【电子与信息科学 / Electronics and Information Science】

# DeepSeek-R1是怎样炼成的?

张慧敏

**摘　要：**简述DeepSeek系列模型在大模型训练中的创新和优化. DeepSeek系列模型的突破主要体现在模型架构、算法创新、软硬件协同优化及整体训练效率的提升. DeepSeek-V3模型采用混合专家(mixture of experts, MoE)模型架构, 通过细粒度设计和共享专家策略, 实现计算资源的高效利用; MoE模型架构中的稀疏激活机制和无损负载均衡策略显著提高了模型训练的效率和性能; 多头潜在注意力(multi-head latent attention, MLA)机制通过减少内存使用和加速推理过程, 降低了模型训练和推理成本; 通过引入多token预测(multi-token prediction, MTP)和8位浮点数(floating point 8-bit, FP8)混合精度训练技术, 提升了模型的上下文理解能力和训练效率; 采用优化并行线程执行(parallel thread execution, PTX)代码显著提高了图形处理器(graphics processing unit, GPU)的计算效率; 所提群体相对策略优化(group relative policy optimization, GRPO)对DeepSeek-R1-Zero模型进行纯强化学习训练, 跳过了传统的监督微调和人类反馈阶段, 显著提升了模型的推理能力. 总体而言, DeepSeek系列模型通过多项创新, 在人工智能领域取得了显著优势, 树立了行业新标杆.

**关键词：**人工智能; DeepSeek; 大语言模型; 混合专家模型; 多头潜在注意力机制; 多token预测; 混合精度训练; 群体相对策略优化

**中图分类号：**TP18　　**文献标志码：**D　　**DOI：**10.3724/SP.J.1249.2025.02226

# How DeepSeek-R1 was created?

**ZHANG Huimin**

**Abstract:** This article summarizes the innovations and optimizations in DeepSeek series models for large-scale training. The breakthroughs of DeepSeek are primarily reflected in model and algorithm innovations, software and hardware collaborative optimization, and the improvement of overall training efficiency. The DeepSeek-V3 adopts a mixture of experts (MoE) architecture, achieving efficient utilization of computing resources through fine-grained design and shared expert strategies. The sparse activation mechanism and lossless load balancing strategy in the MoE architecture significantly enhance the efficiency and performance of model training, especially when handling large-scale data and complex tasks. The innovative multi-head latent attention (MLA) mechanism reduces memory usage and accelerates the inference process, thus lowering training and inference costs. In DeepSeek-V3′s training, the introduction of multi-token prediction (MTP) and 8-bit floating-point (FP8) mixed-precision training technologies improves the model′s contextual understanding and training efficiency, while optimizing parallel thread execution (PTX) code significantly enhances the computation efficiency of graphics processing units (GPUs). In training the DeepSeek-R1-Zero model, group relative policy optimization (GRPO) is used for pure reinforcement learning, by passing the traditional supervised fine-tuning and human feedback stages, leading to a significant improvement in inference capabilities. Overall, DeepSeek series models has achieved significant advantages in the field of artificial

**Received:** 2025-02-05; **Accepted:** 2025-02-08; **Online** (CNKI)**:** 2025-02-11
**Corresponding author:** Professor ZHANG Huimin (zhanghm88@hotmail.com)
**Citation:** ZHANG Huimin. How DeepSeek-R1 was created? [J]. Journal of Shenzhen University Science and Engineering, 2025, 42(2): 226-232. (in Chinese)

intelligence through multiple innovations, setting a new industry benchmark.

**Key words:** artificial intelligence; DeepSeek; large language model; mixture of experts architecture; multi-head latent attention mechanism; multi-token prediction; mixed-precision training; group relative policy optimization

2025 年 1 月，中国杭州深度求索人工智能基础技术研究有限公司（DeepSeek）发布的 DeepSeek-R1 推理模型，震动了全球科技圈，受到各大科技公司，甚至美国政府的密切关注．DeepSeek-R1 模型在编程、数学及逻辑推理等方面都表现出色，性能与目前公开发布的最强推理模型 OpenAI o1 不相上下．更重要的是，DeepSeek-R1 模型以开源形式提供给全球研究人员和开发者使用，展示了真正的开放精神，其影响难以估量．英伟达（NVIDIA）的人工智能（artificial intelligence，AI）科学家 Jim FAN 称赞 DeepSeek-R1 模型是“真正开放、赋能所有人的前沿研究”．

