DeepSeek 是什么

DeepSeek是杭州深度求索人工智能基础技术研究有限公司推出的一款创新大语言模型。公司成立于2023年7月17日，由知名私募巨头幻方量化孕育而生。DeepSeek致力于开发和应用先进的大语言模型技术

深度小助手　　聪明且低成本　　聪明强大能干　　中国本土AI

深度思考

联网搜索

DeepSeek：大语言模型的特点有哪些？

内容 token 化

大模型看到的世界与人看到的不太一样

训练前需要将文本进行处理，比如切割称为Token的基本单元；比如问ai 一个英文单词 illegal 中有几个字母l，有些指令模型回答为2个；

但deepseek r1 推理模型是可以回答正确！

模型训练 存在endtime

大模型训练语料存在一个截止时间

deepseek R1虽然是25年1月发布，但它的知识库截止日期是2023年12月，这就意味着ds可以提供在此日期发布之前的公开信息和常识；需要经过大量清洗、监督微调、反馈强化学习。但对于之后的新闻、事件变化、新事物则无法直接获取或验证。

解决办法是开启联网模式或提示词中 补充说明

无自我认识 无自我意识

网上有个段子是“有人问deepseek你是谁，然后回答是gpt”

目前AI 大模型不知道自己是谁，也不知道自己是采用什么模型。 除非是厂商在后期再微调、或再训练，如果大家问到类似的问题，可能目前的AI 大模型会回答错误。

解决办法是少问 AI是谁、采用什么模型

上下文长度限定 记忆力有限

AI 大模型目前的记忆力大概是64k ~ 128k

目前AI 大模型均有上下文长度限定；deepseek r1 提供64k token上下文长度，对应中文的话大约3万~4万字。目前还不能一次性投喂太长的文档给它，比如：一本完成西游记、或者非常长的文档让它翻译，AI 它是没有办法完整读完

解决办法是分成多次投喂

回答输出 长度有限

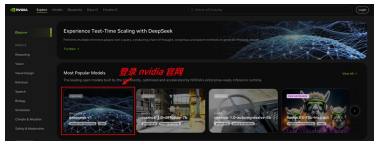
AI 大模型目前的回答 4k ~ 8k，2000~4000字

目前AI 大模型无法一次性完成万字长文，也无法一次性输出5千字，均是模型输出长度限制所致；如果是输出长文，可以尝试先让AI 大模型先生成一个目录，然后再根据目录输出对应模块；如果是长文翻译类，则多次输入，或者拆解后多次调用API

解决办法是将任务分解成多次

DeepSeek 发展由来

布局集卡	投入10亿 手握万卡	GPT 3.5发布	V2 发布 GPT 4o发布	Llama-3.1发布	V3 发布	R1 登录nvidia 官网
2019年	2020年	2022年3月	2024年5月	2024年7月	2024年底	2025年1月31号



补充1：ChatGPT需要上万张 NVIDIA A100显卡，国内主要玩家：百度、字节、腾讯、阿里、商汤、幻方
补充2：nvidia官网 https://build.nvidia.com/explore/discover

来自杭州深度求索公司，系一家成立于2023年。使用数据蒸馏技术，得到更为精炼、有用的数据。由知名私募巨头幻方量化全资孕育而生，专注于开发先进的大语言模型〔LLM〕和相关技术。

DeepSeek为什么火：一个足够优秀的模型变得人人免费拥有

一、技术突破：为什么DeepSeek的模型值得关注？

1. 模型架构与训练效率优化

架构改进：MLA多层注意力架构、FP8混合精度训练框架、DualPipe 跨节点通信
训练策略：采用混合精度训练〔BF16+FP8〕和梯度累积策略

2. 数据质量与领域适配

数据筛选：多模态数据清洗
领域微调："领域渐进式微调"〔Progressive Domain Fine-tuning〕策略

二、开源生态：DeepSeek如何改变开发者社区？

1. 开放模型与工具链

全量开源：DeepSeek开源了完整训练代码、数据清洗Pipeline和领域微调工具包〔如DeepSeek-Tuner〕，极大降低复现和二次开发门槛
轻量化部署：提供模型压缩工具〔如4-bit量化适配TensorRT-LLM〕

2. 社区驱动创新

开发者基于DeepSeek模型快速构建垂直应用
- 金融场景
- 教育场景

三、行业落地：DeepSeek推动的技术范式迁移

1. 从"通用模型"到"领域专家"

传统大模型〔如GPT-3.5〕依赖Prompt Engineering适配行业需求，而DeepSeek通过预训练阶段嵌入领域知识，减少后期微调成本

2. 成本革命

通过模型压缩和高效推理框架，企业可基于单卡部署专业模型，推理成本降至GPT-4 API的1/50

如：某电商客服系统用DeepSeek-7B替代GPT-4，单次交互成本从0.06降至0.001，日均处理量提升10倍。

- 在权威评测集（如MMLU、C-Eval、HumanEval）中，DeepSeek模型在同等参数规模下显著超越主流开源模型：

模型（7B）	MMLU（5-shot）	HumanEval（代码）	C-Eval（中文）
Llama 2	45.2%	12.5%	32.1%
DeepSeek	**53.8%**	**26.7%**	**48.5%**

- 在金融、医疗等垂类评测中（如FinBench、MedMCQA），DeepSeek的领域模型表现接近GPT-4水平。

四、行业竞争格局：DeepSeek的"鲶鱼效应"

1. 倒逼闭源模型降价

DeepSeek的开源策略迫使国际厂商调整定价。例如，Anthropic的Claude 3 Sonnet API价格在DeepSeek开源后下调

2. 催化国产AI芯片生态

DeepSeek与华为昇腾、寒武纪等厂商深度合作，优化模型在国产硬件的推理性能。例如，DeepSeek-7B在昇腾910上的吞吐量比A100高

3. 推动AGI技术民主化

中小企业和研究机构可基于开源模型快速迭代，无需依赖巨头API。例如，非洲某初创团队用DeepSeek-7B开发本地化农业咨询AI，成本仅为GPT-4方案的1/20

挑战及未来

技术挑战

1. 长上下文理解：目前最大支持32K tokens，相比Claude 100K仍有差距。
2. 多模态扩展：尚未开放图像-文本联合模型，需追赶GPT-4V、Gemini。

