# 面向强推理LLM的强化学习进阶：优化、泛化与挑战

陈红阳

2025年3月

之江实验室 ZHEJIANG LAB

之江实验室 ZHEJIANG LAB

# DeepSeek：在正确且困难的事上深度求索

**DeepSeek-V3：MoE架构，671B的大语言模型**

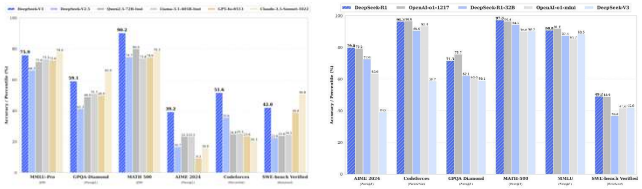
- **训练成本低：**
  - 五百万美元，仅为GPT4的5%

| Training Costs | Pre-Training | Context Extension | Post-Training | Total |
|---|---|---|---|---|
| in H800 GPU Hours | 2664K | 119K | 5K | 2788K |
| in USD | $5.328M | $0.238M | $0.01M | $5.576M |

- **推理性能好：**
  - Benchmark测试持平或超过当时所有主流模型
- **价格优势大：**
  - 约GPT-4o的10%

**DeepSeek-R1：V3+微调+强化学习的深度思考模型**

- 在数学、代码和推理任务中实现了与 OpenAI-o1 相当甚至更好的性能
- 证明了强化学习与Test Time Scaling的有效性



2

2024年12月27日DeepSeek发布了V3，彼时这款产品还没有进入大众视野，但在AI圈子里已经引发了不小的震动。对于DeepSeek，我认为“深度求索”这个名字很贴切，他们的产品完美体现了“在正确且困难的事上深度求索”的工程理念。其实他们在V3上的技术并非颠覆性的，但都非常实用。在技术报告中，可以看到他们把几乎所有技术细节都往前推进了一点点，这些一点点的积累就从量变引出了质变。因此与其说DeepSeekV3是技术的飞跃，不如说是工程的胜利。
V3采用的是MoE架构，总共671B的参数，每次推理将激活其中37B的参数。
它具有如下特点：
训练成本低。整个训练共花费约五百万美元，仅为GPT4的二十分之一。
推理性能好。在包括数学、知识问答、代码等的各项Benchmark测试中几乎都获得了持平甚至超过当时主流大模型的得分。
价格优势大。对于用户来说使用成本仅为GPT-4o的十分之一，而且在这么低的使用成本前提下，官方仍披露其理论成本利润率超过500%，这无疑是因为DeepSeek对模型的优化已经炉火纯青。

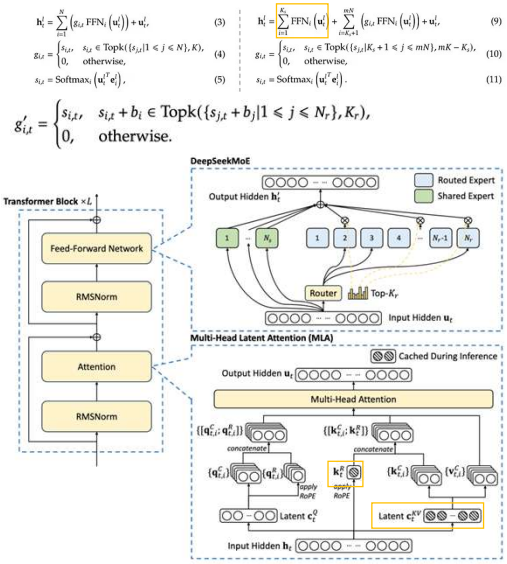
后续1月20日发布的DeepSeek-R1则真正成为了爆款，可谓是2025年农历新年之前最有话题度的产品。它真正打破了OpenAI在推理任务上的垄断，是中国团队在推理模型领域对硅谷巨头的挑战甚至超越。
它是基于V3，以各种技术手段进行微调和强化学习后生成的模型，在数学、代

码和推理任务中实现了与OpenAI-o1 相当甚至更好的性能。更重要的是，它的出现证明强化学习与Test Time Scaling在模型后训练阶段的有效性，其中Test Time Scaling更是自此成为了继Model Size Scaling后最重要的增强模型能力的方式。后面我们将对V3和R1采用的技术进行详细介绍。

# DeepSeek-V3：模型优化

采用MoE（混合专家）和MLA（多头潜在注意力机制）架构实现高效推理及降本增效：

- MoE：
  - **更细粒度的专家分割**：扩大专家数量，减小专家规模
  - **共享专家隔离**：任务强制激活共享专家，提升性能，降低冗余
  - **负载均衡**：根据调用调整bias，均衡负载
- MLA：
  - **压缩输入向量**：大幅降低显存占用
  - **解耦式RoPE**：引入额外查询向量供RoPE使用，避免单独缓存键值投影向量，进一步降低显存