DeepSeek 公司成立于 2023 年 5 月，由中国对冲基金幻方量化创立，创始人为梁文锋，公司研发团队主要由中国本土顶尖大学毕业的博士生组成，团队成员年轻、工作高效且能够紧密合作，擅于快速学习并利用最新技术进行大模型研发．2024 年 12 月底，DeepSeek 发布并开源了 DeepSeek-V3 模型，性能可媲美当时的顶级闭源模型，而不到 600 万美元的训练成本仅为 GPT-4 的 1/20，且训练用时仅 2 个月．其后推出的推理模型 DeepSeek-R1，其性能更是在多项测试中达到或超越了 OpenAI o1 模型．

尽管 DeepSeek-V3 和 DeepSeek-R1 模型的强大功能仍存在争议，但大多数人都同意，两模型的性能分别达到了当时顶尖的闭源大模型 ChatGPT 4o 和 ChatGPT o1 的水平．目前，对 DeepSeek-R1 最大的争议在于，它仅用了 2 048 块 NVIDIA H800 图形处理器（graphics processing unit，GPU）组成的集群，在约 2 个月内完成了对拥有 6 710 亿个参数的混合专家（mixture of experts，MoE）模型的训练，效率比 Meta 等行业领军企业发布的 AI 大模型高出 10 倍，训练成本却仅为 OpenAI o1 的 3% ~ 5%．有评论认为，DeepSeek 可能在技术报告中夸大了其模型训练过程中的效率和资源利用率，但也有人认为，这可能是因为 DeepSeek 在技术上取得了巨大进步，从而实现了这一看似不可能的任务．

DeepSeek 的崛起将对全球的 AI 产业产生什么影响，以及它是否会重塑未来的 AI 产业格局呢？本文通过分析公开的信息和资料，尤其是 DeepSeek-V3 模型和 DeepSeek-R1 模型的技术报告[1-2]，分析 DeepSeek 系列模型，包括 DeepSeek-V3、DeepSeek-R1-Zero 和 DeepSeek-R1 的训练方法，详细介绍 DeepSeek 在现有大模型架构上的探索与改进，深入探讨该系列模型对 GPU 集群负载均衡的改进，采用并行线程执行（pardlel thread execution，PTX）编程语言对 GPU 进行底层优化，以实现包括通信、内存和计算等软硬件系统的协同优化．文章力求用通俗易懂的语言表达笔者的观察和理解，为非 AI 科技圈的读者提供一个客观、公正的分析视角，也为 AI 科技圈的专业人士提供有价值的参考．

## 1　DeepSeek-V3 的高效构架与创新技术

DeepSeek-V3 在编程能力、数学推理、中文理解和长文本理解等领域表现出色．DeepSeek-R1 是基于 DeepSeek-V3 专为复杂推理任务而设计的大语言模型（large language model，LLM），DeepSeek-R1-Zero 则是这两个模型中间的一个过渡性推理模型．虽然 DeepSeek 系列模型仍采用 Transformer 架构，但在架构和算法的各个方面都进行了极致优化，并融入了令人赞叹的最新的创新技术．

### 1.1　高效的模型架构：MoE 模型

DeepSeek-V3 采用的 MoE 模型是目前大多数 AI 大模型都使用的技术，当然也有不依赖 MoE 模型，如美国 AI 企业 Anthropic 的 Claude 和 Meta 的 LLaMA 系列．因信息不透明，目前无法确定 OpenAI 的 ChatGPT 3.5 和 ChatGPT 4.0 是否也采用了 MoE 架构．

MoE 模型通过组合多个专家模型来处理复杂任务，每个专家模型专注于输入数据的不同部分，门控网络决定如何加权这些专家模型的输出．MoE 模型的核心思想是将任务分解为多个子任务，再由不同的专家处理，从而提高了模型的性能，该方法在自然语言处理和计算机视觉等领域表现出色，尤其适合处理大规模数据和复杂任务．通过动态分配计算资源，MoE 模型能够在高效利用硬件资源的同时保持高精度和泛化能力．

DeepSeek 的 MoE 架构独特之处在于细粒度设

计和共享专家策略．其他MoE模型中每个MoE层可能拥有几个到几十个专家，如美国xAI公司的Grok-1采用8个专家的MoE架构，每处理1个token会激活2个专家．而在DeepSeek的MoE架构中，每个MoE层由1个共享专家和256个路由专家组成，每个token会从这些路由专家中选择8个最合适的专家进行处理．

