1 在多个基准测试中展现出与OpenAI-o1相当的性能水平。在Codeforces和MMLU基准测试中与OpenAI-o1-1217得分相近，尤其是在AIME 2024、MATH-500、Swe-Bench等基准测试中，DeepSeek-R1还稍微胜出。

资料来源：DeepSeek、智研咨询整理 www.chyxx.com

核心技术——模型能力蒸馏迁移

（3）模型能力蒸馏迁移：DeepSeek R1 的推理能力可以通过蒸馏技术迁移到更小的模型中，并且小模型的基准测试取得很优秀的表现。在DeepSeek-R1蒸馏出的6个小模型中，在保持模型参数量仅为o1-mini同量级的前提下，其知识理解、代码生成等核心能力实现全面反超。通过对标OpenAI-o1-mini的效果上不难看出DeepSeek在模型轻量化领域的突破性创新，同时也为开源社区提供了兼具高性能与低部署成本的新型解决方案。

DeepSeek-R1蒸馏小模型性能

	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	**1820.0**
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633.0

资料来源：DeepSeek、智研咨询整理 www.chyxx.com

PART 03

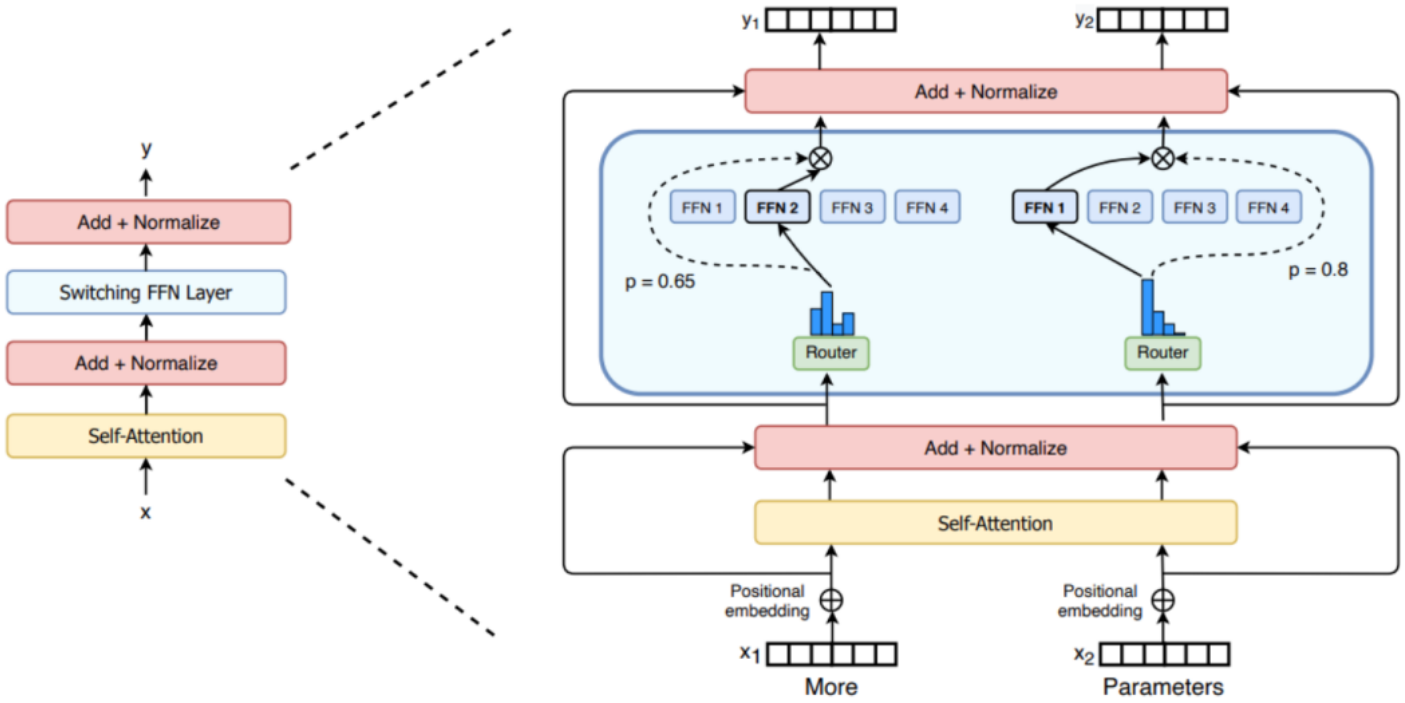
Deepseek技术创新

最全面的产业分析 ● 可预见的行业趋势

MoE架构引入多个独立的专家模型

MoE，全称Mixture of Experts，即混合专家模型，是一种用于提高深度学习模型性能和效率的架构。其核心思想是通过引入多个独立的专家模型（Experts），每个输入数据只选择和激活其中的一部分专家模型来进行处理，从而减少计算量，提高训练和推理速度。MoE的概念在1991年就已提出，训练不容易收敛是其在大模型领域应用的主要障碍。

MoE模型结构



MoE模型的主要组成部分包括：

（1）专家（Experts）：模型中的每个专家都是一个独立的神经网络，专门处理输入数据的特定子集或特定任务。例如，在自然语言处理任务中，一个专家可能专注于处理与语言语法相关的内容，而另一个专家可能专注于语义理解。

（2）门控网络（Gating Network）：门控网络的作用是决定每个输入样本应该由哪个专家或哪些专家来处理。它根据输入样本的特征计算出每个专家的权重或重要性，然后根据这些权重将输入样本分配给相应的专家。门控网络通常是一个简单的神经网络，其输出经过softmax激活函数处理，以确保所有专家的权重之和为1。

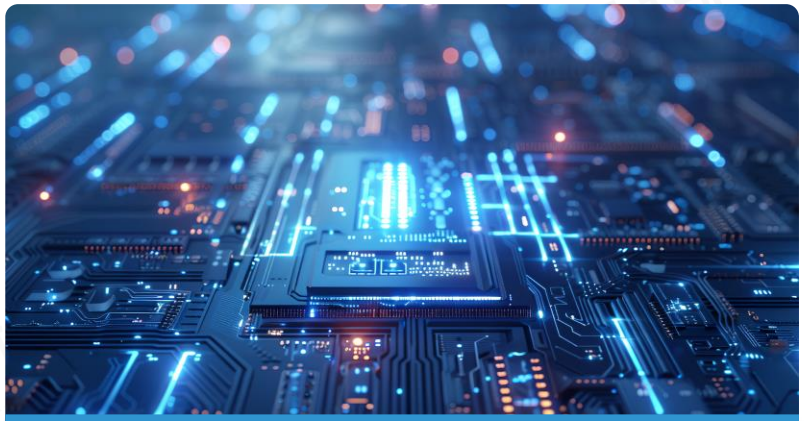
MoE架构可显著提高训练效率

提高模型性能

通过将多个专家的预测结果进行整合，MoE模型可以在不同的数据子集或任务方面发挥每个专家的优势，从而提高整体模型的性能。例如，在图像分类任务中，一个专家可能擅长识别动物图片，而另一个专家可能擅长识别车辆图片，通过门控网络的合理分配，MoE模型可以更准确地对不同类型的图片进行分类。

减少计算成本

与传统的密集模型相比，MoE模型在处理每个输入样本时，只有相关的专家会被激活，而不是整个模型的所有参数都被使用。这意味着MoE模型可以在保持较高性能的同时，显著减少计算资源的消耗，特别是在模型规模较大时，这种优势更为明显。例如，对于一个具有数十亿参数的大型语言模型，采用MoE架构可以在不增加太多计算成本的情况下，通过增加专家的数量来进一步提升模型的性能。



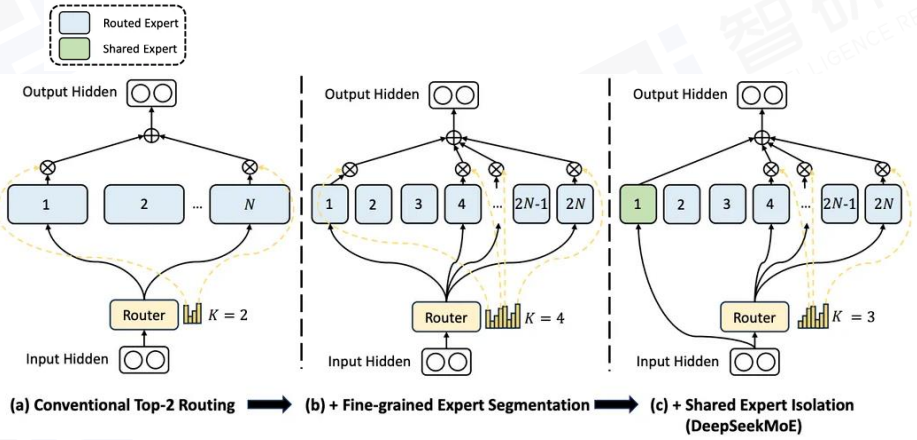
增强模型的可扩展性

MoE模型的架构设计使得它可以很容易地扩展到更多的专家和更大的模型规模。通过增加专家的数量，模型可以覆盖更广泛的数据特征和任务类型，从而在不增加计算复杂度的情况下，提升模型的表达能力和泛化能力。这种可扩展性为处理大规模、复杂的数据集提供了有效的解决方案，例如在处理多模态数据（包含文本、图像、语音等多种类型的数据）时，MoE模型可以通过设置不同的专家来专门处理不同模态的数据，实现更高效的多模态融合。

DeepSeek MoE在传统MoE模型架构上进行了改进

DeepSeek MoE从传统MoE模型架构的基础上，进行了两部分改进：（1）细粒度专家划分：相比传统MoE模型，DeepSeekMoE将每个MoE层细分为更多的细粒度专家，每个专家负责处理更具体的任务。例如，在一个典型的DeepSeekMoE模型中，每个MoE层包含256个专家，每个token会激活其中的8个专家。这种细粒度的分割方式使得每个专家能够专注于特定类型的输入数据，从而提高模型的灵活性和表达能力。（2）共享专家隔离：传统的MoE模型中，所有专家都是独立的，每个专家都需要独立处理输入数据。DeepSeekMoE引入了共享专家的概念，把激活专家区分为共享专家和路由专家时，共享专家和路由专家在数据处理流程上有显著的区别。对于共享专家，输入数据无需经过路由模块的计算，所有数据都会直接通过共享专家进行处理。相反，对于路由专家，输入数据会先经过路由模块，该模块根据输入数据的特征选择最合适的专家进行计算。在这种架构中，路由模块通过计算输入数据与各个专家的匹配概率，选择概率最高的专家进行处理。最终，将路由专家和共享专家的计算结果相加，形成MoE模块的最终输出。通过这种方式，模型能够在处理不同输入数据时，既能捕捉到输入数据的共性，也能关注到输入数据的差异性。这种设计能够提高模型的泛化能力和适应性。