商业化平衡

开源模型可能导致企业版变现困难，需探索类似Red Hat的"开源+服务"模式。

DeepSeek 核心哪些创新大幅降低训练成本

Point1:
大大压缩计算量

Point2:
分布式并行提效

Point3:
模型大、数全、偏科

MLA多层注意力架构
- 原先：每一层有值且内存挨个计算
- 优化：前后合并，使用时再放到内存中

FP8混合精度训练框架
- 原先：32位、16位
- 优化：
  - 不该精确8位(近似值)，需要精确还是32位
  - 每128个位，交给会计总账合计保证精度

DualPipe 跨节点通信
- 原先：需要等前面stage完成才能干活
- 优化：优化为双路计算流水线，传输、计算同时进行
  - 计算+50%、传输+20%

无辅助损失的负载均衡策略
- 原先：每个worker干活一样
- 优化：均衡派单，保证worker有活干

跨节点全对全通信内核

模型够大参数多
- Llama3.1：405 B
- Deepseek：671 B

数据全且精
- 优化：精选数据、清洗干净

MTP技术（Multi-Token Prediction 多令牌预测）
- 传统：一次预测一个Token
- 优化：预测连续多个Token

R1蒸馏技术
- R1推理模型，给出计算逻辑推理
- V3提取推理思路+解题策略
- 用大模型指导小模型训练，降低推理成本

DeepSeek 核心技术架构

1. 模型架构

多模态深度Transformer：支持文本、代码、数学符号的统一理解与生成

动态稀疏激活机制：采用MoE（Mixture of Experts）架构，实现万亿参数级高效推理

2. 核心技术突破

超长上下文建模：支持128K+ tokens窗口，精准捕捉长程依赖

自研训练框架：融合高效分布式训练、混合精度优化与灾难性遗忘抑制技术

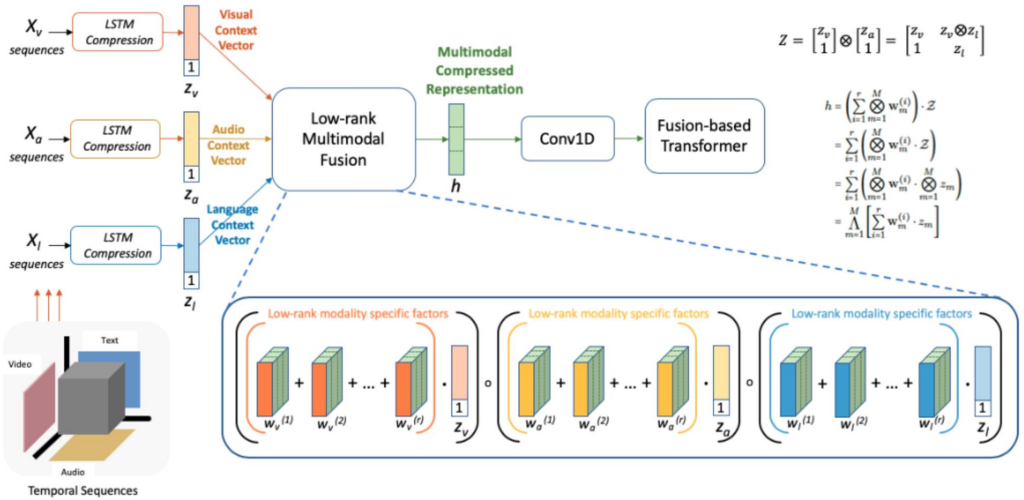
强化学习对齐：基于人类反馈的强化学习（RLHF），提升结果安全性与实用性

3. 核心优势

高效推理：单卡支持千亿参数模型部署，推理速度提升3倍+

多任务兼容：原生支持智能体（Agent）架构，实现工具调用与复杂推理

持续进化：支持参数高效微调（PEFT），快速适配垂直领域需求



4. 应用场景

智能问答 | 代码生成 | 数据分析 | 科研计算 | 多模态交互

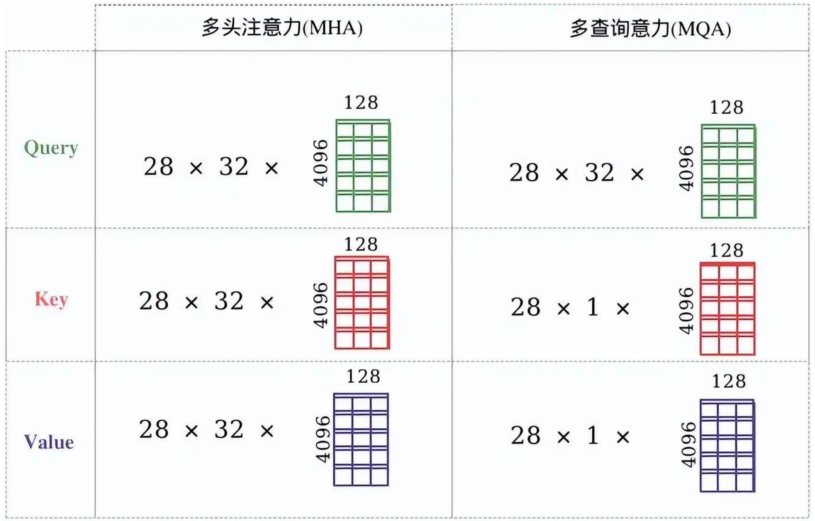
传统 VS MoE架构

特性	稠密模型（Dense）	MoE架构
参数利用率	全参数参与计算	稀疏激活（仅调用部分专家）
扩展方式	增加模型深度/宽度	横向增加专家数量
计算效率	计算成本与参数规模线性增长	计算成本与激活专家数相关
典型场景	中小规模通用任务	超大规模多任务/多模态场景

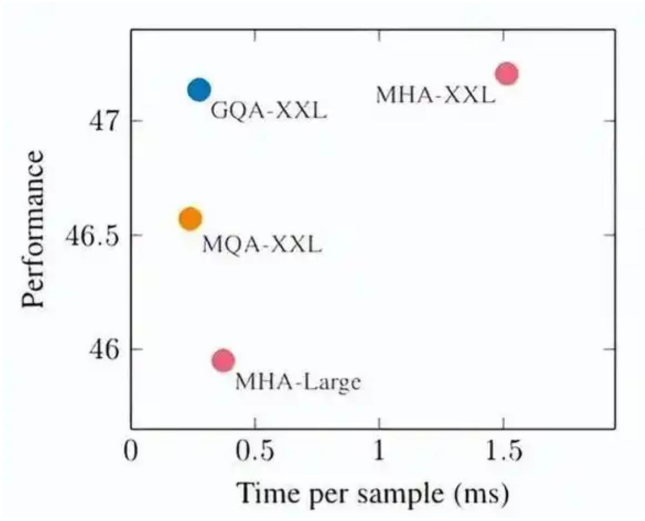
DeepSeek的核心版本（不同版本功能亮点）

——持续迭代的工程与创新

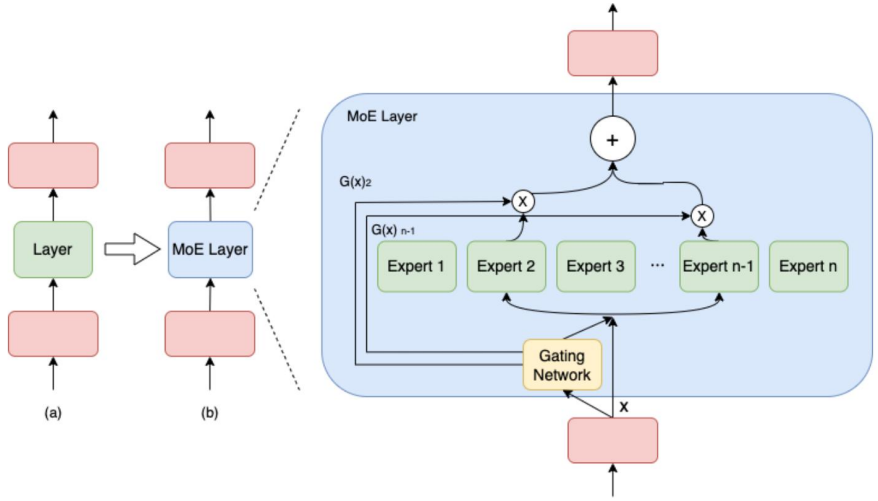
版本	原理	重要功能	特点	关键指标提升
DeepSeek V1	将多头查询（Q）分组共享键值（K/V），减少显存占用	基本沿用LLaMA	奠定基础 GQA + 多阶段训练	训练速度+20%
DeepSeek V2	在潜在空间压缩注意力头维度（如64维→32维），通过低秩分解减少计算量	• 提出DeepSeek MoE • MLA压缩kv减少缓存	效率革命 MoE + 潜在注意力	推理成本-50%
DeepSeek V3	熵最大化路由：约束路由器输出的熵值，自然分散专家负载 梯度掩码：对过载专家暂停梯度更新，促使其"冷却"	• MoE 负载均衡优化 • 引入MTP 技术	负载均衡新范式 无辅助损失均衡	专家利用率+24%
DeepSeek R1	动态路由架构：根据输入类型（文本/代码/数学）自动切换模型分支 混合精度推理：FP16用于注意力计算，INT4用于FFN层，延迟降低35%	冷启动问题的强化学习	全能选手 动态路由 + 混合精度	综合任务得分+15%