共享专家策略是DeepSeek MoE架构中的重要创新．共享专家数量固定且较少，每个MoE层通常仅包含1个始终处于激活状态的共享专家，负责捕获和整合不同上下文中的通用知识，因此减少了知识冗余，提高了参数的使用效率，使得独立路由专家能专注于处理更专业化的知识．共享专家策略提高了模型的泛化能力和整体效率，减少了其他路由专家之间的参数冗余，又通过与细粒度专家分割相结合，实现了高效的模型架构．

设计这种精细的MoE架构在工程上是非常复杂且极具挑战性的，但DeepSeek团队通过精心设计，使模型在缺乏足够的高性能GPU情况下仍能在效率和性能方面达到新的高度，为AI领域树立了新标杆．

在训练过程中，DeepSeek MoE构架的每个token在各个MoE层中都仅激活8个路由专家，且最多可路由4个节点，这种专家激活方法被称为稀疏激活．稀疏激活机制可以在不显著增加计算成本的情况下，大幅扩展模型的容量．

采用细粒度专家系统和稀疏激活机制具有明显的优点．首先，通过减少连接和激活的数量，大大减少了网络的参数量，从而降低了模型的存储需求和计算开销．其次，稀疏的连接和激活模式使模型更具解释性，有助于人们理解模型的决策过程．限制连接和激活还可以降低数据噪声和冗余信息对模型的影响，提高模型对干扰的鲁棒性．通过提取最相关和最重要的特征，不仅增强了模型的泛化能力，还有效减少了模型的过拟合风险．最后，通过只保留最重要的激活值，大幅减少了模型的计算量和内存使用，却几乎不影响其性能．但是，这些方法的缺点也是显而易见的：首先是实现的复杂度较高，不仅需要复杂的路由机制还需要专门的硬件支持；其次，在训练阶段可能需要占用更多的计算资源来优化专家分配和激活模式，这对于资源有限的团队来说是一个挑战；最后，平衡专家数量、激活策略和模型性能三者的关系是一个复杂的过程，往往需要进行大量的实验和调优．

### 1.2 无辅助损失的负载均衡策略

采用MoE架构训练超大规模的LLM，最大的挑战是如何实现负载均衡，这涉及到效率、性能瓶颈、训练稳定性、可扩展性和通信开销等方面的取舍和平衡．若负载不均衡，会使一些专家被过度使用，而其他专家被闲置，从而导致计算资源浪费和训练效率降低．负载不均衡还会导致系统产生性能瓶颈，热门专家负载过高，冷门专家负载过低，进而形成自我强化的循环．随着模型规模增大，负载不均衡还会限制模型的可扩展性，导致收益递减．此外，在分布式训练中，负载不均衡会增加节点间的通信开销，进而降低模型的训练速度．

DeepSeek团队提出的无辅助损失负载均衡(auxiliary-loss-free load balancing，ALFLB)策略通过动态偏差调整路由任务，并根据专家的近期负载调整偏差值，从而实现自适应负载分配．当某个专家过载时，系统会自动降低其接收新任务的概率；反之，则提高该专家接收任务的机会．相比传统的辅助损失方法，ALFLB策略避免了对模型主要训练目标的干扰，能显著提升模型的性能和训练效率，同时减少了对内存和计算资源的消耗，允许模型在不增加键值(key-value，KV)缓存大小的情况下增加注意力头的数量，从而潜在地提高了模型能力．总体而言，该策略通过自然均衡的负载分配，提供了一个高效且低成本的解决方案，显著提升了LLM的性能和训练效率．

### 1.3 创新的注意力机制：多头潜在注意力机制

当ChatGPT生成文本时，它不仅关注刚刚生成的词，还会综合考虑所有已输入的上下文和之前生成的所有词，模型会为这些词分配不同的权重，从而差异化地关注它们对当前生成词的影响．这种动态的差异化关注机制，使模型能够捕捉上下文中的关键信息，生成更加自然连贯且语义丰富的文本，这便是注意力机制[3]的直观体现．

为在训练过程中实现注意力机制，Transformer模型通过引入查询矩阵($\boldsymbol{Q}$)、键矩阵($\boldsymbol{K}$)和值矩阵($\boldsymbol{V}$)来计算注意力，这里$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$都是高维矩阵．在实际的句子生成过程中，首先将$\boldsymbol{Q}$和$\boldsymbol{K}$相乘，计算出前面句子中的不同部分与即将生成的词的关联度，再乘以表示前面句子内容的$\boldsymbol{V}$，进而计算出注意力，并决定下一个词是什么．