DeepSeek MoE与传统MoE的区别



部分开源模型MoE模块配置对比

模型	细粒度	专家分离	共享专家数	路由专家数	激活专家数
Mixtral 8*7B	否	否	0	8	2
Hunyuan-Large	否	是	1	16	1
Qwen1.5-MoE-A2.7B	是	是	4	60	4
DeepSeek-V3	是	是	1	256	8

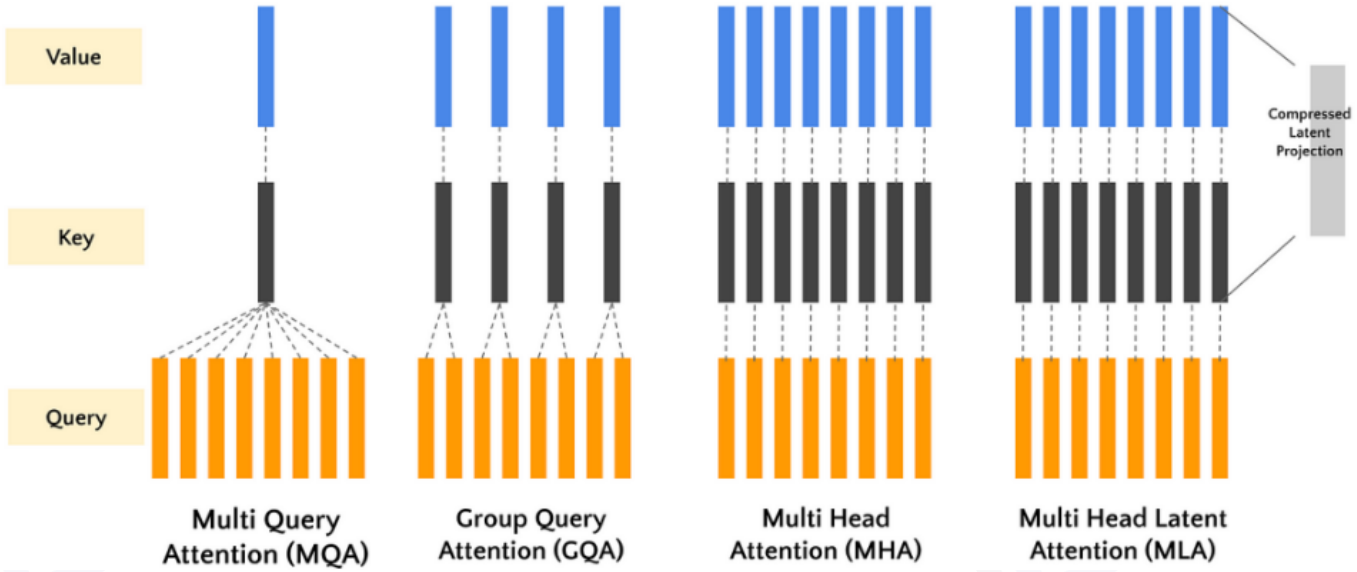
资料来源：智研咨询整理 www.chyxx.com

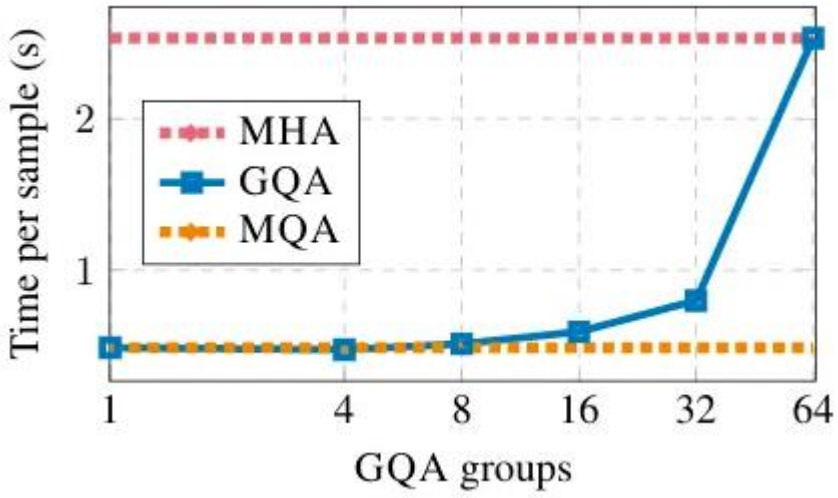
多头潜在注意力MLA进一步减少KV缓存的大小

在标准的Transformer模型中，多头注意力（MHA）机制通过并行计算多个注意力头来捕捉输入序列中的不同特征。每个注意力头都有自己的查询（Q）、键（K）和值（V）矩阵。对于序列中的每一个token，都需要计算各自的QKV，进而计算注意力。在推理过程中，当前大模型所采用的token by token递归生成方式，上文token的KV计算不会受到后续生成token的影响，因此可以缓存下来，避免重复计算，提高推理效率，这就是KV cache的由来。也就是说，当生成第个token时，可以利用之前事先算好的上文个 token的KV值。同样地，位置token的KV值计算出来后也将保存在KV cache中。

目前大模型对于注意力机制做的一些改进，包括MQA 、GQA都是为了想方设法减少KV Cache。DeepSeek提出的MLA的出发点也是如此。减少KV Cache就可以实现在更少的设备上推理更长的Context，或者在相同的Context长度下让推理的batch size更大，从而实现更快的推理速度或者更大的吞吐总量。最终目的都是为了实现更低的推理成本。

MHA、MQA、GQA 与 MLA

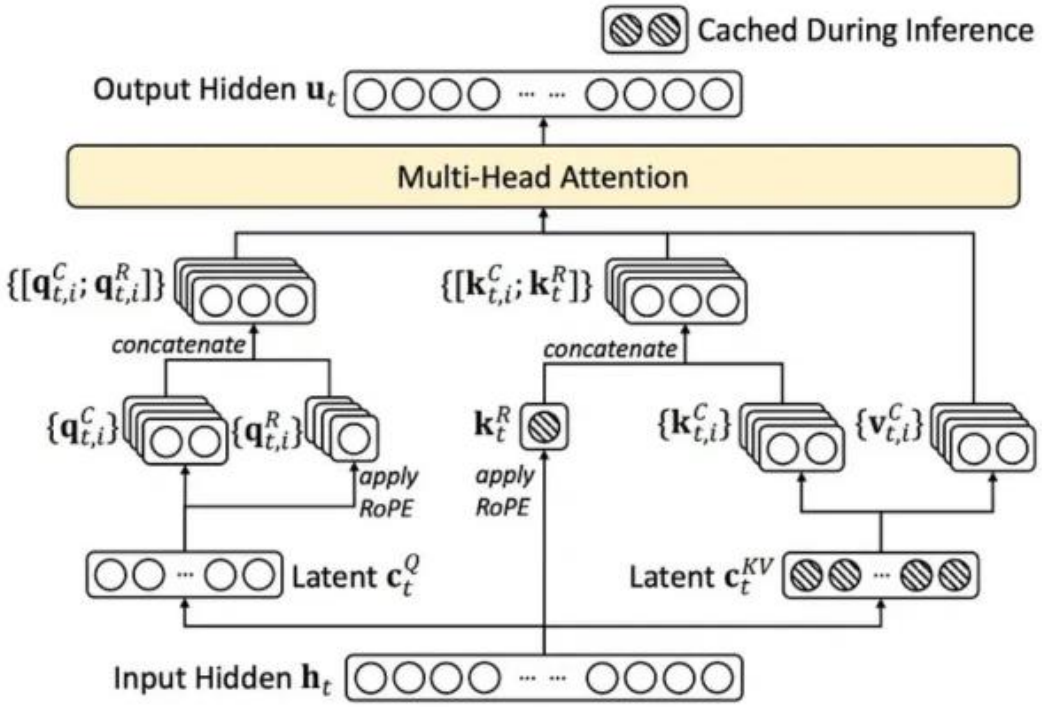


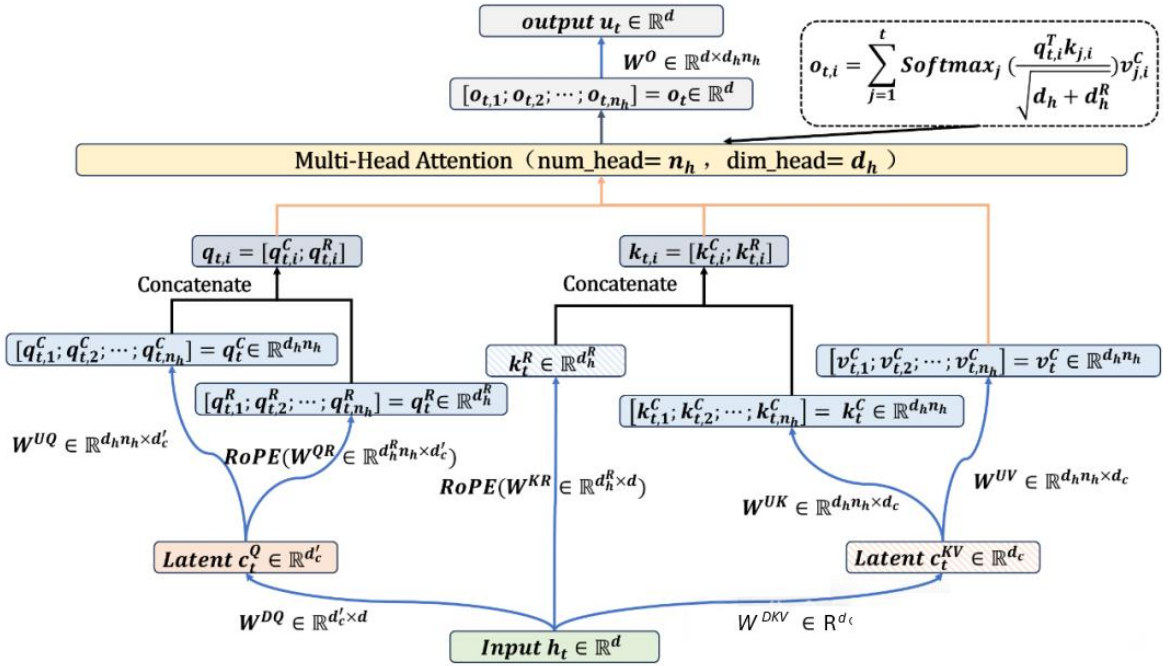


多头潜在注意力MLA实现了更低的推理成本

MQA与GQA的办法是通过共享K、V的注意力头，降低KV的数据维度，但会牺牲模型性能。MLA则是通过对注意力机制中的K、V进行低秩联合压缩，减少推理时的KV缓存；同时对Q进行低秩压缩，减少训练期间的激活内存使用。MLA架构还结合了旋转位置嵌入（RoPE），有效处理了长序列中的位置依赖问题。RoPE通过旋转操作将位置信息嵌入到K和Q中，使得模型能够更好地捕捉长距离依赖关系。尽管MLA通过低秩压缩减少了K、V缓存和激活内存，但它仍然能够保持与标准多头注意力（MHA）相当的性能。在推理过程中，MLA只需要缓存压缩后的键和值，这显著减少了内存占用，使得模型能够处理更长的上下文长度。