DeepSeek-V3在模型优化中下了不少功夫，最有代表性的是MoE和MLA两项技术，凭借这两项技术DeepSeek-V3成功实现了高效推理和降本增效。

MoE。MoE并非DeepSeek首先提出，它的理念源于Hinton 1991年的论文。2017年他和同事一起在LSTM上实现了第一个MoE系统。它的特点是把Transformer中的前馈神经网络（FFN）用MoE层替换。所谓MoE层就是一个门控（gating）操作和一组小型FFN层组成的层。

MoE在大模型中的应用所面对的主要问题有：

1. 其性能高度依赖门控机制的有效性，常常出现少数专家处理大部分输入token，其他专家因缺乏训练机会，无法充分发挥作用，且形成恶性循环的情况。这被称为“专家崩溃”（Expert Collapse）。

2. 权衡专家间的知识聚焦（expert specialization）与知识共享（knowledge sharing）。我们要求专家在自己的专业领域表现出色，但同时也应该具有一定的通用性（不能丧失基本的协作能力）。过度专精与过度冗余都不是理想状态。

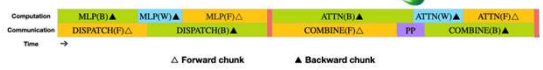
针对这些问题，DeepSeek提出了以下解决方案：第一是更细粒度的专家分割。将专家数量扩大，相应减少每个专家的参数量，使得专家的知识更加聚焦，但不增加计算负担。第二是共享专家隔离，选定其中几个专家为共享专家，每个token的计算都会经过共享专家，从而保持MoE的通用性。第三是无损失辅助的负载均衡，根据训练中专家的调用情况调整专家调用的偏置项，以保证负载均衡。

MLA则是为了解决MHA（多头注意力机制）在对单Token处理中必须要重新进行大量Q、K、V矩阵计算的问题催生的技术，属于对KV-Cache（键值缓存）技术的发展。MLA对原有键值缓存技术的优化主要有：

1. 将输入向量压缩为一个低维的潜在向量，在需要计算时再重新映射回高维空间，大幅降低显存占用。
2. 储存KV键，在后续计算中复用。
3. 解耦式RoPE：如果使用传统RoPE方法，RoPE将出现在Q、K之间，使得我们在做映射和拉回时必须储存额外矩阵。而通过解耦式RoPE，整个算法将只需要储存方框中的两个矩阵，进一步减少显存开销。

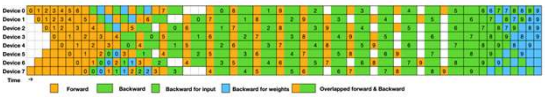
之江实验室 ZHEJIANG LAB

# DeepSeek-V3：训练优化

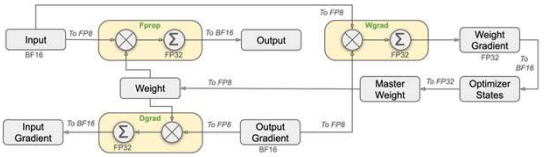



设计重叠策略，隐藏通信开销

- 并行策略：
  - 64路专家并行，16路流水线并行，以及数据并行（ZeRO1），摒弃张量并行
  - 设计 DualPipe，解决并行通信开销问题
  - 开发高效通信内核，充分利用通信带宽
- FP8混合精度：
  - 设计FP8 混合精度训练框架，在大规模模型训练上验证了其**可行性**和**有效性**
  - FP8负责计算密集操作，原始数据格式负责关键操作，平衡训练效率和数值稳定性
  - 实现加速训练和降低 GPU 内存使用




多路流水并行，减少训练气泡




线性混合精度，实现加速训练

4

DeepSeek在训练优化上的探索主要集中在多Token预测（MTP）、并行策略和混合精度计算上。

并行上，DeepSeek使用了16路流水线并行，跨越8个节点的64路专家并行，以及由ZeRO1支持的数据并行。DeepSeek放弃了张量并行，最主要的原因是张量并行会在多个设备间频繁进行张量的切分、合并，通信量会随模型规模和数据量显著增加，降低训练效率。

另外，DeepSeek设计了名为DualPipe的流水线并行策略，包括双流并行计算优化与双向流水调度优化。

双流并行的基本思想是将计算流和通信流重合，让训练Barrier刚好停在两个流任务完成时，避免等待，提高计算效率。

双向流水调度是对传统单向流水调度的优化，同时从流水线起始端和末端进行微批次（micro-batch）处理，减少气泡。

DS采用的通信方案是All-to-all，这样的优势是显存开销小，但通信效率低。因此V3在其上进行了许多优化，包括限制路由范围、优化网络拓扑、优化资源分配等。

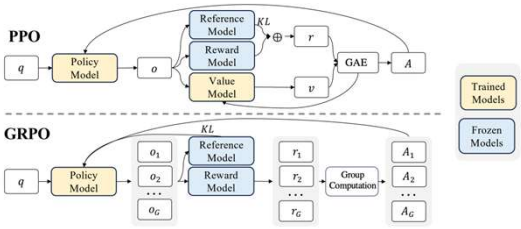
在混合精度（特别是FP8）训练方面，V3设计了混合精度训练框架，将大量GEMM（通用矩阵乘）操作采用FP8进行，但对Token嵌入、注意力操作等依然保持高精度（BF16/FP32）计算。特别地，当在进行数据累加的时候，FP8数据