多头注意力(multi-head attention，MHA)机制是

对自注意力的改进和扩展，可令模型犹如一个多角度的观察者，从多个视角出发同时捕捉不同的特征以及不同的相关性．采用MHA机制不仅扩展了模型的表示空间，还增强了其学习复杂特征的能力．多个注意力头还可以并行计算，因此提高了模型的处理速度，也减少了过拟合风险，从而提升了模型的泛化能力．不同的注意力头关注输入数据的不同特征，这使模型能够更全面地理解语义．通过这种多角度并行处理，MHA使模型能够更全面地理解复杂的语言结构和语义关系，在各种自然语言处理任务中都表现出色．

DeepSeek公司在DeepSeek-V2模型中首次提出了多头潜在注意力(multi-head latent attention，MLA)机制，克服了LLM在训练和推理过程中的瓶颈，特别是KV缓存占用大量内存的问题．MLA机制所需显存仅为MHA的5% ~ 13%，且因减少了KV缓存对资源的占用，模型在处理长序列时的推理速度也更快．与此同时，MLA机制还能实现与MHA机制相当甚至更强的性能，这使得DeepSeek-V2模型在保持高性能的同时，显著降低了训练和推理成本．这项创新为DeepSeek系列模型在LLM领域赢得了显著优势．

MLA机制创新地采用低秩键值联合压缩技术，将传统的MHA机制缓存的键矩阵和值矩阵压缩为一个低维潜在向量，在保留了关键信息的同时显著减少了内存占用，实现了对注意力的高效计算．这种设计令MLA机制在保持或提升模型性能的同时，显著降低了对计算资源的需求，尤其是在处理长序列时效果更明显．这种创新使MLA机制能够在LLM应用中实现更高效的训练和推理，是训练DeepSeek-V3模型的关键之一．

在预训练阶段，MLA机制在扩大模型容量、增加批量大小和优化计算与内存的平衡等方面都展现了显著优势．尽管MLA机制会额外增加计算的复杂性，但节省出的内存资源和潜在的性能改进优势通常超过了这种因计算增加而来的负担，特别是在内存硬件受到限制的情况下．

在推理阶段，MLA机制通过减少KV缓存数量和使用更小的维度，将键矩阵和值矩阵投影至低维潜在空间，从而显著降低内存占用，进而提高了推理效率．尽管这样会增加计算量，但减少了内存带宽和存储需求．此外，MLA机制允许在不增加KV缓存大小的情况下增加注意力头的数量，使模型能够潜在地提高能力而不牺牲推理速度．

### 1.4　多token预测的应用

DeepSeek-V3模型采用多token预测(multi-token prediction，MTP)[4]技术，使其在LLM领域独树一帜．MTP的工作原理是通过使用多个输出头并行地预测多个token，再由主输出头验证预测结果并选择最有可能的结果．模型使用$n$个独立的输出头来预测$n$个未来的token，通过共享同一个主干网络生成上下文的潜在表征，再将该表征送入$n$个独立的头网络．这种设计简单且易于实现，不需要进行复杂的架构改变．

美国互联网公司Meta的研究表明，MTP技术能够通过并行地预测多个token，为模型提供更丰富的监督信号，使其能更快地学习语言的结构和规律．例如，使用4-token预测训练的模型的推理速度比单token模型快3倍．MTP还可以帮助模型学习不同token之间的长距离依赖关系，从而更好地理解上下文信息，此功能在编程任务上表现突出，增强了分布外泛化能力．

然而，MTP的运行需要占用大量的计算资源，尤其是在模型规模较大时，即使只是简单地实现MTP也可能导致内存使用量迅速增加，需要采用特殊的优化技术来解决此问题．因此，在处理某些特定的自然语言处理(natural language processing，NLP)任务时，MTP并不总是优于传统的单token预测，如在一些标准选择题任务中的表现并不佳．

DeepSeek团队率先将MTP技术应用于DeepSeek-V3模型和DeepSeek-R1模型的训练中，通过极致的内存和通信管理，充分发挥了MTP的高效优势，其改进包括提高数据效率、增强预测能力、减少训练时间和提升模型的泛化能力．这种创新显著提升了效率和性能，使DeepSeek在AI技术前沿占据了领先地位．