MLA架构





DeepSeek V3与R1模型采用MIT协议

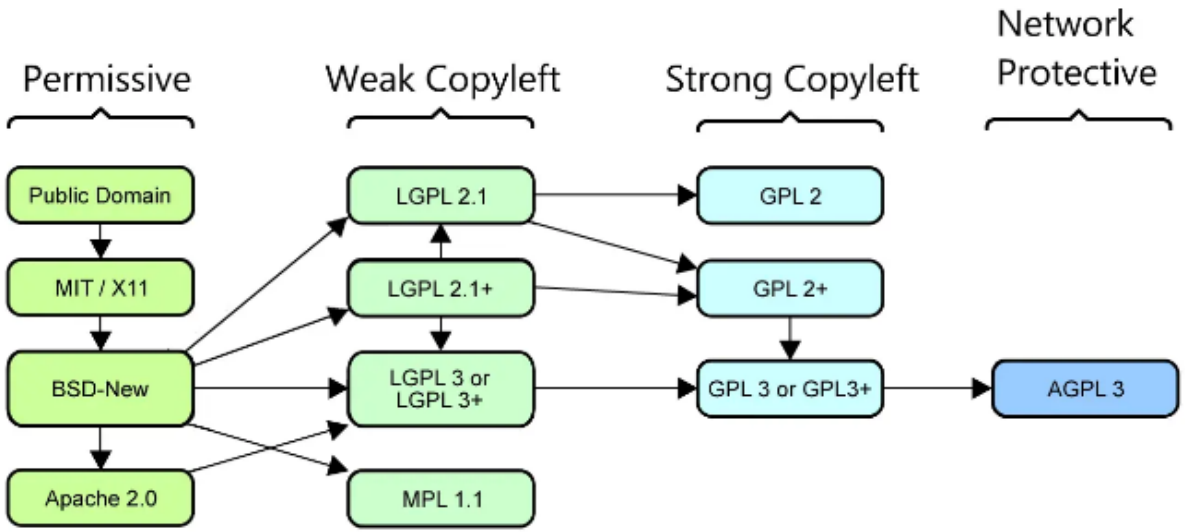
开源即代码层面开源，可以调用与进行二次开发。开源免费调用有助于先行占据市场份额，成为规则制定者，率先拓展生态粘性。如，谷歌将安卓开源，获得了全球80%的移动手机端市场份额，同时也覆盖电视、汽车等使用场景。

DeepSeek V3与R1模型实现了开源，采用MIT协议。 DeepSeek开源模型完全免费，开发者可以利用DeepSeek开源模型开发衍生模型、产品应用以及生成内容。这产生多方面影响：

① 对大模型发展：这提升了世界对中国AI大模型能力的认知，一定程度打破了OpenA1与Anthropic等高级闭源模型的封闭生态。DeepSeek R1在多个测试指标中对标OpenAI o1，通过模型开源，也将大模型平均水平提升至类OpenAI o1等级。

② 对下游生态：优质的开源模型可更好用于垂类场景，即使用者针对自身需求蒸馏，或用自有数据训练，从而适合具体下游场景；此外，模型训推成本降低，将带来使用场景的普及，带动AIGC、端侧等供给和需求。

开源许可协议标准



用户通过获取DeepSeek开源项目中相关信息进行部署/再训练使用，应首先确保满足开源项目对应许可协议。目前，DeepSeek系列开源AI项目，除DeepSeek-R1代码和模型皆遵循MIT开源许可协议外，其他DeepSeek系列开源AI项目皆为代码遵循MIT开源许可协议，模型遵循DEEPSEEK LICENSE AGREEMENT（Version 1.0）。

因此，用户在部署/再训练DeepSeek大模型开源项目时，应首先遵循对应开源许可协议的相关规定，避免开源合规风险。

PART 04

Deepseek商业模式

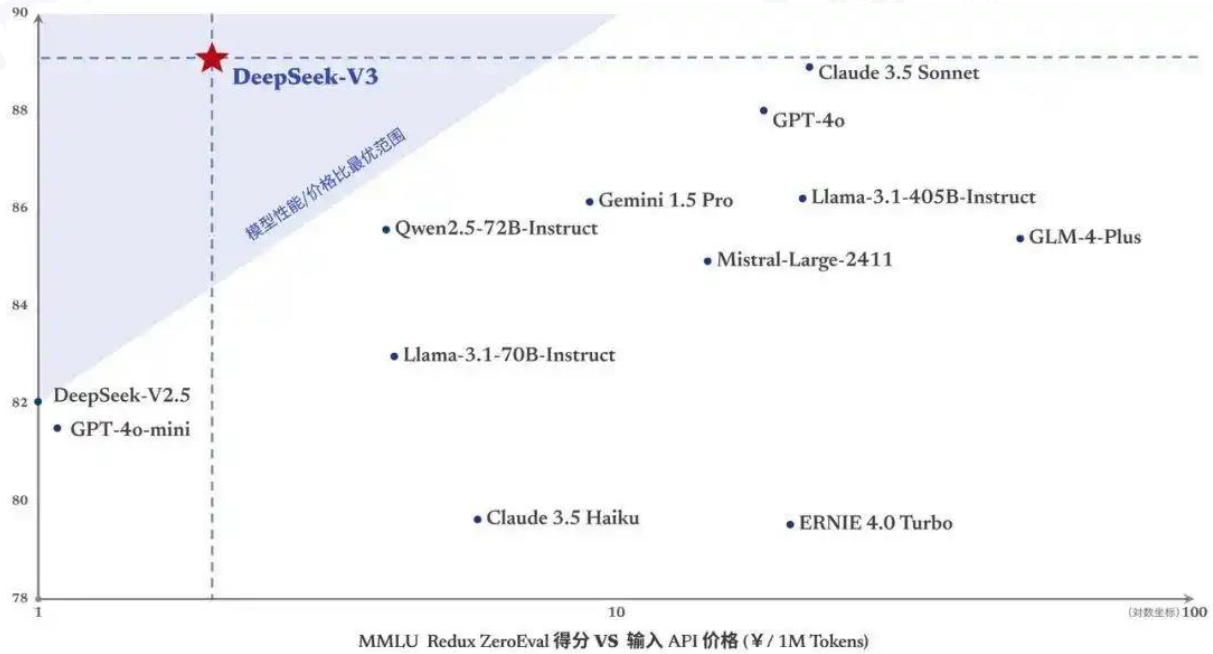
最全面的产业分析 ● 可预见的行业趋势

DeepSeek API 性价比优势明显

DeepSeek API 接入价格

模型	时段	百万tokens输入价格（缓存命中）	百万tokens输入价格（缓存未命中）	百万tokens输出价格输出价格
deepseek-chat（DeepSeek-V3）	标准时段	0.5元	2元	8元
	优惠时端（00:30-8:30）	0.25元	1元	4元
deepseek-reasoner（DeepSeek-R1）	标准时段	1元	4元	16元
	优惠时端（00:30-8:30）	0.25元	1元	4元

DeepSeek-V3 API定价对比海内外主流模型



企业接入DeepSeek大模型的收费方式主要分为两种模式，具体如下：

（1） API接口：按Token计费模式。

标准时段下，deepseek-chat（DeepSeek-V3）API服务定价为百万tokens输入价格0.5元（缓存命中）/2元（缓存未命中）。deepseek-reasoner（DeepSeek-R1） API服务定价为百万tokens输入价格1元（缓存命中）/4元（缓存未命中）。

2月26日，deepseek平台推出错峰优惠活动，在00:30-8:30时间段，DeepSeek-V3降至原价的50%，DeepSeek-R1降至原价的25%。

本地化部署稳定性更强，成为企业重要选择

（2） 本地化部署：把Deep Seek在本地电脑上部署，然后直接在本地访问。本地化部署对硬件要求高、运维更加复杂、成本高昂，下游客户表示，部署一个DeepSeek R1，需要30万~40万元的成本。但本地化部署在稳定性、灵活性、数据安全方面具有显著优势。

DeepSeek本地化部署成本及优劣势

劣势

初期成本高昂

本地化部署需要客户投入大量资金购买高性能硬件设备（如GPU、TPU 等）。此外，还需组建专业团队负责模型的部署、优化和运维。此外还有额外投入，如散热设备、服务器机房的建设和电力消耗。

技术门槛高

部署和优化大模型涉及复杂的技术环节，包括模型压缩、推理加速、分布式计算等。这对技术团队的能力提出了较高要求，需要具备深厚的技术背景和丰富的实践经验。

扩展性有限

本地化部署的计算资源是固定的，难以灵活应对突发性的大规模请求。相比之下，云服务可以按需扩展资源以满足需求。当业务需求超出现有硬件能力时，可能需要追加硬件投资。

生态集成难度

云端服务通常自带丰富的功能（如预训练插件、API 接口等），而本地化部署需要自行开发和集成，这增加了开发和维护的难度和工作量。

deepseek

本地化部署：30-40万元（DeepSeek R1）

大型企业或机构/行业专家团队/高科技创业公司/科研机构

优势

数据隐私与安全性

本地化部署的核心优势在于对数据隐私的高度保障。医院将模型部署在内部系统中，能够完全掌控数据流，避免将敏感信息上传至云端，从而有效降低隐私泄露的风险，更好地满足《数据安全法》的要求。