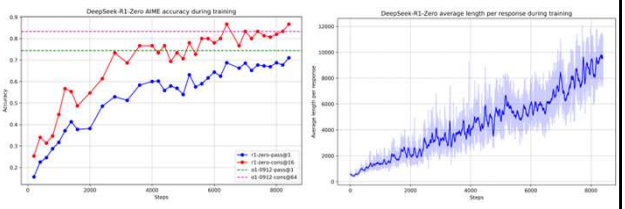
也会使用BF16等格式进行累加，以控制量化误差的积累。
当然V3还有很多其他优化技巧，如选择重计算、EMA显存优化等等。

之江实验室 ZHEJIANG LAB

# DeepSeek-R1-Zero：强化学习解锁思维能力

- 首次（公开）验证了纯强化学习在 LLM 中显著增强推理能力的可行性：
  - 无监督学习，仅通过 RL 即可使模型自发具备长链推理和反思等能力。
- 从**PRM**到**ORM**：基于规则的奖励设计
  - 正确性奖励
  - 格式奖励
- **GRPO**（Group Relative Policy Optimization）：
  - 分组采样
  - 相对优势估计
  - 替换传统的PPO算法
  - 无额外奖励模型，显著减少训练资源
- 以及搭配GRPO的一些小巧思：
  - 组相对优势估计(Group Relative Advantage Estimation):
    - $A_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)}$



- 训练持续推进，性能稳步提升
- 自我进化：学会以更多的思考来解决推理任务

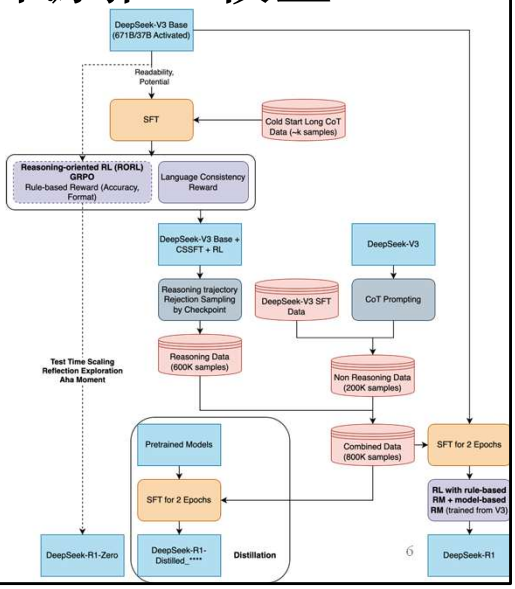
5

DeepSeek-R1的最大亮点之一莫过于它对强化学习的应用，这也是首次有论文公开验证仅使用强化学习也可以在大模型训练中显著增强其推理能力，并且让大模型自发地掌握长链推理和反思（主动回溯、顿悟等）能力。
为了达成这一目标，DeepSeek也对如何使用强化学习进行了多方面的改进。
首先对于强化学习来说，环境、算法、数据、奖励是缺一不可的。而可定制的奖励函数是其区别于其他非强化学习算法的一大显著特征。DeepSeek没有采用PRM，而是使用了ORM，是因为在实验中往往无法对中间步骤进行有效的界定，手动标注难以规模化。ORM则简单许多，只需要正确性与格式两个要点，而且在实验中效果很好。
DeepSeek也从常见的PPO转移至GRPO，其特点是使用分组采样和相对优势估计，避免额外训练奖励模型和价值函数，显著减少训练资源开销。PPO则需要Reward Model和Value Model，而且因为奖励劫持等问题通常效果不佳。
相应地，DeepSeek也加入了一些小巧思，比如，因为不再需要Value Model，也就不需要复杂的GAE估计，可以使用组相对优势估计进行反向传播，进一步减少开销和误差。
其结果是随着训练的持续推进，模型的性能可以稳步提升，并且显示出自我进化的特征：包括Aha Moment（主动反思以及顿悟）和主动领悟以更多的思考来解决推理任务。

之江实验室 ZHEJIANG LAB

# DeepSeek-R1: 引领时代的开源推理模型

- R1 = V3 + 冷启动 + 强化学习 + 多阶段训练
  - 解决 DeepSeek-R1-Zero训练不稳定、输出内容可读性差的问题
  - 更加用户友好，进一步提升推理能力。
- **少量SFT热身**：
  - 几千条长CoT数据提升推理性能
- **改进GRPO**：
  - 加入语言一致性奖励
- 拒绝采样和 SFT：
  - 生成高质量的推理（60万条）和非推理（20万条）数据，并用这些数据对模型进行微调。侧重点是提升模型的综合能力，使其在写作、事实问答等多种任务上表现良好
- 全场景强化学习：
  - 使模型在所有场景下都能表现良好，包括推理任务和非推理任务，并且保证模型的安全性和无害性
  - 此时引入基于模型的奖励模型（偏好模型），对齐人类偏好