### 1.5　混合精度训练

DeepSeek-V3模型的一项重大创新是引入8位浮点数(floating point 8-bit，FP8)混合精度训练框架，即对精度要求相对较低的数据使用8 bit的浮点数表示数据，而对精度要求较高的数据，采用32 bit(FP32)或16 bit(FP16)的浮点数来表示，使得训练出的模型相较于传统的采用FP32和FP16格式训练模型，虽然精度降低了，但占用空间更小，且计算速度也更快．混合精度策略采用FP8来实现大部分的核心计算内核，包括前向传播、激活反向传播

和权重反向传播．输出结果则采用BF16或FP32格式，向量激活值以FP8格式存储用于反向传播．这种方法可显著提升模型的计算速度，同时大大降低内存消耗．

DeepSeek通过创新的误差累积解决方案，令FP8混合精度训练将精度损失控制在0.25%内，几乎不影响模型的性能．首次在超大规模模型上验证了FP8混合精度训练的有效性，使DeepSeek-V3模型在降低GPU内存占用量和计算开销的同时，仍能保持高水平的性能，进一步提高了单位GPU小时的计算利用率，降低了整体训练成本．

混合精度训练的实际操作是相当困难的，需要设计团队对大模型训练过程中的每一个环节和细节的计算精度都有全面且精准的把握．因此，许多大型AI公司，尤其是拥有巨资又手握数以10万计GPU的AI巨头们发布的大模型并未采用混合精度训练．DeepSeek系列模型采用混合精度训练方法，实属迫不得已，而成功地实现了这一点，可以说是绝处逢生．

### 1.6 通过直接编写和优化PTX代码来提升GPU的计算效率

PTX是NVIDIA架构统一计算设备(compute unified device architecture，CUDA)的中间表示语言，介于高级GPU编程语言(如CUDA C/C++)和低级机器代码SASS(syntactically awesdme systle sheets)之间．PTX提供了更接近底层的指令集架构，允许开发者进行细粒度的优化．DeepSeek团队在训练DeepSeek-V3模型时，通过编写和优化PTX代码提高了GPU的计算效率，包括将132个流式多处理器(streaming multiprocessors，SMs)中的20个专用于服务器间的通信，从而绕过通信带宽的限制；通过优化寄存器分配和线程调度，减少了数据搬运的开销．这些优化给DeepSeek-V3模型带来了显著的性能提升，实现了比Meta高出10倍的GPU计算效率．通过直接控制寄存器和线程调度，充分发挥了GPU潜力；通过对特定硬件(如H800)进行深度优化，获得了极致的性能表现．然而，由于PTX代码更接近汇编语言，需要开发人员具有深厚的硬件知识和编程能力；代码难以阅读和维护，这令程序的可维护性差，不利于团队协作和后续的升级；可移植性低，针对特定硬件优化的PTX代码难以在不同型号的GPU之间迁移．因此，这种方法并未被广泛采用．

### 1.7 数据并行和模型并行

DeepSeek-V3模型的并行策略非常复杂，包括16路流水线并行、跨8个节点的64路专家并行等．DeepSeek-V3模型创新地引入了双向流水线并行算法DualPipe，显著减少了流水线停滞现象，并实现了计算与通信阶段的重叠，从而大大提高了GPU利用率并减少了通信开销．在专家并行方面，DeepSeek-V3模型由256个路由专家和1个共享专家组成，每个token会激活8个专家，并确保最多被发送到4个节点．这种多层次的并行策略不仅能够充分利用硬件资源，还可大幅提高训练效率，使DeepSeek-V3模型能在较短时间内完成大规模模型的训练．此外，模型在软硬件架构联合设计、内存和计算能力的合理调配以及负载均衡策略上也已达到了极致．

通过这些技术的综合应用，DeepSeek在有限的GPU资源和较短的训练时间内，成功训练出了通用语言大模型DeepSeek-V3．

## 2 创新性新算法GPRO的应用：从DeepSeek-V3到DeepSeek-R1-Zero

大模型首先需要经过预训练，此过程非常昂贵，需要海量的训练数据集、足够大的计算机群，以及相当长的训练时间．预训练目的是将海量的训练数据中的知识压缩到大模型的上亿个参数中，得到一个通用的LLM，如DeepSeek-V3和ChatGPT4.0．尽管这种通用语言大模型几乎无所不知，但推理能力仍是有限的．