定制化能力

本地化部署允许客户根据自身需求对模型进行深度微调和优化。例如，医院可以针对特定领域的知识对模型进行额外训练，从而提升其适用性和准确性。

性能稳定

本地化部署无需依赖外部网络连接，避免了因网络延迟或云端服务中断导致的业务中断。对于需要实时响应的应用场景，本地部署通常能显著降低延迟，提供更高的服务稳定性。

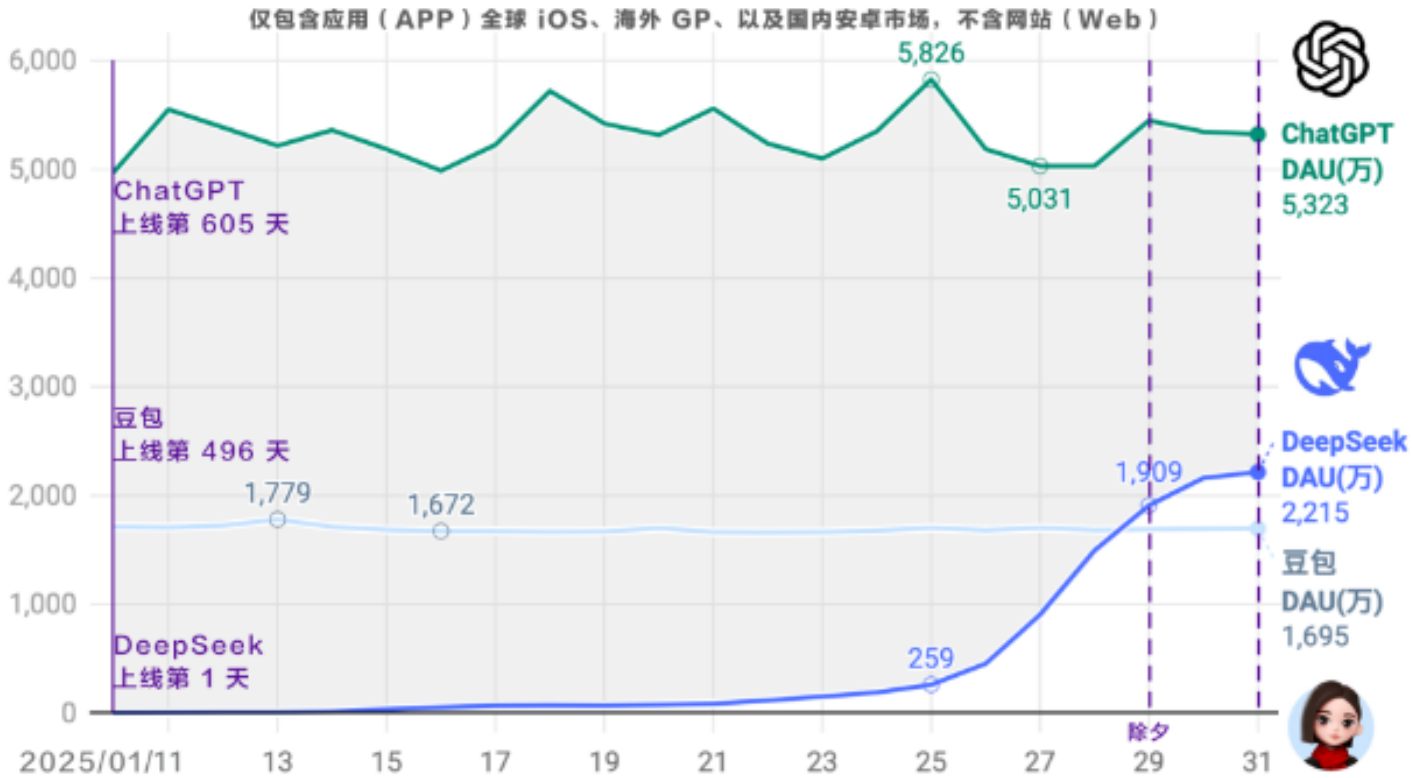
长期成本控制

对于高频、大规模使用的场景，本地化部署在长期内可能比持续使用云端 API 更具有成本效益。

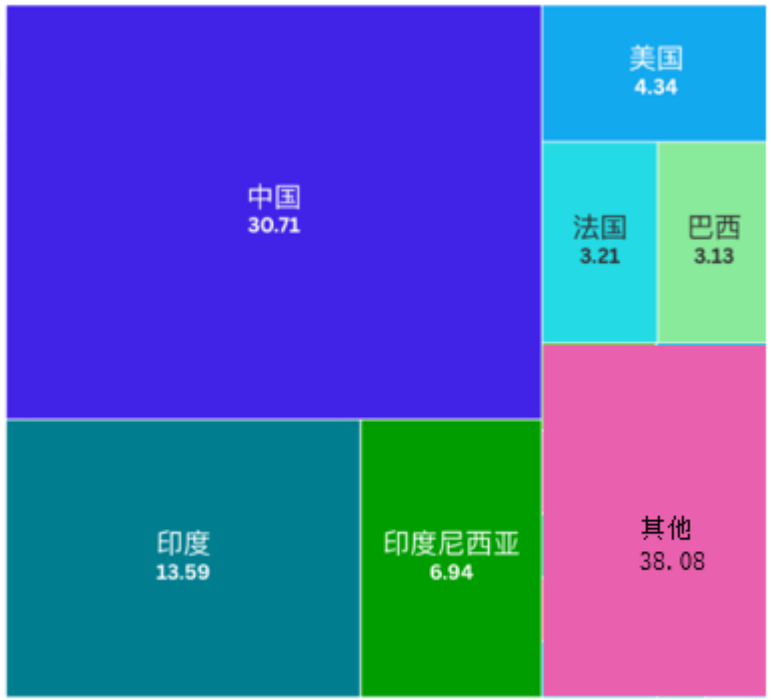
DeepSeek App用户规模迅速增长

DeepSeek App自2025年1月11日上线以来，截至2月9日，累计下载量已突破1.1亿次。其中，1月20日至1月26日，DeepSeek App的周下载量达到226万次，而在随后的一周内，下载量激增至6300万次，环比增长超过2700%。这一增长主要得益于其开源推理模型DeepSeek-R1的发布。

2025年1月DeepSeek日活跃用户DAU



DeepSeek 应用(APP) MAU 月活跃用户分布

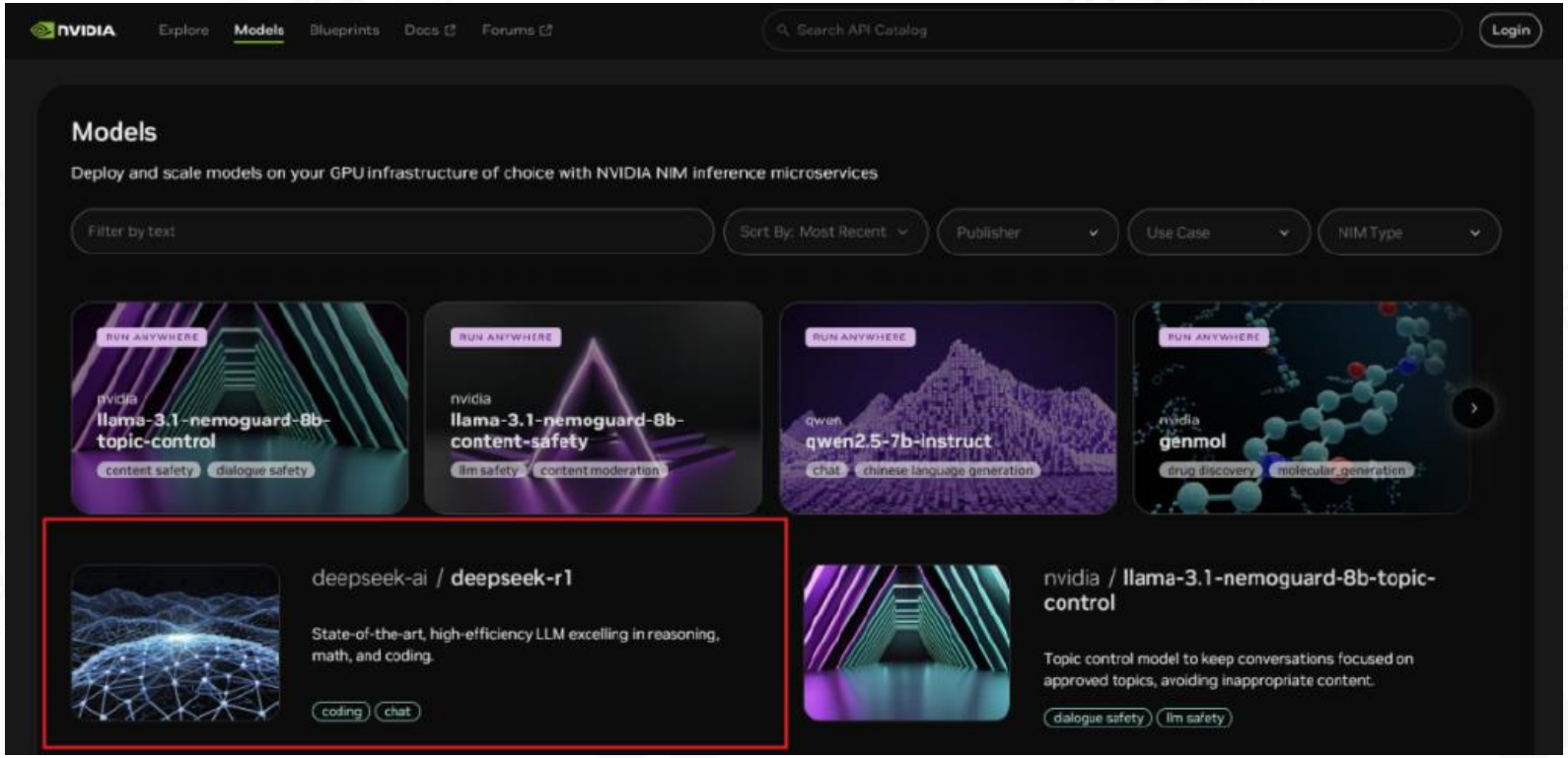


资料来源：智研咨询整理 www.chyxx.com

海外科技巨头纷纷宣布上线DeepSeek大模型

DeepSeek热度持续席卷全球，微软Azure、英伟达等海外科技巨头纷纷宣布上线DeepSeek大模型。1月，微软最早宣布将DeepSeek-R1模型添加到云平台Azure AI Foundry，开发者可用于构建基于云的应用程序和服务。1月25日， AMD宣布已将新的DeepSeek-V3模型集成到Instinct MI300X GPU上，该模型旨在与SGLang一起实现最佳性能。1月30日，美国人工智能巨头英伟达（Nvidia）在官网宣布，DeepSeek-R1模型可作为NVIDIA NIM微服务预览版使用。1月31日，亚马逊云科技官方公告，DeepSeek的R1模型已正式在Amazon Bedrock及Amazon SageMaker AI平台上全面推出。