MHA和MQA的原理差异



GQA和MQA优化后和原始模型推理速度对比



MoE 原理图

参考链接：http://www.360doc.com/content/24/0203/15/170868_1113191680.shtml

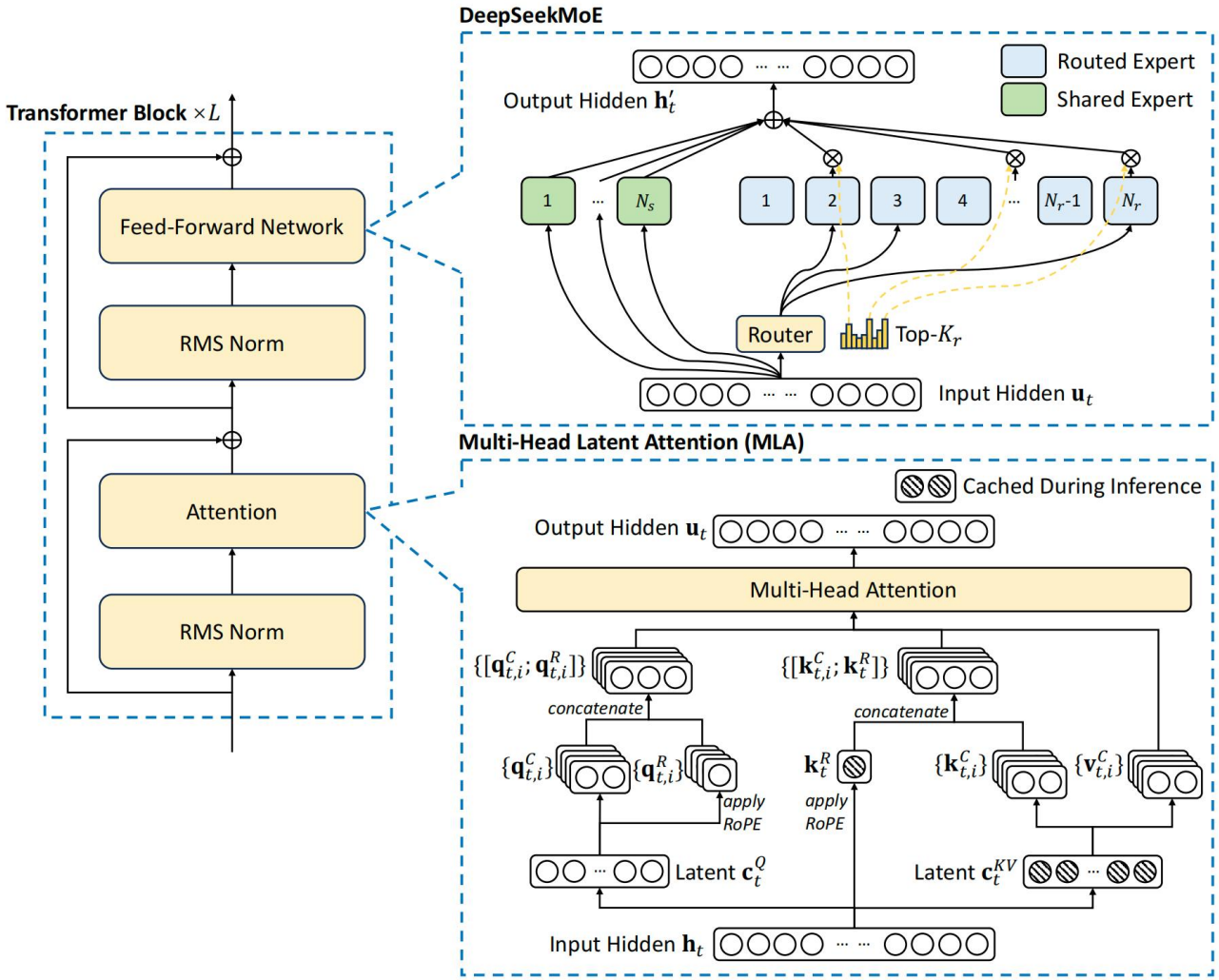
DeepSeek的核心技术（架构示意图）

——持续迭代的工程与创新

对Transformer框架内的注意力模块和前馈网络（FFNs）进行了优化，采用了我们提出的**多头潜在注意力（MLA）**和**DeepSeekMoE技术**。

- 在注意力机制方面，多头注意力（MHA）的键值（KV）缓存对大型语言模型（LLMs）的推理效率构成了重大障碍。人们已经探索了多种方法来解决这个问题，包括分组查询注意力（GQA）和多查询注意力（MQA）。然而，这些方法在试图减少KV缓存时，往往会牺牲性能。为了实现两全其美，我们引入了MLA，这是一种具备低秩键值联合压缩功能的注意力机制。实证研究表明，MLA相较于MHA具有更优的性能，同时显著减少了推理过程中的KV缓存，从而提高了推理效率。

- 对于前馈网络（FFNs），我们采用了DeepSeekMoE架构，该架构采用细粒度的专家分割和共享专家隔离策略，以实现更高的专家专业化潜力。与传统的MoE架构相比，DeepSeekMoE架构具有显著优势，使我们能够以较低的成本训练出强大的模型。由于我们在训练过程中采用了专家并行策略，还设计了补充机制来控制通信开销并确保负载均衡。



参考文章：《DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model》

DeepSeek的核心技术 -- MLA 减少kv 缓存占用空间

——MLA改进MHA，从而压缩KV缓存，提高推理速度

MLA诞生背景：

传统的Transformer模型通常采用多头注意力〔MHA〕，但在生成过程中，其庞大的键值〔KV〕缓存会成为限制推理效率的瓶颈。为了减少KV缓存，提出了多查询注意力〔MQA〕和分组查询注意力〔GQA〕它们需要的KV缓存规模较小，但性能不及MHA。

配备多头注意力〔MHA〕、分组查询注意力〔GQA〕和多查询注意力〔MQA〕的70亿参数密集型模型在四个困难基准测试上的评估结果

Benchmark (Metric)	# Shots	Dense 7B w/ MQA	Dense 7B w/ GQA (8 Groups)	Dense 7B w/ MHA
# Params	-	7.1B	6.9B	6.9B
BBH (EM)	3-shot	33.2	35.6	**37.0**
MMLU (Acc.)	5-shot	37.9	41.2	**45.2**
C-Eval (Acc.)	5-shot	30.0	37.7	**42.9**
CMMLU (Acc.)	5-shot	34.6	38.4	**43.5**

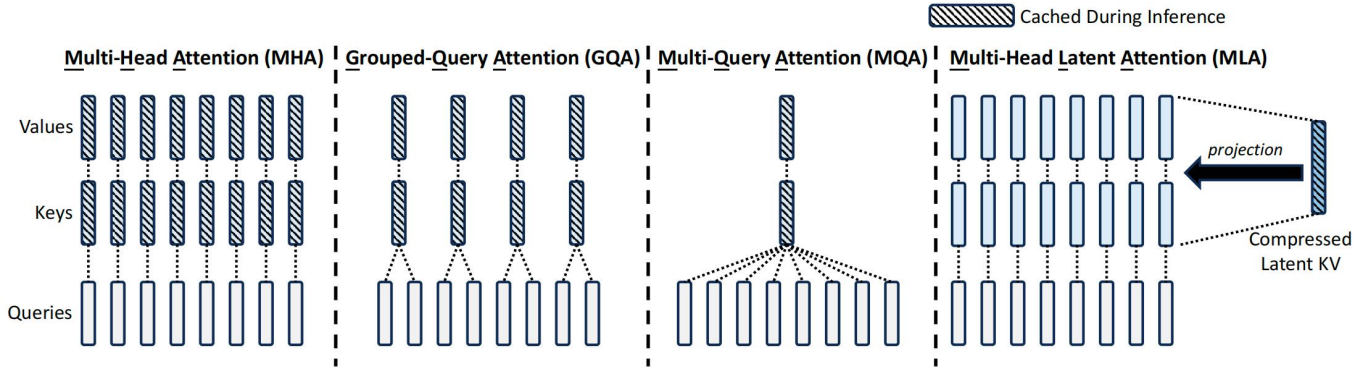
对于DeepSeek-V2，我们设计了一种创新的注意力机制，称为多头潜在注意力〔MLA〕。MLA配备了低秩键值联合压缩功能，其性能优于多头注意力〔MHA〕，但所需的键值缓存〔KV cache〕量显著减少。

不同注意力机制每个 token 的 KV 缓存比较

Attention Mechanism	KV Cache per Token (# Element)	Capability
Multi-Head Attention (MHA)	$2n_h d_h l$	Strong
Grouped-Query Attention (GQA)	$2n_g d_h l$	Moderate
Multi-Query Attention (MQA)	$2d_h l$	Weak
MLA (Ours)	$(d_c + d_h^R)l \approx \frac{9}{2} d_h l$	Stronger