### 2.1 监督微调和强化学习

为增强大模型的推理能力，人们开发了多种训练方法，其中最重要和常用的是有监督微调(supervised fine-tuning，SFT)和强化学习(reinforcement learning，RL)．

SFT是在预训练模型的基础上，使用标注数据进行进一步训练，以提升模型在特定任务或领域上的表现．SFT需要大量高质量、标注好的特定任务数据，这需要雇佣专业人员进行数据标注和处理，此过程既耗时又昂贵．SFT过程也需要占用大量计算资源，为达到理想效果，还可能需要进行多次迭代和优化，这进一步增加了成本．因此，业界常说："天下苦SFT久矣．"

RL是一种机器学习方法，大模型通过与环境

交互，根据环境反馈的奖励信号，学习最优策略以最大化累积奖励．在大模型后训练中，将 RL 与人类反馈（human feedback，HF）相结合的 RLHF 方法更为常用．

RLHF 和传统 RL 方法在框架和优化策略与迭代式的学习上相似，但在奖励来源、学习目标和训练过程上有所不同．传统 RL 方法依赖于预定义规则或环境，而 RLHF 方法通过将 HF 转化为奖励，训练奖励模型以预测人类偏好，使模型输出更符合人类价值观．训练过程中，RLHF 包含预训练、奖励模型训练和 RL 微调等阶段，更适用于难以用算法定义但人类却易判断其质量的任务，如生成引人入胜的故事．

与 RL 相比，RLHF 需要大量高质量的 HF 数据，还需多次执行模型训练和部署，因此会消耗更多计算资源，其成本有时甚至比 SFT 更高，这对于资源有限的企业来说是一个挑战．

### 2.2　群体相对策略优化算法

2024 年 2 月，DeepSeek 团队提出了一种创新的 RL 算法——群体相对策略优化（group relative policy optimization，GRPO）[5]．该算法旨在提升 LLM 的推理能力，在数学和编程等复杂任务中表现尤其突出．GRPO 算法的主要特点是不依赖独立的价值函数模型，而是通过用多个输出的平均奖励作为基准进行优化．该算法不仅简化了模型的训练过程，还减少了内存消耗和计算开销，在某些任务上更是取得了显著的性能提升．

DeepSeek-R1-Zero 模型采用 GRPO 算法，完全跳过了消耗计算时间和资源的 RLHF 和传统的 SFT 过程，使训练过程在高效和低耗方面都效果显著．在 AIME2024 测试集上，模型得分从 15.6% 提升至 71.0%，展现了出色的性能和资源节省能力．

### 2.3　DeepSeek-R1-Zero 模型的意义

DeepSeek-R1-Zero 模型通过纯 RL 从一个基模型开始训练．训练过程中，首先给模型一些提示（prompt），要求它在 2 个 thinking 标签之间进行思考，并在 2 个 answer 标签之间给出答案．然后，根据最终结果的正确性和格式作为奖励（reward），进而优化模型行为．随着训练步骤的增加，DeepSeek-R1-Zero 模型逐渐涌现出长思维链（chain-of-thought，CoT）能力，令推理路径变得越来越长．此外，模型在训练过程中还会出现“顿悟时刻（aha moment）”，即自我发现并修复以前的推理错误，此“顿悟时刻”可视为涌现能力的一种具体表现．

这种纯 RL 的训练方法展示了 DeepSeek-R1-Zero 模型无需 HF 即可提升推理能力的突破．DeepSeek-R1-Zero 模型完全抛弃了预设的思维链模板和 SFT，仅依靠简单的奖惩信号来优化模型行为，打破了传统训练方法对人类标注数据的依赖．

令人惊艳的是，在训练过程中，DeepSeek-R1-Zero 模型又展现出了类似人类的反思、探索和多步验证等复杂推理行为的替代解决方案，具备了通用人工智能（artificial general intelligence，AGI）的重要特征——自主学习能力．

以上自我纠错和深度思考的能力，令 DeepSeek-R1-Zero 模型可根据问题的复杂度自然调节相应长度，表明该模型真正理解了所提问题的难度．这些复杂行为充分展示了 AI 系统在没有明确编程的情况下自主开发高级问题解决策略的能力．