英伟达接入DeepSeek-R1模型



PART 05

Deepseek应用场景

最全面的产业分析 ● 可预见的行业趋势

能源企业“牵手”DeepSeek已成为一股新风潮

近来，能源企业“牵手”DeepSeek已成为一股新风潮。据不完全统计，能源领域的央企如中国石化、中国石油、中国海油、中国中化、国家能源集团、中国核电、中广核、华能集团、国家电投、华电集团、南方电网等多家能源企业相继宣布，已完成DeepSeek 大模型私有化部署，全面接入企业自有的 AI 大模型。

DeepSeek作为一款具有强大算法优化能力的人工智能平台，将为能源领域提供更加精准和高效的数据分析与处理方案。这意味着，能源企业不仅可以在日常管理中更好地应对复杂的能源系统问题，还能够通过智能化手段提升能源业务的运营效率。

部分能源企业DeepSeek部署情况

企业	动态
中国石油	2月8日，中国石油昆仑大模型正式完成DeepSeek大模型私有化部署。
国家管网集团	2月10日，国家管网集团完成满血版DeepSeek模型的私有化部署。
国家能源集团	2月11日，在国家能源集团科信部指导下，信息技术公司（集团数据中心）顺利完成DeepSeek-R1系列大模型在国能企业云平台本地化部署并正式上线。
龙源电力	2月12日，龙源电力宣布，新能源数字化平台部署上线DeepSeek-R1系列大模型。
南方电网	2月12日，南方电网人工智能创新平台完成了开源大模型DeepSeek的本地化部署。
中国中化	2月13日，中国中化人工智能平台完成DeepSeek系列模型部署，通过私有化部署方式面向全公司提供开放服务。
中广核	2月13日消息，中广核AI大模型完成了DeepSeek的全面接入，实现了DeepSeek模型在集团的本地化部署。
中南电力	2月14日，中南电力顺利完成了DeepSeek开源大模型在本地的部署工作。
中国华能集团	2月15日，中国华能集团有限公司完成了DeepSeek系列模型的本地化部署，并推出了“睿智小能”AI助手与“iHN+”移动门户。
正泰新能源	2月18日，正泰新能源宣布，正泰新能源售电交易事业部与信息管理部联合，使DeepSeek人工智能系统成功应用于上海市松江区虚拟电厂项目。

三大运营商相继宣布全面接入DeepSeek

中国电信、中国移动、中国联通三大运营商相继宣布全面接入DeepSeek，在通信与AI融合领域激起千层浪。运营商具备全国最大的流量通道和数据积累，同时在云业务硬件基础上具有较高的普及度，这些因素使得DeepSeek的全面接入有望加速AI应用的发展，推动云业务的持续增长。

运营商通过接入DeepSeek，利用其在深度学习和多场景适应能力上的优势，旨在提升网络管理效率和客户服务质量。通过与DeepSeek的深度合作，运营商能够在网络优化、智能客服、个性化服务等领域实现突破，进一步巩固其在通信行业的领先地位。

部分能源企业DeepSeek部署情况

中国电信 CHINA TELECOM

中国电信通过天翼云全场景上架DeepSeek，提供从部署到推理、微调的全流程服务。用户可在天翼云智算产品体系——息壤-科研助手、天翼AI云电脑、魔乐社区、“息壤”智算平台、GPU云主机/裸金属开启智能新体验。在科研场景，“息壤-科研助手”基于DeepSeek构建WebUI应用服务，帮助科研工作者提高学术资源检索和文献阅读分析效率；普惠AI场景中，天翼AI云电脑接入DeepSeek，提供智能会话服务，赋能办公、教育和生活等多元化场景。

中国联通 China unicom

联通云已基于星罗平台实现国产及主流算力适配多规格DeepSeek-R1模型，兼顾私有化和公有化场景，提供全方位运行服务保障。联通云基于A800、H800、L40S等多款主流算力卡，预置DeepSeek-R1多尺寸模型，用户可按需灵活选择、快速扩展，快速搭建DeepSeek-R1推理和微调环境。

中国移动 China Mobile

中国移动移动云全面上线DeepSeek，实现全版本覆盖、全尺寸适配、全功能使用。中国移动覆盖全国的13个智算中心全面上线上述能力，用户可选择任一智算资源池进行部署、蒸馏、智能体编排等操作。此外，移动云深度集成DeepSeek模型，搭载自研的COCA算力原生平台，实现“开箱即用”的便捷性，还为DeepSeek-R1模型定制算力方案。

DeepSeek在金融领域的应用场景较为广泛

自DeepSeek V3/R1模型发布以来，金融机构纷纷将其视为提升技术实力和市场竞争力的重要抓手，并加速推进部署与应用。已有多家金融机构宣布接入或部署DeepSeek。DeepSeek通过底层数据的深度关联和逻辑推演，为用户提供可靠的数据支持，还可将复杂的投资问题拆解为清晰的分析步骤，深入理解市场逻辑。 此外，DeepSeek赋能效率提升，成本优势凸显，尤其对人力和技术资源相对有限的中小金融机构更为友好。相比较其他AI工具，DeepSeek性价比较高，再加上开源策略，金融机构可根据自身需求进行定制和优化，降低对外部供应商的依赖。当前DeepSeek在金融领域的应用场景较为广泛，覆盖智能服务、智能投研、风险管理、文档处理、供应链金融等诸多核心业务场景。

部分金融企业DeepSeek部署情况

	01	02	03	04	05
应用方向	智能服务	智能投研	风险管理	文档处理	供应链金融
主要场景	数据分析、策略生成、风险评估	违约识别、欺诈检测、合规监控	智能客服、个性化推荐、需求分析	报告生成、合同审核、文档分类	信用评估、风险监控、融资服务
技术特点	时间序列模型、NLP技术、强化学习	LSTM模型、图神经网络、异常检测	多模态交互、用户画像、知识图谱	预训练模型、语义匹配、规则引擎	ERP数据分析、动态评分、实时监控
优势	提升研究效率和决策准确性	降低风险事件发生率	提升客户满意度和响应效率	提升处理效率	优化了供应链金融业务流程
布局者	邮储银行、重庆银行等	国泰君安、汇添富基金等	江苏银行、苏商银行等	兴业证券、北京银行等	联易融科技集团等

多地将DeepSeek应用于政务系统

作为日常生活中与公众交互最密切、最频繁的场景之一，政务服务与人工智能大模型在信息收集、文本总结、智能交互等方面的能力高度契合。近日，多地宣布，已将DeepSeek应用于政务系统，面向用户开展应用。当前，数字政府建设已进入深化提质阶段，政务应用与人工智能结合，或将成为未来重要发展趋势。 DeepSeek不仅在内容生成、智能交互等方面提升办公效率，还能够与政务系统深度融合，助力城市治理和公共服务升级，从而推动政府从传统管理模式向智能化、数字化管理模式的转变。

政府部门DeepSeek部署情况

昌吉回族自治州上线全疆首个基于DeepSeek—R1智慧政务大模型

乌海市海南区成功完成 DeepSeek 人工智能平台的本地化部署

呼和浩特市大数据管理局组织内蒙古大数据产业发展集团技术团队，完成了DeepSeek-R1在政务外网的本地化部署

昌都市经济和信息化局(数据管理局)携手中电信昌都分公司完成DeepSeek-R1-70B模型在市大数据中心政务云上的部署

北京市丰台区政务和数据局完成了政务云本地部署DeepSeek大模型环境，上线了“丰小政”数智助手

山东烟台市大数据局成功完成了国产大模型DeepSeek的本地化部署

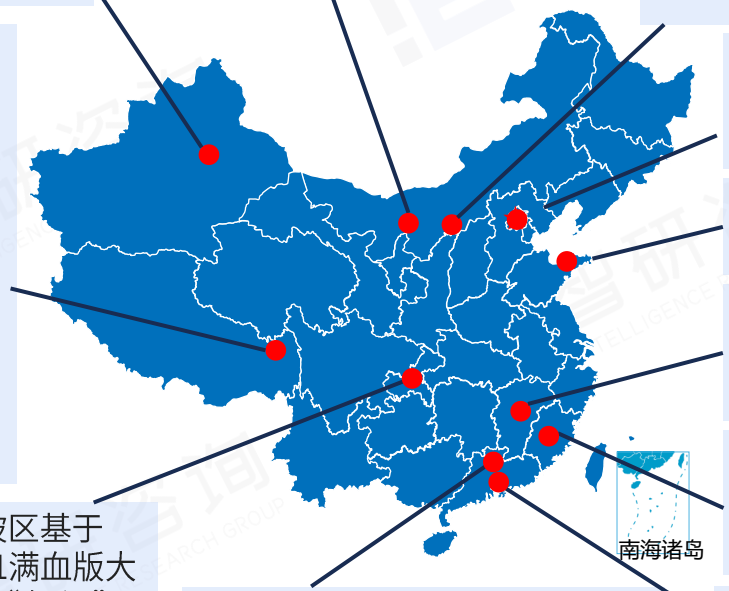
江西赣州市已完成政务环境DeepSeek系列大模型部署，成为全省首个成功部署上线该大模型的地级市

泉州市德化县完成基于国产大模型DeepSeek的政务审批信息智能导询系统本地化部署

重庆九龙坡区基于DeepSeek R1满血版大模型建设的“九i问”在区政府官网上线

广州市政务服务和数据管理局在政务外网正式部署上线DeepSeek-R1、V3 671B大模型

深圳市基于政务云环境面向全市各区各部门正式提供DeepSeek模型应用服务



DeepSeek大模型让政务工作提质增效,政务智能体也赋能DeepSeek,丰富了DeepSeek的应用场景和训练资源,让模型更“聪明”。政务服务与人工智能的双向赋能,形成了一个互利共赢的循环。