多头注意力〔MHA〕、分组查询注意力〔GQA〕、多查询注意力〔MQA〕和多头潜在注意力〔MLA〕的简化示意图

通过将键和值联合压缩到一个潜在向量中，MLA在推理过程中显著减少了键值缓存〔KV cache〕



在困难基准测试中，MLA与MHA的比较

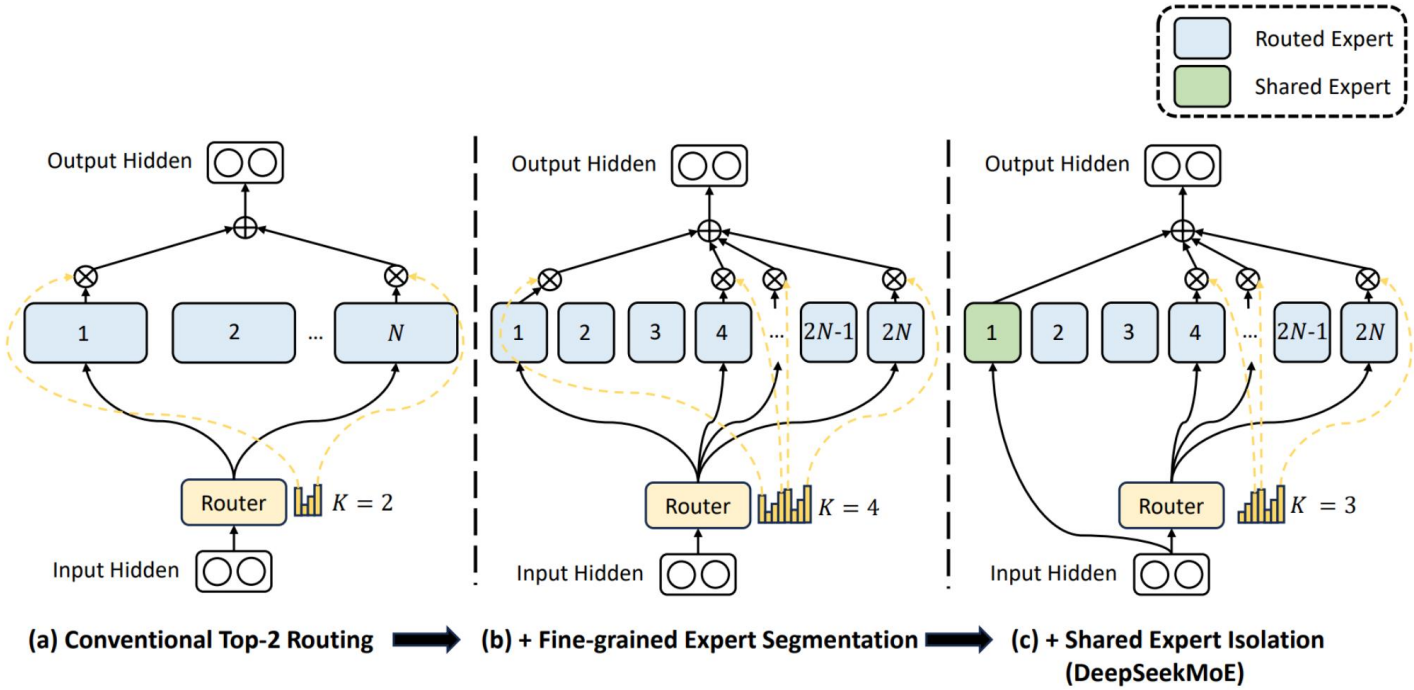
DeepSeek-V2的MLA性能优于MHA，但所需的键值缓存〔KV cache〕量显著减少

Benchmark (Metric)	# Shots	Small MoE w/ MHA	Small MoE w/ MLA	Large MoE w/ MHA	Large MoE w/ MLA
# Activated Params	-	2.5B	2.4B	25.0B	21.5B
# Total Params	-	15.8B	15.7B	250.8B	247.4B
KV Cache per Token (# Element)	-	110.6K	15.6K	860.2K	34.6K
BBH (EM)	3-shot	37.9	**39.0**	46.6	**50.7**
MMLU (Acc.)	5-shot	48.7	**50.0**	57.5	**59.0**
C-Eval (Acc.)	5-shot	**51.6**	50.9	57.9	**59.2**
CMMLU (Acc.)	5-shot	52.3	**53.4**	60.7	**62.5**

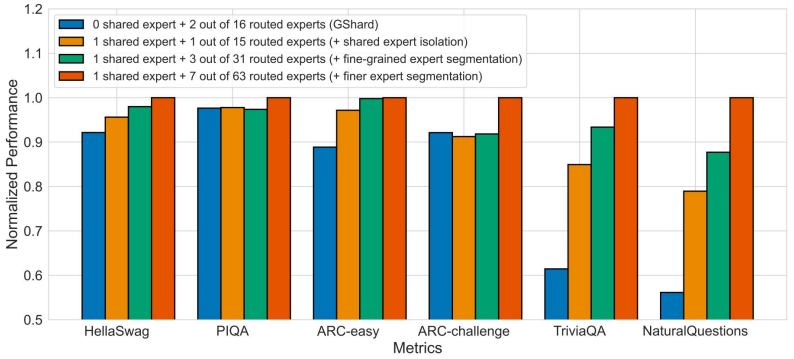
参考文章：《DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model》

DeepSeek的核心技术 -- DeepSeekMoE细粒度分割与共享隔离

——细粒度expert分割，优化路由，多级别负载均衡，提升模型性能

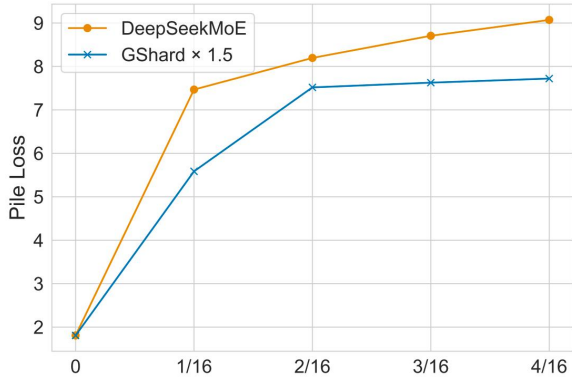


DeepSeekMoE 的共享 expert性能研究



为清晰展示，性能以最佳表现为基准进行了归一化处理。所有对比模型的参数数量和激活参数数量均相同。发现：细粒度的expert分割和共享expert隔离都有助于提升整体性能

禁用Top路由不同比例expert的损失数据



在不同禁用Top路由Expert比例下的堆叠损失。值得注意的是，DeepSeekMoE对禁用Top路由Expert的比例更为敏感，这表明DeepSeekMoE中被路由Eepert之间的冗余度较低。

基础 MoE

基础的MoE将原来的每个token的单个FFN层变成多个并行的FFN层〔对应多个expert〕，并根据输入生成一个路由到各个FFN的打分，选择topN个Expert，实现在单token运算量没有显著提升的前提下，扩大模型的参数空间的目的。

VS

DeepSeek MoE

DeepSeekMoE相比MoE有2个核心优化

1、FFN维度调小，增加Expert数量 〔 细粒度的Expert分割 〕

将expert细分到更细的粒度，以实现更高的专家专业化程度和更准确的知识获取

1、增加提取公用Expert并共享化，其它Expert专注于差异化

增加了几个所有token都走的公用Expert，通过让所有token都走这些Expert，让这些Expert提取通用信息，隔离一些共享expert，以减轻路由专家之间的知识冗余，其他Expert就能更专注于提取差异化的信息。

参考链接1: 《DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models》
参考链接2: 《DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model》