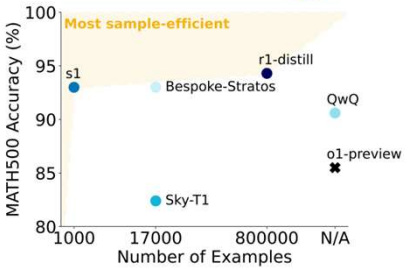

在R1-Zero的训练中仍有一些问题，其答案可读性不佳，并且存在较严重的中英文混合的情况。这导致模型的测试表现虽好，却不能作为一款成功的人类可用的产品推出。为了修正这些问题，DeepSeek结合了冷启动、强化学习以及多阶段训练，最终得到了DeepSeek-R1这个引领时代的开源推理模型。它对用户更加友好，答案更加可读，而且性能较R1-Zero又有进一步提升。

第一步是使用少量（几千条）长思维链数据热身，提升推理性能的同时增强模型的可读性。之后将模型放入改进后的GRPO，此时额外加入了语言一致性奖励，避免中英文混合。下一步使用训练完成的模型构造共八十万条高质量推理、非推理数据，并使用这些数据对V3再进行两个Epoch的微调，使其的综合能力，包括写作、事实问答等与推理无关的能力进一步增强。

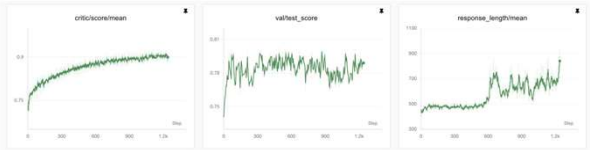
最后加入全场景的强化学习，以保证模型的安全性和无害性。这里既使用基于规则的奖励模型来衡量那些有明确答案的问题，也使用偏好模型来对齐人类偏好。此处加入偏好模型是较为常见的做法，原因也很简单：人类的偏好很难规则化，用模型进行拟合效果较好。

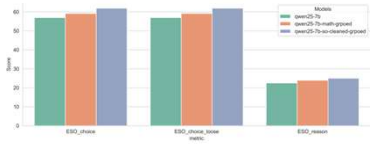
之江实验室 ZHEJIANG LAB

# 学界及我们的跟进工作




- SFT：少量高质数据 is All You Need
  - s1: Simple test-time scaling （1000条数据）
  - LIMO: Less is More for Reasoning （871条数据）
- RL：
  - Understanding R1-Zero-Like Training: A Critical Perspective
  - LIMR: Less is More for RL Scaling
  - DeepScaleR: Surpassing O1-Preview with a 1.5B Model by Scaling RL



7

在DeepSeek-R1发布后，可以说对LLM推理能力的增强研究进入了一个全新的阶段。因此在这里分享一些对DeepSeek-R1后续的跟进工作。

首先是对SFT的研究。这些研究打破了之前SFT需要的数据量比较多的认知，通过实验证明，也许在SFT中数量不是关键，质量才是关键。当然，这个前提是基础模型必须具备一定的能力，两篇文章选用的基座模型都是Qwen2.5-32B-Instruct。我们团队在7B上进行了实验，效果就不如论文那么好。

第二是对RL的扩展研究。第一篇文章指出基座模型对反思能力和Aha Moment的重要性，他们发现Aha Moment只在Qwen系列模型中出现，且出现在第一个Epoch的训练中，而在LLaMA系列的8B模型中不出现。并且对于参数较小的模型，很容易出现无效反思，回复长度的增加的直接原因是基于规则的奖励函数。因此，回答长度增加只是大模型推理能力增强的必要条件，而非充分条件。这和我们的实验观察到的也具有一致性：平均回答长度增加并不总是导致模型能力增强。

第二篇文章指出，对强化学习来说，数据也并非越多越好。而且他们发现，对于小模型（7B）来说，强化学习的效果比SFT更好。我们的部分实验结果也证明了这点。以地学奥赛题作为评判标准，在7B模型的微调中，小样本强化学习就可以提升模型约10%的性能，而SFT却做不到相应的能力提升。

第三篇文章则进一步强调了RL对于小参数大语言模型的有效性，并且给出一个具有参考价值的结论：逐步增加模型的有效上下文长度可以实现更有效的回复

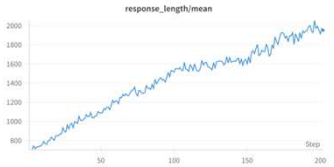
平均长度扩展，进而获得更长的推理思维链，从而有效提升模型。
还有很多重要的跟进工作，让我们对强化学习的认识更加深入，此处不再一一列举。

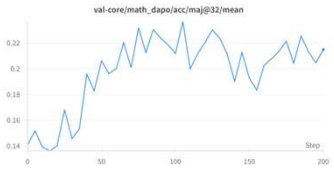
之江实验室 ZHEJIANG LAB

# 我们的跟进工作：对GRPO的改良训练

- 参考DAPO，针对GRPO做了以下改进：
  - **解耦裁剪范围的上下限**：避免熵崩溃，提高了系统的探索能力
  - **动态采样**：保证每组采样到有效的梯度信号（避免一个组里全是正确或全是错误答案，导致无效传播）
  - **词元级策略梯度损失**：在整个批次的所有词元上计算平均损失。每个词元的学习信号权重相对均等，不受其所在序列长度的影响。
  - **超长奖励调整和软性超长惩罚**：一方面过滤超长样本的奖励，另一方面加入惩罚函数避免长度增长过快，稳定训练。