资料来源：智研咨询整理 www.chyxx.com

PART 06

AI大模型市场现状

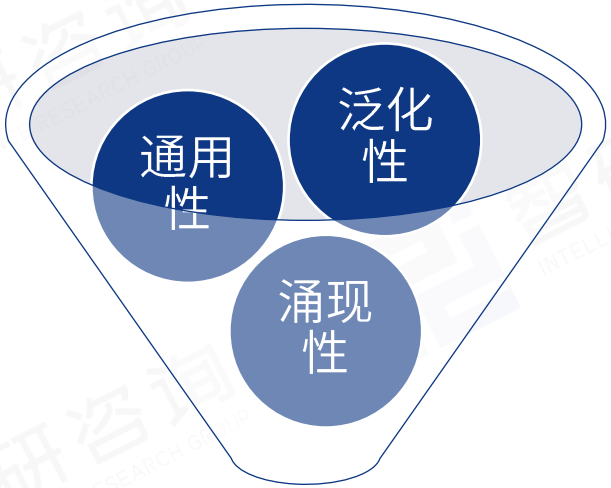
最全面的产业分析 ● 可预见的行业趋势

AI大模型被视作通往通用人工智能的重要路径

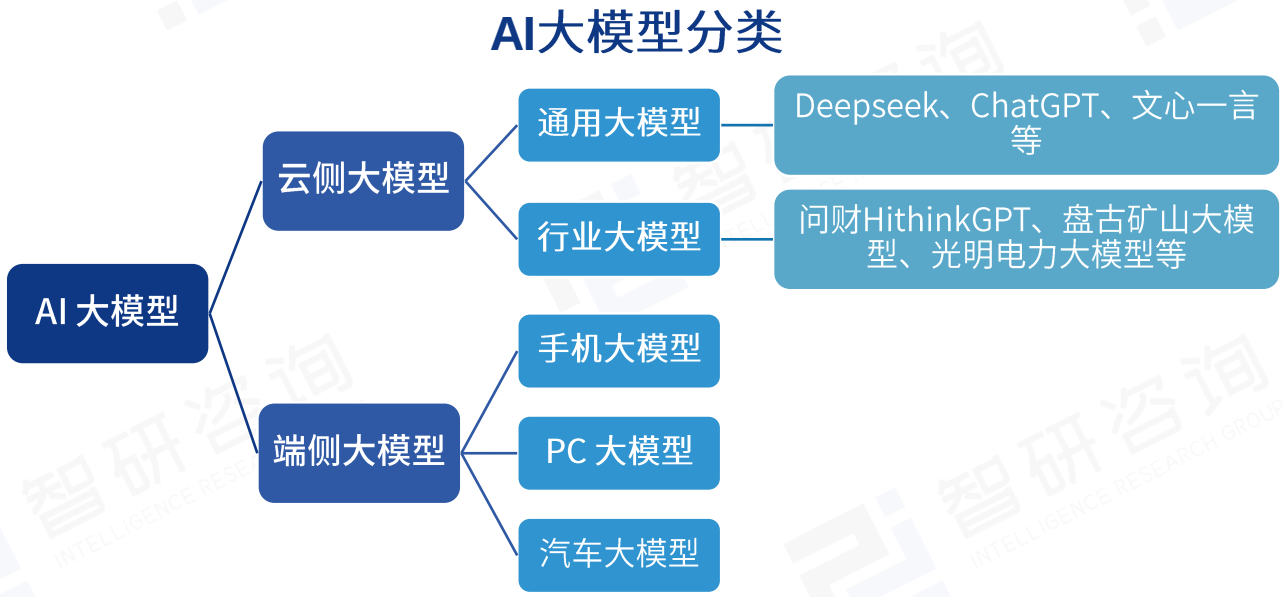
2022 年底，由OpenAI 发布的语言大模型 ChatGPT引发了社会的广泛关注。在“大模型+大数据+大算力”的加持下，ChatGPT能够通过自然语言交互完成多种任务，具备了多场景、多用途、跨学科的任务处理能力。以 ChatGPT 为代表的大模型技术可以在经济、法律、社会等众多领域发挥重要作用。大模型被认为很可能像PC时代的操作系统一样，成为未来人工智能领域的关键基础设施，引发了大模型的发展热潮。

AI大模型具有泛化性（知识迁移到新领域）、通用性（不局限于特定领域）以及涌现性（产生预料之外的新能力）特征。以 ChatGPT 为代表的 AI 大模型因其具有巨量参数和深度网络结构，能学习并理解更多的特征和模式，从而在处理复杂任务时展现强大的自然语言理解、意图识别、推理、内容生成等能力，同时具有通用问题求解能力，被视作通往通用人工智能的重要路径。

按照部署方式划分，AI大模型主要分为云侧大模型和端侧大模型两类，云侧大模型分为通用大模型和行业大模型，端侧大模型主要有手机大模型、PC 大模型等。



AI大模型的三大特征



资料来源：智研咨询整理 www.chyxx.com

2024年中国AI大模型商业发展加速

1956-2006年，深度学习和神经网络技术的提出和发展，为AI大模型的出现奠定了技术基础，大模型技术萌芽；2006年后自然语言处理技术、Transformer架构的发展，为大模型预训练算法技术和架构奠定了基础；2018年OpenAI和Google分别发布GPT-1与BERT，预训练大模型成为自然语言处理领域的主流；2022年底，OpenAI推出ChatGPT引发全球大模型发展热潮，2023年中国国内大模型训练开始井喷，出现“百模大战”现象；2024年中国政策加大行业落地推动力度，商业发展加速。

AI大模型发展历程

1. 萌芽期（1950-2005）

- 1956年：计算机专家约翰·麦卡锡首次提出“人工智能”概念，标志着AI领域的诞生。
- 1980年：卷积神经网络（CNN）的雏形诞生，为后续的深度学习奠定了基础。
- 1998年：LeNet-5的出现，标志着机器学习从浅层模型向深度学习模型的转变，为自然语言处理和计算机视觉等领域的研究奠定了基础。

2. 探索沉淀期（2006-2019）

- 2013年：Word2Vec模型的诞生，首次提出将单词转换为向量的“词向量模型”，极大地推动了自然语言处理技术的发展。
- 2014年：对抗式生成网络（GAN）的诞生，标志着深度学习进入了生成模型研究的新阶段。
- 2017年：Google提出了基于自注意力机制的Transformer架构，为大模型的预训练算法架构奠定了基础。
- 2018年：OpenAI和Google分别发布了GPT-1与BERT，标志着预训练大模型成为自然语言处理领域的主流。

3. 迅猛发展期（2020-至今）

- 2020年：OpenAI推出了GPT-3，模型参数规模达到1750亿，成为当时最大的语言模型，并在零样本学习任务上实现了巨大性能提升。
- 2022年11月：搭载了GPT-3.5的ChatGPT发布，以其逼真的自然语言交互和多场景内容生成能力，迅速成为互联网上的热门话题。
- 2023年3月：GPT-4的发布，这是一个超大规模的多模态预训练大模型，具备了多模态理解与多类型内容生成能力，标志着大数据、大算力和大算法的完美结合，大幅提升了大模型的预训练和生成能力。2023年，中国掀起“百模大战”，发布各类大模型数量超过100个，涵盖通用大模型、行业大模型等。
- 2024年中国政策加大行业落地推动力度，商业发展加速；同时，大模型产品使用价格进一步下降，为大模型广泛商用落地提供了基础。

国家和地方各级政府对AI大模型的创新发展给予了有力支持

在政策层面，国家和地方各级政府对AI大模型的创新发展给予了有力支持，推动传统产业数字化转型。近年来，我国始终高度重视人工智能发展机遇和顶层设计，发布多项人工智能支持政策，国务院于2017年发布《新一代人工智能发展规划》。科技部等六部门也于2022年7月印发《关于加快场景创新 以人工智能高水平应用促进经济高质量发展的指导意见》对规划进行落实。伴随人工智能领域中大模型技术的快速发展，我国各地方政府相继出台相关支持政策，加快大模型产业的持续发展。

中央地区AI大模型相关政策

- 2022年7月，科技部等六部门发布《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》

场景创新成为人工智能技术升级、产业增长的新路径，场景创新成果持续涌现，推动新一代人工智能发展上水平。

- 2024年1月，十七部门印发了《“数据要素×”三年行动计划（2024—2026年）》

以科学数据支持大模型开发，深入挖掘各类科学数据和科技文献，通过细粒度知识抽取和多来源知识融合，构建科学知识资源底座，建设高质量语料库和基础科学数据集，支持开展人工智能大模型开发和训练。