DeepSeek的核心技术 -- Multi-Token 预测（MTP）

——MTP 一次预测多个token，训练更长更多数据，提升大模型的训练和推理效率

token-by-token生成序列
主流大模型

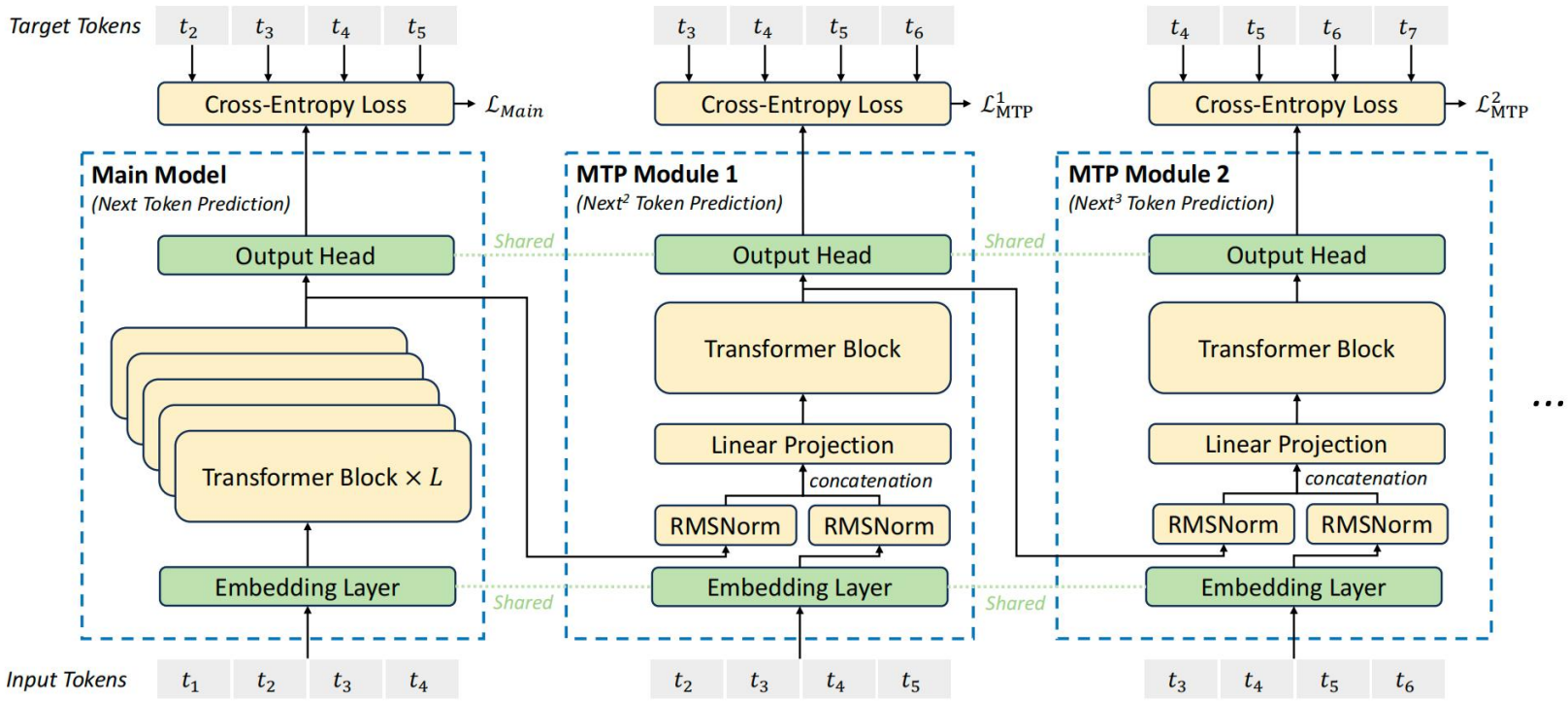
主流大模型token-by-token生成序列，而每次token生成需要频繁与访存交互，从而因为访存效率形成训练或推理的瓶颈

VS

单token 优化为多token
MTP方法

MTP主要将单token的生成，转变成多token的生成，提升训练和推理的性能；

MTP使训练信号更加密集，可能会提高数据效率；还使模型预先规划，以便更好地预测未来的Token



多标记预测（MTP）实现的示意图，保留每个深度的每个标记预测的完整因果链

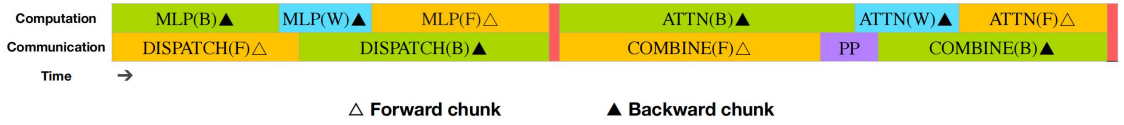
数据来源：《DeepSeek-V3 Technical Report》

DeepSeek的核心技术 -- DualPipe调度策略 + 细粒度的混合精度框架

——基础设施方面进行一定优化，提升效率

管道并行算法 DualPipe

除了基础架构，DeepSeek还在基础设施方面进行了一定优化。例如设计了一种创新的管道并行算法 DualPipe，在每一对前向和后向块内重叠计算和通信，提高通信效率、加速了模型训练



一对单独的前向和后向计算块的重叠策略〔变换器块的边界未对齐〕。橙色表示前向计算，绿色表示"针对输入的后向计算"，蓝色表示"针对权重的后向计算"，紫色表示管道并行〔PP〕通信，红色表示屏障。全连接分发和管道并行通信均可被完全隐藏。

示例：在两个方向上，针对8个管道并行〔PP〕等级和20个微批次的双管道〔DualPipe〕调度方案。
反向方向的微批次与正向方向的微批次是对称的，为简化示意图，我们省略了反向方向微批次的批次编号。
由共享黑色边框包围的两个单元格，其计算和通信过程是相互重叠的。

不同管道并行方法中管道气泡和内存使用的比较

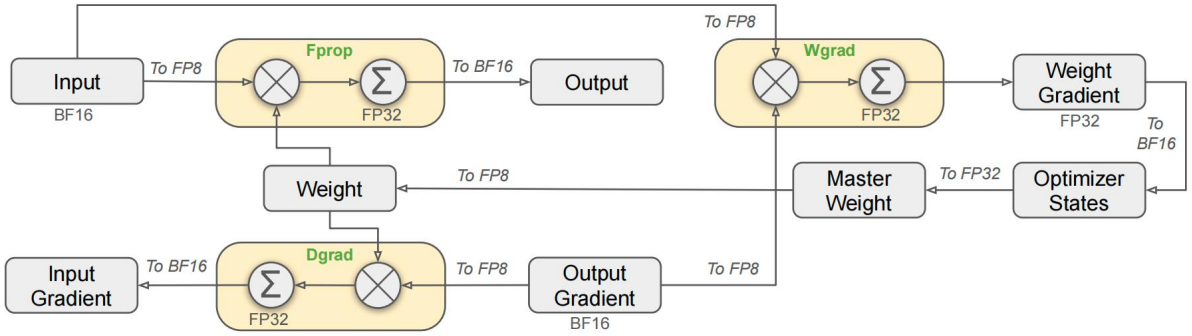
Method	Bubble	Parameter	Activation
1F1B	$(PP-1)(F+B)$	1×	PP
ZB1P	$(PP-1)(F+B-2W)$	1×	PP
DualPipe (Ours)	$(\frac{PP}{2}-1)(F\&B+B-3W)$	2×	$PP+1$

双管道显著减少了管道气泡
无论微批次数量如何增加，管道气泡和激活内存都不会增加

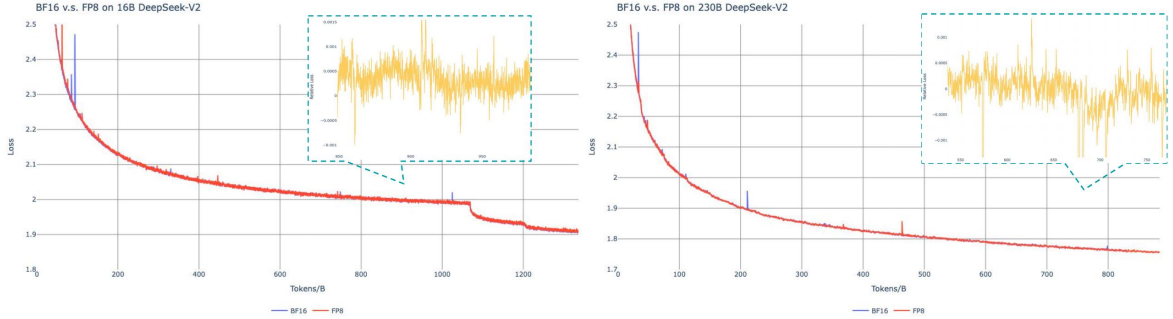
数据来源：《DeepSeek-V3 Technical Report》

细粒度的混合精度框架

DeepSeek提出了一种用于 FP8 训练的混合精度框架，其中大多数计算密集型操作在 FP8 精度下进行，而一些关键操作则战略性地保持在原始数据格式以平衡训练效率和数值稳定性；训练过程中，采用英伟达 PTX〔并行线程执行〕汇编级编程替代标准 CUDA 方案，实现了硬件级深度优化，减少了计算冗余，提高了推理速度。



BF16和FP8训练的损失曲线对比



通过与在不同规模下的两个基准模型上进行BF16训练作对比，对我们的FP8混合精度框架进行了验证
- 在小规模情况下，我们在1.33万亿个token上训练一个包含约160亿总参数的基准MoE模型
- 在大规模情况下，我们在约0.9万亿个token上训练一个包含约2300亿总参数的基准MoE模型

上图展示了训练曲线，并**证明了通过高精度累积和细粒度量化策略，相对误差保持在0.25%以下。**

DeepSeek的核心技术 -- R1-zero 基础模型上的强化学习

——R1-Zero验证纯强化学习（RL）对推理能力的提升

强化学习算法：采用了群体相对策略优化

摒弃了通常与策略模型大小相同的评论家模型，而是从群体得分中估算基线

最终实现训练集上的平均响应长度持续提升，自然地学会了通过更多的思考时间来解决推理任务

奖励建模：准确性奖励、格式奖励

训练模板：要求先给出推理过程，然后给出最终答案

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

Table 1 | Template for DeepSeek-R1-Zero. prompt will be replaced with the specific reasoning question during training.