为了应对奖励劫持（尽管模型在训练上奖励和平均长度一直上升，但在测试集上的表现却反而下降），我们采取Early Stopping，14B模型性能可与DeepSeek-V3（671B）相当。





8

之江实验室 ZHEJIANG LAB

# 本组跟进工作：跨语言、跨专业泛化能力实验

- 中文训练数据集：
  - TAL-SCQ5K：中英文数学竞赛试题（5千题）
- 中文测试数据集：
  - CMath（小学数学应用题）
  - Agieval-gaokao-mathcloze（高考填空题）
  - agieval-gaokao-mathqa（高考选择题）
- 医学训练数据集：
  - MEDQA-USMLE：美国医学执照考试题目（8千题）
- 医学测试数据集：
  - PubMedQA-L：1000道医学专家级MedMCQA：印度医学入学考试（约6150道）
  - PubMedQA-L：1000道医学专家级别判断题
  - XMedBench：从 MedQA、MedMCQA、MMLU 数据集精选了约 7330道题

不同模型在中文数学推理基准上的性能对比

| 模型 | CMath | 高考填空 | 高考选择 | 平均 |
|---|---|---|---|---|
| *大规模开源模型 (>70B)* | | | | |
| Llama-3.1-70B (few-shot) | 85.5 | 55.9 | 72.6 | 71.3 |
| Qwen2-Math-72B (few-shot) | 86.4 | 72.9 | 69.5 | 76.3 |
| Qwen2.5-Math-72B (few-shot) | 89.7 | 72.9 | 86.3 | 83.0 |
| *小规模开源模型 (7B)* | | | | |
| Qwen2.5-7B (zero-shot) | 76.9 | 34.7 | 47.3 | 53.0 |
| Qwen2.5-7B-Instruct (zero-shot) | 93.7 | 62.7 | 68.9 | 75.1 |
| Qwen2.5-Math-7B (zero-shot) | 75.7 | 43.2 | 61.8 | 60.2 |
| Qwen2.5-Math-7B-Instruct (zero-shot) | 93.3 | 78.8 | 75.8 | 82.6 |
| *基于规则的强化学习 (GRPO)* | | | | |
| Qwen2.5-7B-CN-Zero (zero-shot) | 90.7 | 55.6 | 71.1 | 72.5 |
| 提升幅度 | *+13.8* | *+20.9* | *+23.8* | *+19.5* |
| Qwen2.5-Math-7B-CN-Zero (zero-shot) | 89.8 | 80.5 | 81.2 | 83.8 |
| 提升幅度 | *+14.1* | *+37.3* | *+19.4* | *+23.6* |

注 1：数值表示百分比准确率；提升幅度是指相对于对应基础模型的绝对性能提升。
注 2：大规模开源模型 (>70B) 的数据来源于 Qwen2.5-Math 技术报告[76]。
注 3：大规模模型 (>70B) 采用少样本设置评估：CMath 为 6-shot，高考填空为 5-shot，高考选择为 4-shot；而所有 7B 模型均采用零样本 (zero-shot) 评估。

不同模型在医学推理基准上的性能对比

| 模型 | MedQA-USMLE | MedMCQA | PubMedQA-L | XMedBench | 平均 |
|---|---|---|---|---|---|
| Qwen2.5-7B | 45.7 | 50.1 | 56.1 | 55.8 | 51.9 |
| Qwen2.5-7B-Instruct | 55.2 | 53.6 | 52.9 | 60.2 | 55.5 |
| Qwen2.5-7B-med-Zero | 54.0 | 52.9 | 59.9 | 59.1 | 56.5 |
| 提升幅度 | *+8.3* | *+2.8* | *+3.8* | *+3.3* | *+4.6* |

注：数值表示百分比准确率；提升幅度是指相对于基础模型 Qwen2.5-7B 的绝对性能提升。

9

常见的GRPO训练集和测试集都为英文数据，为全面评估方法的有效性和泛化能力，避免语言上的偏差，我们在中文数学推理任务上也进行了实验。此处我们选用的训练数据集是TAL-SCQ5K，包括从小学到高中的数学竞赛题，其中三千题为选择题，两千题为填空题，题目有中英文两个版本，互为对照。我们选用了其中的中文数据子集作为训练集。
测试集我们采用C-Math（1700道中文数学应用题）、118道高考数学填空题和351道高考数学选择题。
作为补充，验证方法在专业领域推理任务上的有效性，我们选择了医学推理作为测试领域。相比数学推理，医学推理不仅需要逻辑推理能力，还需要大量专业知识和临床判断。在医学推理任务中，我们使用MEDQA-USMLE作为主要训练数据集，该数据集源自美国医学执照考试题目，包含复杂的临床场景和多选题。通过与DeepSeek-V3进行答案共识过滤，我们从原始的10718条数据中精选出约8000条高质量样本用于训练和评估。
除MEDQA-USMLE外，我们还使用了以下医学数据集作为评估基准:
（1）是来自印度医学入学考试的约6150道多项选择题数据集，专为解决实际医学入学考试问题而设计；（2）是包含1000个手动注释的是/否/可能问答数据，要求对生物医学研究文本进行推理，特别是其定量内容；（3）是从各医学问题数据集精选的约7330道多项选择题数据集。
实验结论—