- 2024年7月，四部门发布了《国家人工智能产业综合标准化体系建设指南（2024版）》

规范大模型训练、推理、部署等环节的技术要求，包括大模型通用技术要求、评测指标与方法、服务能力成熟度评估、生成内容评价等标准。

- 2024年12月，四部门发布《中小企业数字化赋能专项行动方案（2025—2027年）》

建设一批适用于中小企业的垂直行业大模型，强化中小企业大模型技术产品供给。

地区AI大模型相关政策

2023年8月，成都市发布《成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施》

2024年10月，河南省发布《河南省推动“人工智能+”行动计划（2024—2026年）》

2024年7月，北京市发布《北京市推动“人工智能+”行动计划（2024-2025年）》

2024年6月，山东省发布《关于加快大模型产业高质量发展的指导意见》

2025年2月，武汉市发布了《武汉市促进人工智能产业发展若干政策措施》

2023年11月，上海市发布《上海市推动人工智能大模型创新发展若干措施（2023-2025年）》

2025年1月，贵州省发布《贵州省推动人工智能高质量发展行动方案（2025—2027年）》

2024年9月，湖南省发布《湖南省人工智能产业发展三年行动计划（2024-2026年）》

2024年5月，广东省发布《广东省关于人工智能赋能千行百业的若干措施》

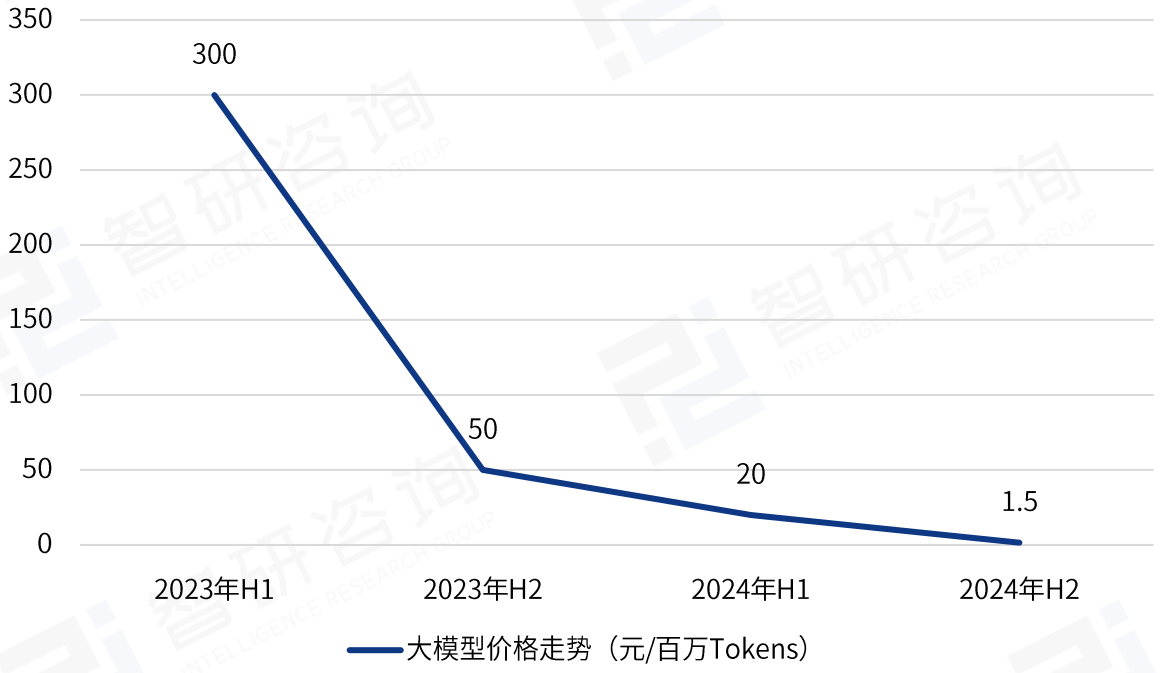
资料来源：智研咨询整理 www.chyxx.com

AI大模型应用规模不断壮大

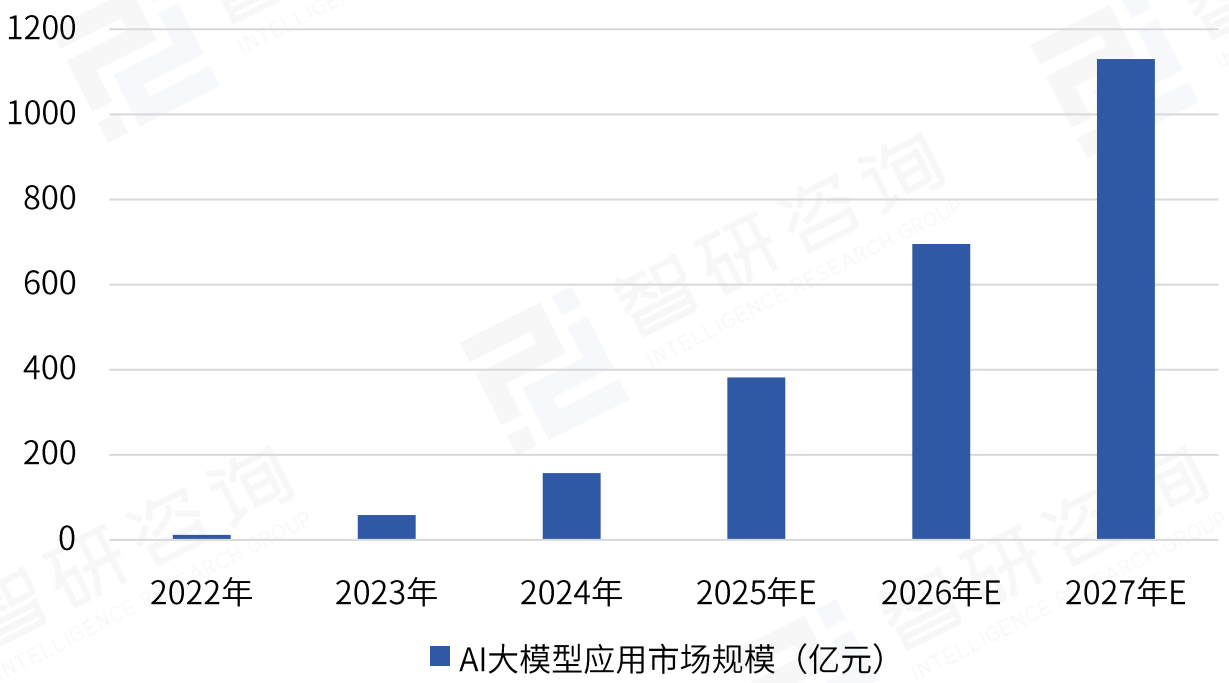
价格方面，中国大模型价格下降趋势仍在继续。截止到2024年底，我国典型AI大模型的输入价格下降至1.5元/百万Tokens以内。再到DeepSeek的横空出世，一度将百万Tokens的输入价格拉进“毛时代”。

应用方面，随着大模型技术成本的持续下降和应用场景的不断拓展，AI大模型正迎来从高门槛专业技术向大众化、普惠化转变的关键节点，应用规模持续壮大，2022-2027年中国AI大模型应用市场规模复合增长率将达到148%。2024年中国AI大模型应用市场规模达157亿元，预计到2027年市场规模将超1100亿元。

2023-2024年中国国产AI大模型价格走势



2022-2027年中国AI大模型应用市场规模（亿元）

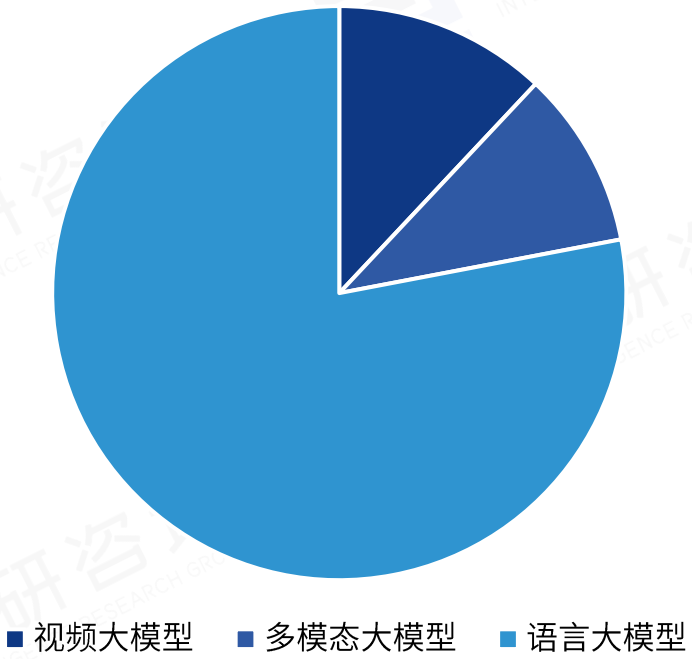


资料来源：智研咨询整理 www.chyxx.com

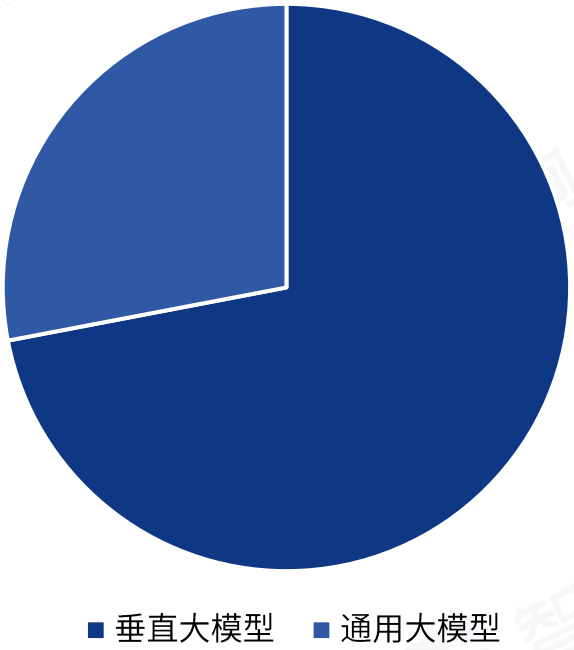
语言大模型为市场主流

为促进生成式人工智能服务创新发展和规范应用，2024年，网信部门会同有关部门按照《生成式人工智能服务管理暂行办法》要求，持续开展生成式人工智能服务备案工作。截至2024年12月31日，共302款生成式人工智能服务在国家网信办完成备案，其中2024年新增238款备案；对于通过API接口或其他方式直接调用已备案模型能力的生成式人工智能应用或功能，2024年共105款生成式人工智能应用或功能在地方网信办完成登记。从模态结构来看，语言大模型为市场主流，占比78%；从类型结构来看，通用大模型占比28%，垂直大模型占比72%。

模态结构



类型结构



资料来源：智研咨询整理 www.chyxx.com

AI大模型架构不断完善

AI大模型架构包括基础设施层、模型层、应用技术层、应用层。基础设施层包括GPU、CPU、存储和网络等硬件设施。这些硬件设备为AI大模型的训练与推理提供了关键的运算资源和存储能力。模型层包含各种AI大模型，如大语言模型、视觉-语言模型等，具备强大的学习和推理能力。应用技术层包括Agent智能体技术、检索增强生成技术、大模型微调、提示词工程、思维链技术等。这些技术利用大模型的推理能力对任务进行规划拆解，并使用外部工具完成复杂任务。应用层展示了AI大模型在具体场景中的应用，如增强检索类应用、智能体类应用、事务处理类应用等。

AI大模型架构图

应用层	增强检索类应用	智能体类应用	事务处理类应用	分析决策类应用	
应用技术层	智能体	检索增强生成	微调	数据获取	数据向量
		提示词工程	思维链	数据清洗	访问控制
模型层	大语言模型	语言-视觉大模型	语音-语言大模型	召回、排序模型	文本理解模型
		语音-视觉大模型	图片识别模型	多模态监测与分割大模型	
基础设施层	GPU	CPU	存储	网络	算力调度

大模型应用需求落地一般分为四个阶段

大模型应用需求落地一般分为四个阶段：1、场景需求评估：评估企业当前的大模型技术、应用场景和能力，做好大模型应用落地的准备，包括技术能力评估、应用场景梳理、能力分析等。2、部署能力建设：设计和构建符合战略规划和业务需求的大模型能力体系，包括大模型建设方案设计、系统研发和功能测试、数据与算法准备等。3、大模型应用部署：将大模型部署到具体的业务场景中，提供定制化的智能解决方案，实现大模型的商业化应用，包括定制化优化与应用开发、效能评估与闭环管理、全生命周期管理等。4、大模型运营管理：建立大模型运营管理体系，保障大模型的长效运行，并通过实时监测和反馈机制提升运营效率，包括实时监测与动态追踪、持续优化与管理体系完善等。

AI大模型应用部署方式

全栈构建（定制大模型）	定制模型	扩展应用	能力嵌入	直接调用
应用软件	应用软件	应用软件	应用软件	应用软件
大模型工具链（数据检索、提示词工程等）	大模型工具链（数据检索、提示词工程等）	大模型工具链（数据检索、提示词工程等）	大模型工具链（数据检索、提示词工程等）	大模型工具链（数据检索、提示词工程等）
AI大模型（基础模型及微调）	AI大模型（基础模型及微调）	AI大模型（基础模型及微调）	AI大模型（基础模型及微调）	AI大模型（基础模型及微调）
算力基础设施	算力基础设施	算力基础设施	算力基础设施	算力基设施