自我思考能力：自发学会了重新评估其初始回答，更多的思考时间

这种“反思”的特性能够一定程度解决大模型幻觉问题〔大模型逐token输出，过去没有机制去纠正已经输出的错误，反而会继续用错误掩盖先前的问题，带来幻觉问题〕

DeepSeek-R1-Zero与OpenAI o1模型在推理相关基准测试上的比较

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

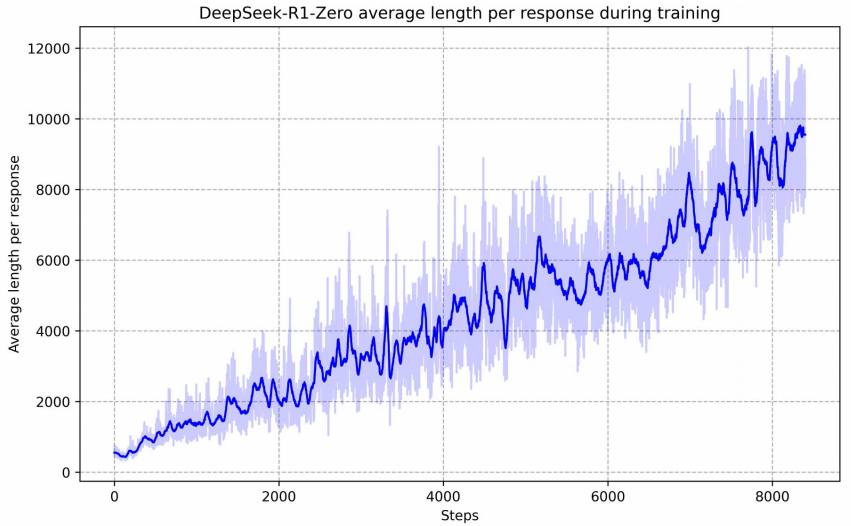
R1-Zero的特别之处在于，其无需任何监督微调数据即可获得强大的推理能力，反映了模型仅通过强化学习就能有效学习和泛化的能力。

尽管R1-Zero模型展现了强大的推理能力，但仍面临可读性差和语言混合等挑战，R1模型则通过冷启动和多阶段训练解决了上述问题。

数据来源：《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》

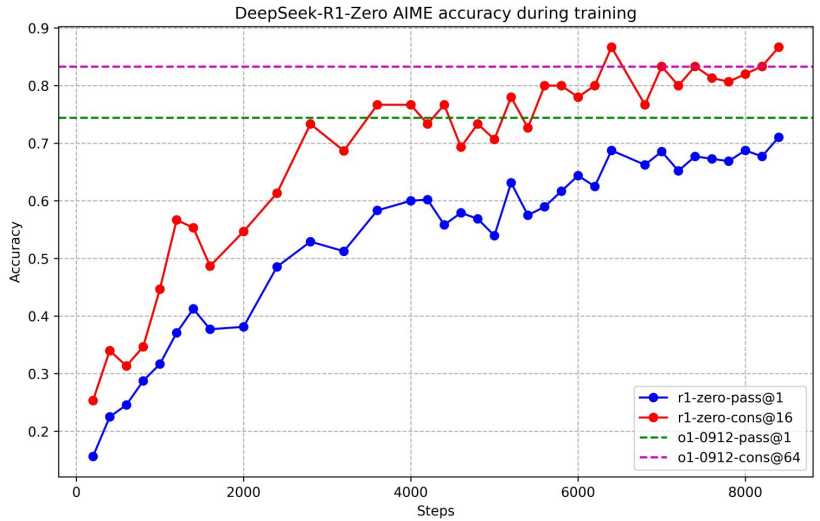
在强化学习过程中，DeepSeek-R1-Zero在训练集上的平均响应长度

DeepSeek-R1-Zero能够自然地学会利用更多的思考时间来解决推理任务



DeepSeek - R1 - Zero在训练期间的AIME准确率

对于每个问题，我们抽取16个回答并计算总体平均准确率，以确保评估的稳定性。



DeepSeek的核心技术 -- R1具有冷启动的强化学习

——R1-Zero验证纯强化学习（RL）对推理能力的提升

背景： 尽管DeepSeek - R1 - Zero展现出了强大的推理能力，并且能够自主发展出出人意料且强大的推理行为，但它仍面临一些问题。例如，DeepSeek - R1 - Zero在可读性差以及语言混杂等方面存在困难。为了使推理过程更具可读性，并能与开源社区共享，我们探索了DeepSeek - R1方法，该方法利用带有对人类友好的冷启动数据的强化学习。

冷启动机制

可读性增强
性能提升

基于长CoT示例的少样本提示
直接提示生成包含反思验证的详细答案
整理DeepSeek-R1-Zero的规范化输出
人工标注后处理优化

拒绝采样与监督微调

推理数据构建
对RL训练检查点执行拒绝采样生成推理轨迹
非推理数据整合在写作、事实QA、自我认知和翻译等领域，采用DeepSeek-V3流程和部分SFT数据
全场景强化学习
优化人类偏好对齐，实施第二阶段RL训练，着重提升模型实用性、安全性和推理能力

推理强化学习优化

引入语言一致性奖励机制，着重提升模型的推理能力，尤其是在涉及有明确解决方案的明确定义问题的推理密集型任务中，例如编程、数学、科学和逻辑推理等任务

蒸馏：赋予小模型推理能力

采用DeepSeek-R1生成的80万训练样本，对Qwen和Llama等开源模型进行直接微调，旨在将DeekSeek-R1的推理能力迁移至计算效率更高的小型模型。

实验结果表明，这种直接知识蒸馏方法能显著提升小型模型的推理性能。研究选用的基础模型包括：Qwen2.5-Math-1.5B、Qwen2.5-Math-7B、Qwen2.5-14B、Qwen2.5-32B、Llama-3.1-8B和Llama-3.3-70B-Instruct

DeepSeek - R1蒸馏模型与其他可比模型在推理相关基准测试上的比较

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

如上表所示，仅仅对DeepSeek - R1的输出进行蒸馏，就能使高效的DeepSeek - R1 - 7B［即DeepSeek - R1 - 蒸馏 - Qwen - 7B，下文类似简称］在各个方面都优于像GPT - 4o - 0513这样的非推理模型。

DeepSeek - R1 - 14B在所有评估指标上都超过了QwQ - 32B - 预览版，而DeepSeek - R1 - 32B和DeepSeek - R1 - 70B在大多数基准测试中显著超过o1 - mini。这些结果展示了蒸馏的强大潜力。此外，我们发现对这些蒸馏模型应用强化学习［RL］能带来显著的进一步提升。我们认为这值得进一步探索，因此在此仅展示简单监督微调［SFT］蒸馏模型的结果。