Qwen2.5-7B模型在所有基准测试中均取得显著提升，平均性能提高19.5个百分点（53.0%→72.5%），其中高考选择题和填空题分别提升23.8和20.9个百分点。其数学专用版本Qwen2.5-Math-7B表现更为突出，平均提升23.6个百分点（60.2%→83.8%），高考填空题提升达37.3个百分点（43.2%→80.5%），超越了72B参数大模型，且参数量仅为后者的十分之一，验证了基于规则的强化学习方法对结构化推理任务的有效性。

值得注意的是，Qwen2.5-Math-7B-CN-Zero在零样本设置下达到83.8%的平均性能，分数与Qwen2.5-Math-7B-Instruct相当。这些结果表明，通过优化训练策略，小规模模型可在特定任务上超越大规模或专用模型。

实验结论二

Qwen2.5-7B-med-Zero 在医学基准测试中表现提升显著：

•平均提升 **4.6%**（51.9% → 56.5%），MEDQA-USMLE 提升 **8.3%**（45.7% → 54.0%）。

•相比数学任务，医学领域提升幅度较小，可能因医学知识更专业、复杂，依赖临床经验积累。

之江实验室 ZHEJIANG LAB

# 本组跟进工作：KL 散度惩罚系数 $\beta$ 的影响

- GRPO 算法中，KL 散度惩罚系数 $\beta$ 是一个关键的超参数，它控制了训练过程中模型偏离初始分布的程度。
- 较大的 $\beta$ 值使模型更倾向于保持接近初始策略，而较小的 $\beta$ 值则允许模型更自由地探索新策略。
- KL散度鼓励策略在多个好的动作上都有一定的概率，而不是只集中在一个动作上。

不同 KL 散度惩罚系数 $\beta$ 值对模型性能的影响

Table 4-9 Impact of Different KL Divergence Penalty Coefficient $\beta$ on Model Performance

| 超参数设置 | MATH-500 | GSM8K | AIME 2024 | AMC | Minerva | OlympiadBench | AIME2025 | 平均 |
|---|---|---|---|---|---|---|---|---|
| $\beta = 0.04$ | 60.4 | 84.8 | 6.7 | 30.1 | 33.5 | 25.8 | 0.0 | 34.5 |
| $\beta = 0.01$ | 60.4 | **91.0** | 6.7 | **39.8** | 29.8 | 31.7 | 0.0 | 37.1 |
| $\beta = 0.005$ | **60.6** | 88.5 | **10.0** | 36.1 | **35.7** | **32.2** | **3.3** | **38.1** |

注：所有列的数值均表示百分比准确率；每列的最高值以粗体标出。所有实验均在没有采用课程学习策略的情况下进行。


10

GRPO 算法中，KL 散度惩罚系数 $\beta$ 是一个关键的超参数，它控制了训练过程中模型偏离初始分布的程度。较大的 $\beta$ 值使模型更倾向于保持接近初始策略，而较小的 $\beta$ 值则允许模型更自由地探索新策略。KL散度鼓励策略在多个好的动作上都有一定的概率，而不是只集中在一个动作上。这有助于提高策略的鲁棒性和多样性。

在单一推理任务的 GRPO 场景下，较低的 $\beta$ 值（如 0.005 ）通常能带来更好的性能，特别是对于高难度的推理任务。这是因为 GRPO 算法本质上是朝着增大奖励的方向优化，即鼓励模型生成高质量的推理过程，而较低的 $\beta$ 值允许模型更充分地探索这一方向。

AIME2025 基准的结果尤其具有启发性——这是一个极具挑战性的测试集，包含了最新的高水平竞赛题目。在这一基准上，只有 $\beta = 0.005$ 的模型能够取得非零的成绩（ 3.3% ），而其他配置的模型完全无法解答。这进一步凸显了适当降低 KL 散度惩罚对于解决超高难度推理问题的重要性。

而从实验图中可以观察到以下关键趋势：
（1）回复长度变化：第一幅图展示了不同 $\beta$ 值对模型生成长度的影响。 $\beta = 0.04$ 的模型（黄色曲线）生成长度基本保持稳定且较短，这表明较大的 $\beta$ 值限制了模型偏离初始行为的能力。 $\beta = 0.005$ 和 $\beta = 0.01$ 的模型表现出更多的长度变化，能够探索更复杂的推理过程。 $\beta = 0.003$ 的模型（青色曲线）在训练后期（约 230 步后）出现了长度的急剧增长，虽然这表明模型具有很强的探索能力，但也增加了训练不稳定的风险。