重量部署 ← → 轻量部署

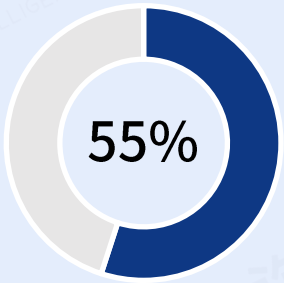
AI大模型市场应用的商业化模式逐渐清晰

定制化

本地化部署：软硬件一体、提供预训练和微调等服务（80%在B端）

云部署：提供算力服务和大模型预训练及微调服务

适用于党政、金融、能源、工业等行业的大中型企业和组织机构。如中国华能集团有限公司完成了DeepSeek系列模型的本地化部署

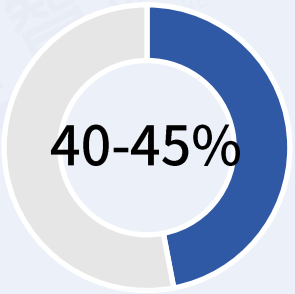


定制化模式面向大型政企：大型企业在应用AI大模型时，更倾向于定制化，并采用本地化部署模式。当前，大模型技术和产品达代迅速，定制化模式下，客户会要求大模型服务商定期迭代更新服务。

API及订阅

SaaS、PaaS、MaaS等方式调用服务，按流量、Tokens、产出内容、时间等方式计费

适用于电商、医疗、教育等中小型用户，如

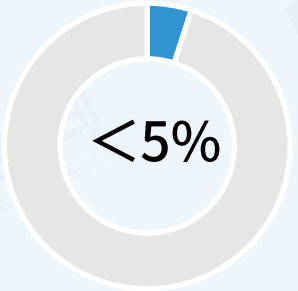


采用API及订阅模式，具有节省资源、快速集成、实时更新和可扩展性强等优势，适用于中小型企业。

广告

嵌入智能终端或APP，用户免费使用，向广告主收取广告费

适用于智能终端等面向大规模人群的ToC场景，如Kimi已开始涉足大模型的广告投放业务



随着大模型应用的普及，将大模型嵌入智能终端和APP中，向广告主收取广告费的模式，将成为大模型产品变现的重要方式，未来大模型应用将成为互联网广告提升和优化流量的重要抓手。

资料来源：智研咨询整理

PART 07

Deepseek对AI行业影响总结

最全面的产业分析 ● 可预见的行业趋势

对AI模型层：开源与价格优势将导致大模型层竞争加剧

打破已有过度依赖算力与标注数据的训练模式

打破已有过度依赖算力与标注数据的训练模式，显著降低大模型准入壁垒，利好模型追赶者。在DeepSeek-V3和R1模型之前，大模型行业信奉“算力即权力，规模即护城河”的逻辑，DeepSeek打破了这种传统路径依赖，展示了通过改进模型架构和训练方法，如大规模使用强化学习技术，即使在数据标注量少的情况下，也能极大提升模型推理能力。

为其他模型研发者提供了新的技术思路

架构上的“捷径”对于利用大算力与标注数据作为护城河的领先模型是巨大的挑战，为其他模型研发者提供了新的技术思路和追赶方式，预计将引发一波模仿、探索高效训练方法和创新模型架构，从而加速追赶的趋势。

促进开源生态发展

加剧大模型从能力、迭代周期到性价比全面竞争，促进开源生态发展。DeepSeek不仅主打高性价比还将模型全部开源，打破了闭源模型在性能和应用上的优势神话，促使更多人重新审视开源模型的价值。多家团队已宣布复现其训练过程，这将极大推动开源生态的繁荣，也意味着模型层竞争更加激烈，闭源模型不再拥有绝对优势，促使模型开发者不断提升模型性能、降低成本，以在市场中拥有更多客户和使用量。

对AI算力层：短期降低对先进算力需求预期

短期内缓解算力压力 01

DeepSeek通过创新的训练方法，如在预训练阶段加入强化学习，用较少的计算资源就达到了接近 GPT-o1的性能，这使业界开始反思大算力在AI发展尤其是大模型训练过程中的必要性，部分企业预计会减少对大规模算力基础设施的激进投入。短期内可能会局部缓解算力压力，但长期来看，随着AI能力的边界扩展（如多模态、复杂推理、通用人工智能）以及应用场景的爆发式扩展，算力需求仍将增长。

为国产显卡和ASIC芯片带来了机会 02

另一方面，也为国产显卡和ASIC芯片带来了机会。因为DeepSeek的RL策略对并行计算需求下降40%，这使得国产算力硬件有机会凭借成本和服务优势在市场中占据一席之地。客户可以根据实际应用场景灵活进行定制化芯片开发，算力市场预计走向多元化发展。

对云厂商：利好云厂商下游需求增长，有望进一步提升国产云厂商利润率

DeepSeek为云厂商提供了更高效、低成本的API调用方案/AI解决方案

目前云厂商自身集算力供给、大模型研发与AI应用为一体，DeepSeek 高性价比、开源模型的发布虽然削弱模型层竞争壁垒，加大AI云格局的不确定性，但为云厂商提供了更具性价比的AI方案。DeepSeek高性价比、开源模型的发布削弱了云厂商/大模型厂商在AI模型服务层面的壁垒，让大模型差距更小。但DeepSeek利好国内外大模型向OpenAI等一流模型追赶。同时，DeepSeek的高性价比开源模型为云厂商提供了更高效、低成本的API调用方案/AI解决方案，如R1上线短短两周，腾讯云、华为云、微软Azure和亚马逊AWS均已上线DeepSeek-R1相关服务，并提供了便捷的部署和调用方式。

Deepseek将加速企业数字化转型上云，规模效应下进一步提升云业务利润率

对国内云厂商，Deepseek将加速企业数字化转型上云，规模效应下进一步提升云业务利润率。AI背景下数字化和云化是必然的趋势，且AI云的技术壁垒、相关配套服务的利润空间和整体市场空间显著高于传统云，而Deepseek模型的出现加速了各行业的数字化转型进程。预计将带动国内云厂商利润率向海外云厂商靠拢。

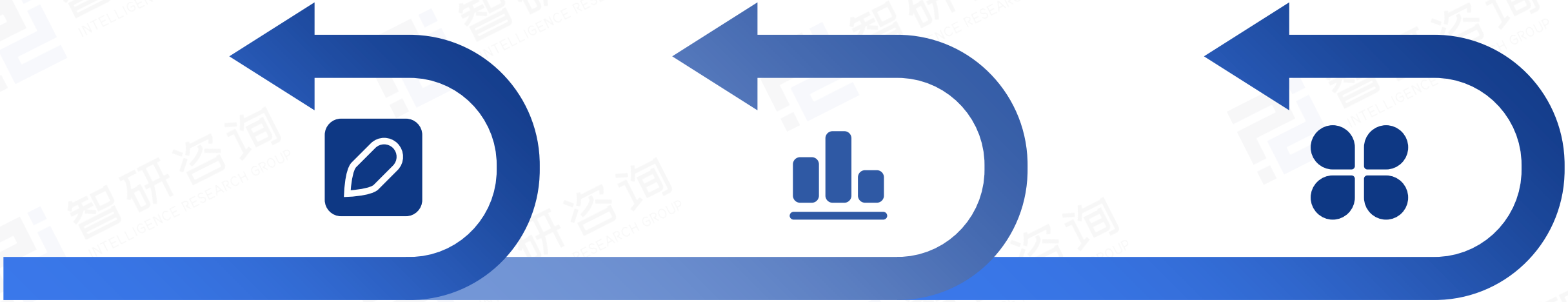
Deepseek缩小了云厂AI前期投入与应用兑现之间的时间与资源成本

Deepseek的技术路线使得云厂在高额前期投入的重压下有了喘息之机，更好地去评估AI板块的ROI，更加注重模型的成本效益和实用性，加大在模型部署、优化和管理的投入，加强对AI应用场景的拓展和落地。

Deepseek拓展AI应用场景，带动AI云增长

DeepSeek模型的普及将赋能更多应用场景，从而推动了云服务厂商的业务增长，云服务厂商既是技术降本的受益者，也是放大降本效应的推动者

对AI应用层：降低AI应用研发与落地的成本，加速AI应用发展



降低垂类模型/应用开发门槛，加速AI应用/Agent在各个场景落地

DeepSeek 模型的低成本优势使得开发利用大模型训练、调优的门槛降低，企业无需投入巨额资金用于模型训练就能获取高性能模型，加速垂类模型发展，利好AI在各行业的渗透，如医疗、教育等领域，催生出更多创新的AI应用场景和商业模式。且DeepSeek-R1具备深度思考和出色的推理能力、且成本低，有望成为互动场景或工作任务的 “Agent智能体”大脑，利于AI Agent在各个场景普及。

显著降低推理成本，提升应用端盈利能力

DeepSeek高性价比的模型使得AI应用研发和使用成本显著降低，从而提升企业盈利能力，应用厂商也可以有更多资源进行进行产品优化和市场拓展。

将同等模型能力所需的算力极度压缩，为AI端侧落地提供技术基础

Deepseek将同等模型能力所需的算力大幅压缩，模型提供的高性价比和高效推理能力使其能够更广泛地应用于端侧设备，预计将加速了端侧AI应用的落地。

新质生产力报告推荐

01 2025年中国HBM行业市场现状分析及未来趋势研判报告

02 2025年中国DDR5行业市场研究分析及产业需求研判报告

03 2025年中国智能算力行业发展现状及未来趋势分析报告

04 2025年中国端侧AI行业市场现状分析及发展趋势展望报告

05 2025年中国AIDC行业市场发展态势及产业需求研判报告

智研咨询领域优势

数据优势
Data advantages

拥有全国百万家企业基础数据库

权威渠道
Authoritative channel

我们的第三方数据渠道有国家统计局、国家海关、商务部、相关行业协会等权威机构

专业服务
professional services

全国各地分支网络和严格的调查控制流程，使我们有足够的知识和能力向客户提供高质量服务

成功案例
Success cases

超过200多个研究项目的成功案例

研究领域
Research field

研究领域覆盖能源、化工、机械、汽车、电子、医疗等诸多行业

全球客户
Global customers

我们很荣幸的为国内外知名企业和机构提供过咨询服务

产业研究报告

定制报告

可行性研究报告

商业计划书

（公众号）

（微信客服）

（智研小程序）

最全面的产业分析 • 可预见的行业趋势