数据来源：《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》

DeepSeek的应用场景

——AI技术驱动的场景化赋能

零售领域：数据驱动的精准运营

客户需求预测

- 技术方案：

  融合Transformer时序模型与外部环境变量〔天气、节假日〕，动态预测区域级商品需求。

  结合联邦学习技术，保护隐私的同时整合多门店数据，提升预测泛化能力。

- 业务价值：

  → 降低预测误差率、降低缺货率；

  → 支持动态补货策略，降低仓储成本。

金融领域：智能风控系统

多模态图神经网络+动态对抗训练

年损失减少亿元级

教育领域：自适应学习生态

智能辅导系统

- 核心技术：

  多模态交互：语音识别〔ASR〕+手势识别，支持低龄学生自然交互解题辅导。

  认知诊断：基于DKT〔深度知识追踪〕模型量化学生知识状态，动态生成学习路径图谱。

- 落地场景：

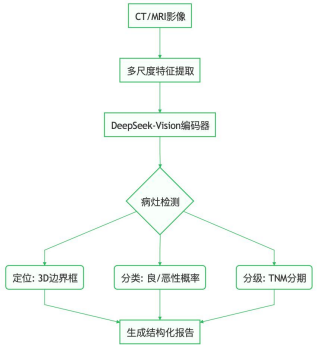
  → 小学数学辅导场景，知识点掌握度预测准确率超90%；

  → 自动批改作文并生成多维反馈〔语法/逻辑/创意〕，节省教师70%批改时间。

医疗领域：影像辅助诊断

MoE架构+多模态对齐

早期癌症检出率提升



DeepSeek的技术发展趋势

——通用智能与垂直场景的双重进化

通用人工智能（AGI）的渐进式突破，推动科技新变革

大模型持续扩展

参数规模：千亿级→万亿级参数演进，混合专家〔MoE〕架构提升效率
能力泛化：从单模态到多模态统一建模〔文本/图像/视频/传感器数据联合学习〕

自主决策能力增强

世界模型构建：通过物理仿真与真实数据融合，提升对复杂环境的理解能力
因果推理升级：从统计相关性向因果机制建模跨越〔如反事实干预推演〕

人机协作深化

具身智能：机器人+AI深度融合，实现物理世界交互〔如仓储分拣、实验操作〕
伦理安全框架：构建价值观对齐与风险可控的AGI系统

训练成本高昂

超级App，
DAU >= 五千万

AGI提供通用认知能力
2025年底进入快速发展
进程比垂直领域AI慢

垂直领域AI的深度渗透，重塑行业格局

行业大模型专业化

领域知识注入〔如医疗术语库、零售供应链图谱、教育知识图谱〕提升任务精度
轻量化部署：模型蒸馏+硬件适配技术推动边缘场景落地

实时化与个性化

AI推理加速：端侧实时推理〔<10ms延迟〕支持毫秒级决策〔如线下实时推荐〕
数据互通与个性化：联邦学习保障隐私安全下的跨机构数据协同，支持个性化

闭环增强系统

"感知-决策-执行"全链路自动化〔如教育场景的"学习-测评-推荐"闭环〕
在线学习实现模型动态迭代〔天级甚至小时级更新〕

训练成本显著降低

中小型App，
DAU <= 五千万

垂直AI通过领域知识增强精准性
已进入快速发展
进程比通用人工AGI快

数据参考：推理成本预降数十倍，2025中国将从“百模大战”转向应用大战

与同行技术的比较

——性能、场景与创新的多维竞争力

一、性能对比：效率与精度双突破

架构设计

DeepSeek：采用混合专家〔MoE〕架构，激活参数仅占模型总量的5.5%〔如DeepSeek-R1激活370亿参数，总量6710亿〕，显著降低计算资源消耗，推理延迟压至10ms级

GPT系列：基于纯Transformer架构，依赖全参数激活〔如GPT-4约1万亿参数〕，推理成本高且响应速度较慢
Claude系列：强调安全对齐性，但模型灵活性和多任务处理能力受限，推理速度低于DeepSeek

任务表现

中文场景：DeepSeek在C-Eval〔86.5%〕、C-SimpleQA〔64.1%〕等中文评测中显著优于GPT-4〔中文任务偏差率降低30%+〕

代码生成：HumanEval-Mul得分82.6%，超越GPT-4o〔78.2%〕和Claude 3.5〔80.1%〕，尤其在函数调用和API集成上表现突出

多模态推理：Gemini在多模态任务领先，但DeepSeek通过强化学习后训练〔RLHF〕在纯文本逻辑推理〔DROP 91.6%〕上超越同类模型17

资源效率

训练成本仅550万美元〔GPT-4估算超1亿美元〕，单位算力能耗降低80%

支持FP8量化和动态稀疏训练，边缘设备可部署百亿参数模型〔如零售终端AR推荐〕

二、应用场景对比：垂直优化与通用泛化

模型	核心优势场景	局限性
DeepSeek	中文任务、代码生成、实时决策〔零售库存联调、教育个性化推荐〕	多模态支持较弱，长上下文处理〔64k vs Claude 200k〕
GPT系列	创意写作、长文本生成〔法律文档、学术研究〕	中文语义偏差，部署成本高〔API价格超DeepSeek 10倍〕
Claude系列	安全敏感场景〔医疗咨询、法律合规〕	灵活性与创造力不足，推理速度慢
Gemini	多模态分析〔视频描述、跨媒体检索〕	纯文本任务表现平庸，模型臃肿