（2）准确率奖励变化：第二幅图显示了不同 $\beta$ 值对准确率奖励的影响。总体趋势表明，较小的 $\beta$ 值（特别是 0.003 和 0.005 ）能够获得更高的准确率奖励峰值。这符合 GRPO 的基本原理：在单一推理任务场景中，较小的 $\beta$ 值允许模型更充分地朝着增大奖励（即生成高质量推理）的方向优化。 $\beta = 0.04$ 的模型获得了最低但最稳定的准确率奖励，表明过大的 $\beta$ 值会严重限制模型的优化空间。
（3）格式奖励变化：第三幅图展示了格式奖励的变化趋势。所有 $\beta$ 值配置都能在训练早期（约 50 步内）快速学习并遵循格式要求。值得注意的是，$\beta = 0.01$ 在训练中期出现了格式奖励的下降，同样 $\beta = 0.003$ 在训练后期也出现了类似情况，这表明较低的 $\beta$ 值在优化准确率的同时，可能会导致训练崩溃。

之江实验室 ZHEJIANG LAB

# 本组跟进工作：课程学习策略的影响

- 英文训练数据集：
  - MATH-lighteval：源自 MATH 数据集（7500题）
- 英文测试数据集：
  - MATH-500：包含不同难度的数学问题,覆盖代数、几何、数论等领域。
  - GSM8K:1319道小学数学题数据集。
  - AIME2024、AIME2025；
  - AMC数据集包含83道来自AMC12 2022和AMC12 2023（美国大学生数学竞赛）的题目；
  - Minerva-Math：包含谷歌研究团队提出的272道数学难题
  - OlympiadBench:675道奥赛级别数据集。

课程学习策略在英文数学推理任务上的消融实验结果

| 模型 | MATH-500 | GSM8K | AIME2024 | AMC | Minerva | OlympiadBench | AIME2025 | 平均 |
|---|---|---|---|---|---|---|---|---|
| 基于 *Qwen2.5-7B* 的实验 | | | | | | | | |
| 无课程学习 | 60.4 | 84.8 | 6.7 | 30.1 | 33.5 | 25.8 | 0.0 | 34.5 |
| 双阶段课程 | 64.6 | 79.8 | **16.7** | 33.7 | 38.9 | **31.4** | **6.7** | 38.8 |
| 五阶段课程 | **65.6** | **92.0** | 13.3 | **39.8** | **44.9** | 30.7 | 3.3 | **41.4** |
| 基于 *Qwen2.5-Math-7B* 的实验 | | | | | | | | |
| 无课程学习 | 74.0 | 81.5 | 23.3 | **55.4** | 41.5 | 38.1 | 10.0 | 46.3 |
| 双阶段课程 | 73.8 | 83.2 | **26.7** | 54.0 | 42.3 | **41.2** | **13.3** | 47.8 |
| 五阶段课程 | **75.2** | **85.3** | 26.7 | 54.2 | **44.1** | 40.0 | 10.0 | **47.9** |

课程学习策略在中文数学推理任务上的消融实验结果

| 模型 | CMath | 高考填空 | 高考选择 | 平均 |
|---|---|---|---|---|
| 基于 *Qwen2.5-7B* 的实验 | | | | |
| 无课程学习 | 88.3 | 51.7 | 66.4 | 68.8 |
| 双阶段课程 | **90.7** | **55.6** | **71.1** | **72.5** |
| 提升幅度 | *+2.4* | *+3.9* | *+4.7* | *+3.7* |
| 基于 *Qwen2.5-Math-7B* 的实验 | | | | |
| 无课程学习 | 85.2 | 74.6 | 79.5 | 79.8 |
| 双阶段课程 | **89.8** | **80.5** | **81.2** | **83.8** |
| 提升幅度 | *+4.6* | *+5.9* | *+1.7* | *+4.0* |