## 3　监督微调+强化学习：从 DeepSeek-V3 到 DeepSeek-R1

DeepSeek-R1 的训练过程采用阶段策略．首先是冷启动，利用数千个长思维链（CoT）样本对基础模型进行 SFT，以提供初始的推理能力；接着是面向推理的强化学习，通过大规模强化学习提升模型的推理能力，尤其是在编程、数学、科学和逻辑推理任务上；然后是重构和数据生成，利用拒绝采样和 CoT 提示生成高质量训练数据，包括推理和非推理任务数据；最后是最终进化，再次进行 SFT，并引入人类偏好奖励，以提高模型的通用能力、可用性和安全性．

这种训练策略的优点有：通过设计易读的输出格式，过滤掉不友好的响应，提高了输出的可读性；通过人类先验设计的模式，增强了模型的推理能力；平衡了模型的多种能力，在提升推理能力的同时，也注重了模型的通用性和安全性．当然，这种策略也存在缺点，如训练过程复杂，需要更多的时间和资源管理；因引入了人类设计的模式，可能给输出带来偏见．

DeepSeek 系列模型取得成功的关键在于：① 合理的项目规划，通过将研发周期细分为多个阶段，每个阶段都有明确的目标和时间节点；② 有效的团队协作，采用多智能体协同学习机制，提高了团队的整体效率；③ 创新的算法设计，如引入

多智能体协同学习和经验回放技术，提升学习效率；④平衡监督学习和强化学习，通过分阶段训练策略，结合了两种学习方法的优势．这种训练方法不仅提高了模型性能，还显著降低了训练成本，为AI行业的发展提供了新的思路．

## 4 结 论

1）DeepSeek系列模型在AI领域的突破主要体现在模型和算法的创新、软硬件协同优化，以及整体训练效率的提升．DeepSeek-V3模型采用MoE架构，通过细粒度设计和共享专家策略，实现了高效的计算资源利用．MoE架构中的稀疏激活机制和无辅助损失的负载均衡策略显著提高了模型效率和性能，尤其是在处理大规模数据和复杂任务时．创新的MLA机制通过减少内存使用和加速推理过程，在处理长序列时表现出色，降低了模型的训练和推理成本．

2）在训练DeepSeek-V3模型时，团队引入了MTP和FP8混合精度训练等技术．MTP通过一次性预测多个词汇，显著提升了模型的上下文理解能力和训练效率．FP8混合精度训练策略采用8位浮点数表示数据，使模型在保持高性能的同时，降低了内存消耗和计算开销．为最大限度地提高GPU计算效率，团队还直接编写和优化PTX代码，使模型的计算效率远超竞争对手．这些创新方法有效提升了模型的整体训练效率，提高了单位GPU小时的计算利用率．

3）在对DeepSeek-R1-Zero模型的训练过程中，团队采用了全新的强化学习算法——GRPO，跳过了传统的SFT和RLHF阶段．GRPO算法通过优化多个输出的平均奖励，简化了训练过程，显著提升了模型的推理能力．DeepSeek-R1-Zero模型又通过纯强化学习，从1个基模型开始训练，展示了其在自我学习和推理方面的突破能力．

总之，DeepSeek团队在有限的、相对低效的GPU资源下，在较短的时间内成功训练出世界一流的开源推理大模型，为全球的AI研发开创了新的道路，这一突破不仅证明了DeepSeek团队的技术实力，也预示着未来AI格局的彻底改变，开启了一扇通往无限可能的大门．

**作者简介：**张慧敏(zhanghm88@hotmail.com)，教授、博士，自由撰稿人．

**引　　文：**张慧敏．DeepSeek-R1是怎样炼成的?[J]．深圳大学学报理工版，2025，42(2)：226-232.

**参考文献 / References：**

[1] GitHub. DeepSeek-V3 technical report [R/OL]. (2024-12-24) [2025-02-01]. https://github. com/deepseek-ai/DeepSeek-V3/blob/main/DeepSeek_V3.pdf.

[2] GitHub. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning [R/OL]. (2025-01-25) [2025-01-01]. https://github. com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. (2017-06-12) [2024-02-01]. https://arxiv. org/pdf/1706. 03762.

[4] GLOECKLE F, IDRISSI B Y, ROZIÈRE B, et al. Better & faster large language models via multi-token prediction [EB/OL]. (2024-04-30) [2025-02-01]. https://arxiv. org/pdf/2404.19737.

[5] DeepSeekMath: pushing the limits of mathematical reasoning in open language models [EB/OL]. (2024-02-05) [2025-02-01]. https://arxiv.org/pdf/2402.03300.

【中文责编：英子；英文责编：木柯】