三、创新能力对比：开源生态与技术前瞻性

技术突破

低成本训练

仅用1/11算力〔对比Llama-3-405B〕实现同等性能，FP8量化技术压缩训练能耗70%

开源战略

完全开源模型代码与训练框架

吸引超10万开发者贡献；降低企业AI开发成本

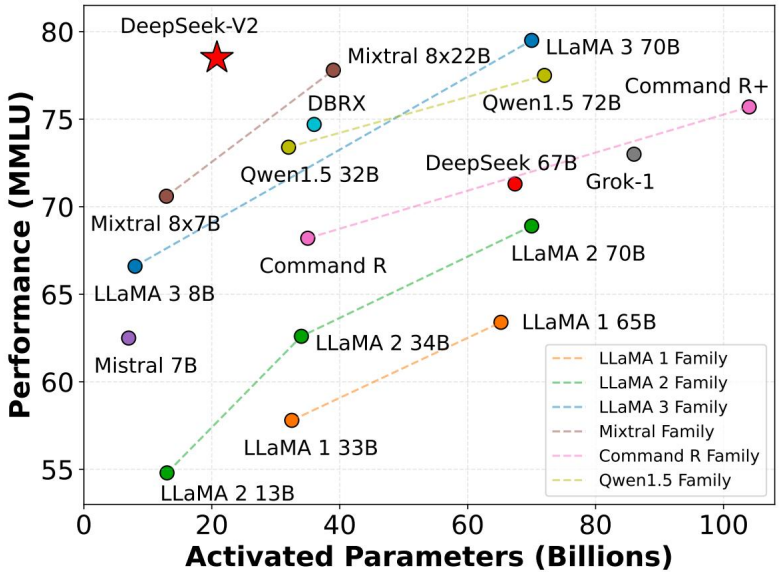
未来方向

AGI基座
研发万亿参数MoE架构+

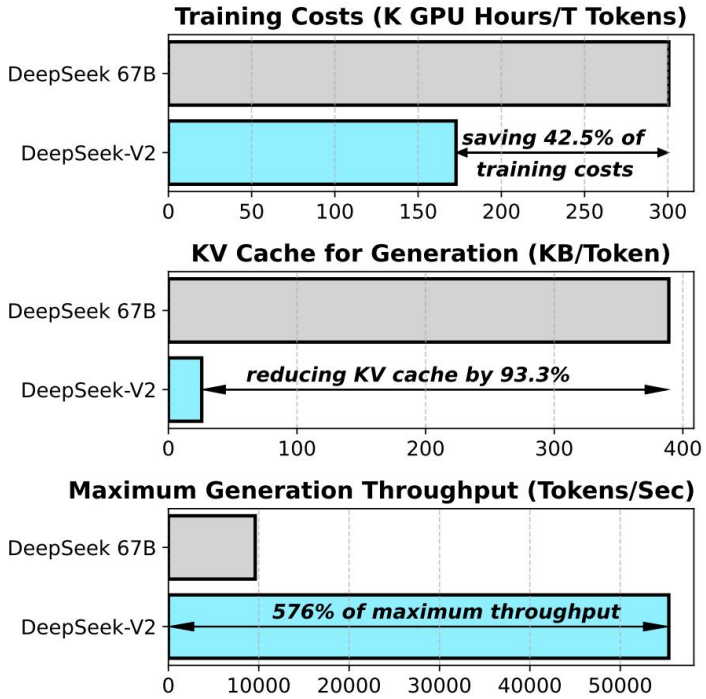
垂直领域AI
中小公司如春笋般涌现

各大模型按总分降序排列

在不同开源模型中
MMLU〔大规模多任务语言理解评估基准〕准确率与激活参数的关系



DeepSeek - 76B〔稠密型〕模型和
DeepSeek - V2模型的训练成本及推理效率



DeepSeek - V2与其他代表性开源模型的比较

Model 模型	Overall 总分	Reasoning 中文推理			Language 中文语言						
		Avg. 推理总分	Math. 数学计算	Logi. 逻辑推理	Avg. 语言总分	Fund. 基本任务	Chi. 中文理解	Open. 综合问答	Writ. 文本写作	Role. 角色扮演	Pro. 专业能力
GPT-4-1106-Preview	**8.01**	**7.73**	**7.80**	**7.66**	8.29	7.99	7.33	**8.61**	**8.67**	**8.47**	**8.65**
DeepSeek-V2 Chat (RL)	**7.91**	7.45	7.77	7.14	**8.36**	**8.10**	8.28	8.37	8.53	8.33	8.53
ERNIEBot-4.0-202404*（文心一言）	**7.89**	7.61	**7.81**	7.41	8.17	7.56	**8.53**	8.13	8.45	8.24	8.09
DeepSeek-V2 Chat (SFT)	**7.74**	7.30	7.34	7.26	8.17	8.04	8.26	8.13	8.00	8.10	8.49
GPT-4-0613	**7.53**	*7.47*	7.56	7.37	7.59	7.81	6.93	7.42	7.93	7.51	7.94
ERNIEBot-4.0-202312*（文心一言）	**7.36**	6.84	7.00	6.67	7.88	7.47	7.88	8.05	8.19	7.84	7.85
Moonshot-v1-32k-202404*（月之暗面）	**7.22**	6.42	6.41	6.43	8.02	7.82	7.58	8.00	8.22	8.19	8.29
Qwen1.5-72B-Chat*	**7.19**	6.45	6.58	6.31	7.93	7.38	7.77	8.15	8.02	8.05	8.24
DeepSeek-67B-Chat	**6.43**	*5.75*	5.71	5.79	7.11	7.12	6.52	7.58	7.20	6.91	7.37
ChatGLM-Turbo（智谱清言）	**6.24**	*5.00*	4.74	5.26	7.49	6.82	7.17	8.16	7.77	7.76	7.24
ERNIEBot-3.5（文心一言）	**6.14**	*5.15*	5.03	5.27	7.13	6.62	7.60	7.26	7.56	6.83	6.90
Yi-34B-Chat*	**6.12**	4.86	4.97	4.74	7.38	6.72	7.28	7.76	7.44	7.58	7.53
GPT-3.5-Turbo-0613	**6.08**	*5.35*	5.68	5.02	6.82	6.71	5.81	7.29	7.03	7.28	6.77
ChatGLM-Pro（智谱清言）	**5.83**	*4.65*	4.54	4.75	7.01	6.51	6.76	7.47	7.07	7.34	6.89
SparkDesk-V2（讯飞星火）	**5.74**	*4.73*	4.71	4.74	6.76	5.84	6.97	7.29	7.18	6.92	6.34
Qwen-14B-Chat	**5.72**	*4.81*	4.91	4.71	6.63	6.90	6.36	6.74	6.64	6.59	6.56
Baichuan2-13B-Chat	**5.25**	*3.92*	3.76	4.07	6.59	6.22	6.05	7.11	6.97	6.75	6.43
ChatGLM3-6B	**4.97**	*3.85*	3.55	4.14	6.10	5.75	5.29	6.71	6.83	6.28	5.73
Baichuan2-7B-Chat	**4.97**	*3.66*	3.56	3.75	6.28	5.81	5.50	7.13	6.84	6.53	5.84
InternLM-20B	**4.96**	*3.66*	3.39	3.92	6.26	5.96	5.50	7.18	6.19	6.49	6.22
Qwen-7B-Chat	**4.91**	*3.73*	3.62	3.83	6.09	6.40	5.74	6.26	6.31	6.19	5.66
ChatGLM2-6B	**4.48**	*3.39*	3.16	3.61	5.58	4.91	4.52	6.66	6.25	6.08	5.08
InternLM-Chat-7B	**3.65**	*2.56*	2.45	2.66	4.75	4.34	4.09	5.82	4.89	5.32	4.06
Chinese-LLaMA-2-7B-Chat	**3.57**	*2.68*	2.29	3.07	4.46	4.31	4.26	4.50	4.63	4.91	4.13
LLaMA-2-13B-Chinese-Chat	**3.35**	*2.47*	2.21	2.73	4.23	4.13	3.31	4.79	3.93	4.53	4.71

数据来源：《DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model》

DeepSeek-R1 性能评分

DeepSeek-R1 与其他代表性模型的比较

	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	88.0	89.1	86.7	-	**92.9**
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	**84.0**
	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)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	**92.3**
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	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	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	**79.8**
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	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	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

DeepSeek-R1 蒸馏模型评估

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

蒸馏模型与强化学习（RL：Reinforcement Learning）模型在推理相关基准测试中的比较

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	**72.6**	**83.3**	**94.3**	**62.1**	**57.2**

数据来源：《DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning》

如何高效用好DeepSeek？（一）

Step1: 前提需要先了解清楚大语言指令模型、推理模型工作原理与局限

指令模型： open ai的gpt4o、字节豆包，用于遵循指令生成任务；**需要较完善的提示词，才能激发模型的表现**
推理模型： deepseek r1、gpt-o1 专注于逻辑推理问题解决，自主处理多步骤、因果推断或者解决复杂决策的
清晰明确表达你的需求即可

Step2: 在和DS交流时，当它当成是你极其聪明超过10年工作经验的助理，需要交待清楚你的诉求是什么

高效向 DeepSeek 提问模版

1. 我的角色/背景: （例如：我是蜜雪冰城的产品运营）
2. 我的问题场景: （例如：希望通过12个月的周期提升客单价）
3. 我的目标: （例如：提升客单价至15元，同时稳住市场份额）
4. 我的限制条件: （例如：能接受短期业绩波动）
5. 期望的回答形式: （例如：需要具体的执行方案，包括产品升级、套餐设计等）

个人使用建议：

- 如果需要分析的是23年12月之前的不太需要联网模式；
- 如果是近期、或实时新闻事件则需要开启联网模式

如何高效用好DeepSeek？（二）

1. 明确问题背景

将复杂问题拆解为多个小问题，或列出关键点。

示例：

低效提问：“怎么提升客单价？”
高效提问：“我是蜜雪冰城的产品运营，希望通过12个月的周期提升客单价，同时稳住市场份额，能接受短期业绩波动，有什么具体方案？”

2. 提供具体信息

包括：您的角色（如产品经理、学生、创业者等）。
问题的具体场景（如“针对老年人团队的重庆旅游攻略”）。

您的目标（如“提升客单价”“设计一个2天行程”）。
限制条件（如“预算有限”“不接受过度劳累”）。

示例：
低效提问：“重庆怎么玩？”

高效提问：“我计划2月初带老年人团队去重庆玩2天，希望行程轻松、避开人流高峰，有什么推荐路线和注意事项？”

3. 结构化描述问题

将复杂问题拆解为多个小问题，或列出关键点。

示例：
低效提问：“怎么运营一个品牌？”

高效提问：“我想运营一个新茶饮品牌，目前有以下问题：
如何定位目标用户？
如何设计产品线？
如何通过社交媒体吸引第一批用户？”

避免无效向 DeepSeek 提问

1. 避免过于宽泛的问题

无效示例：告诉我一些有趣的事情。
改进建议：明确具体领域或主题。
有效示例：能推荐一些适合初学者的编程学习资源吗？

2. 提供足够的上下文

无效示例：帮我写个方案。
改进建议：说明背景、目标和限制条件。
有效示例：我需要为一家新茶饮品牌设计一个营销方案，目标用户是18-25岁的年轻人，预算有限，希望聚焦社交媒体推广。

3. 避免过于复杂或冗长的描述

无效示例：一段长达500字、包含多个不相关问题的描述。
改进建议：将复杂问题拆解为多个小问题，或聚焦核心需求。
有效示例：如何设计一个吸引年轻人的品牌logo？、如何通过社交媒体推广新品牌？

4. 避免模糊的指令

无效示例：给我一些建议。
改进建议：明确需要建议的具体方向。
有效示例：我想提升工作效率，能给我一些时间管理的建议吗？

5. 避免矛盾或不切实际的要求

无效示例：帮我写一篇1000字的文章，但只能用50个字。
改进建议：确保需求合理且可实现。
有效示例：我需要一篇500字左右的文章，介绍如何提升团队协作效率。

6. 避免使用歧义或模糊的词汇

无效示例：给我一些‘好’的建议。
改进建议：明确“好”的具体标准。
有效示例：我需要一些低成本、易执行的营销活动建议。

7. 避免重复提问

无效示例：多次提问相同或类似的问题。
改进建议：如果对回答不满意，可以补充更多细节或调整问题方向。
有效示例：关于时间管理，除了番茄工作法，还有其他适合职场新人的方法吗？

DeepSeek 大解密

最后，特别兴奋国产DeepSeek大幅降低训练成本且性能出色 让AI进一步融入日常生活，服务大众

一起学习，一起交流，跟随技术奔跑，共勉！

技术发展进步很快，未来ds不一定是最完美那个，但一定会在AI 历史长河中留下浓墨重彩的一笔