注：所有列的数值均表示百分比准确率；每组实验中的最高值以粗体标出。中文实验中采用了双阶段课程学习，按难度将训练数据划分为两个子集。

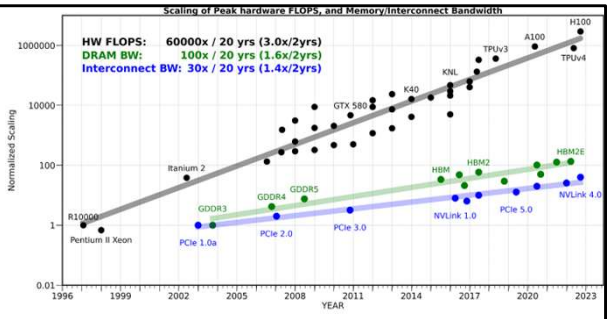
11

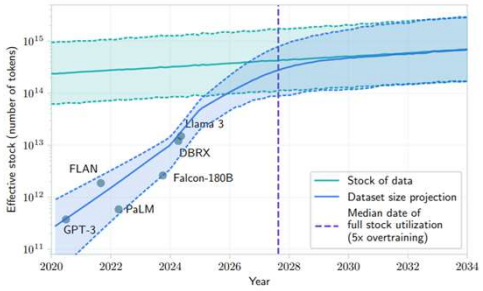
对于极端困难的任务，直接进行强化学习的效果是不佳的，模型很难给出有意义的答案。课程学习对解决极端难度任务具有明显的促进作用。在AIME2025这一极具挑战性的基准上，未采用课程学习的Qwen2.5-7B模型完全无法解答任何问题，而采用双阶段课程学习后，该模型成功解答了6.7%的题目，实现了从``完全无能力"到``初步具备能力"的质变。这表明课程学习能够帮助模型逐步建立解决超高难度问题所需的基础能力，进而实现在极端难度任务上的突破。

我们在中文数据集上同样进行了测试。结果表明，课程学习的有效性能够很好地跨语言泛化。在中文数学推理任务上，我们采用了将训练数据划分为两个难度子集的课程学习策略，为不同类型的模型都带来了显著提升。特别是在高考填空题这一具有挑战性的任务上，课程学习使 Qwen2．5－Math－7B 模型的准确率从 74．6%提升至 80.5%，提高了 5.9 个百分点。这一结果证明，课程学习作为一种基于认知理论的训练范式，其有效性不受语言边界的限制。

# 值得思考的问题

- **如何推进强推理模型的工程革命?**
  - 内存墙：DRAM带宽增长慢
  - 通信墙：通信带宽增长慢
- **如何深化强化学习的应用?**
  - 对GRPO/反思能力的深入探索
  - 探索其他可Scaling Up的强化学习方案：
    - MCTS
    - DAPO
- **如何创造更多的高质量数据?**
  - 提升数据质量
  - 新数据生成




12

AlphaGo的横空出世让很多非专业人员第一次了解人工智能的力量，DeepSeek-R1的出现则让更多人享受到了人工智能带来的革命性变革，而两者背后，有着诸多相似之处：工程革命和强化学习。没有GPU参与模型训练和后续Google开发的TPU，就不会有MCTS的真正大规模应用，也就不会有AlphaGo；没有DeepSeek对H100性能的极限压榨和显存优化方法，就不会有MoE+GRPO的技术路线，也就不会有DeepSeek-R1。

大模型的Scaling Law并未失效，更大的模型、更长的回答、更多更好的训练数据总是指向更好的表现。而限制Scaling Law的无疑是硬件。不仅仅是GPU算力，内存带宽和通信带宽也越来越成为Scaling Law的枷锁。GPU在计算时必须把数据从DRAM送至计算核，而内存带宽的提升速度远小于计算能力的提升速度。并行也不可避免地引入GPU间和节点间的通信，通信带宽成为限制性能的一个越来越重要的原因，通信带宽的增长甚至更慢于DRAM的增长。因此DeepSeek花了大量精力优化它们。在DeepSeek之后，我们还能做些什么？

第二方面是对强化学习的进一步探索。首先，对于以GRPO为代表的强化学习算法，还有很多问题我们无法回答，包括一些最基本的问题：反思到底是Base模型本身即具备的，还是由GRPO赋予的？是只有GRPO才能带来这种能力，还是其他的强化学习、甚至SFT都能带来这种能力？出现这种能力的必要条件是什么？参数量和长度的阈值是多少？大模型是不是可以像人一样，把书读厚（通过GRPO增加长度）再把书读薄（强制大模型用更简洁的推理解决问题），进一步

增强能力？
MCTS无疑是一个好算法。它在围棋上是有效的，因为每步剪枝以后可选的答案数量不多。而在大模型中，目前Token Level的MCTS将会面对很大的工程困难，因为可选答案太多，而Step Level的MCTS又需要人工标注的优质数据，两条路线都存在无法scaling up的问题。如果有办法Scale Up MCTS，很可能再带来一场LLM领域的革命。
也有很多基于GRPO的改进，比如DAPO，提出使用动态采样的方式来提高训练过程中数据的质量，提升训练的效果。还有很多有待探索的强化学习方案，但可Scaling Up始终是最重要的。
更大的问题是数据：R1及其后续工作再一次证明了最终大模型的训练仍将依靠数据，特别是高质量的数据。数据决定上限，算法逼近上限，诚如所言。预计在2027年人类原创数据就将被全部囊括进大模型的预训练中。理论上，所有的思维链都会隐藏在这些预训练数据中，大模型只是没能显式地提取出这些思维模版。GRPO则很可能帮助大模型将这些思维模版提取了出来，下一步就是归纳、强化、泛化这些思维模版，对数据质量的提升关键也将集中于这些领域。
如何创造更多的数据？在未来，新数据很可能是人机共创，甚至是独立由大模型完成的。实际上，大模型已经深度参与现阶段我们开展的许多数据相关工作。而且大模型已经在各个领域开始解决人类无法解决的问题。期待着大模型生成新知识被工程化的那一天。

# 谢谢

欢迎交流，集思广益，共同探讨

之江实验室 ZHEJIANG LAB

13

结束语