

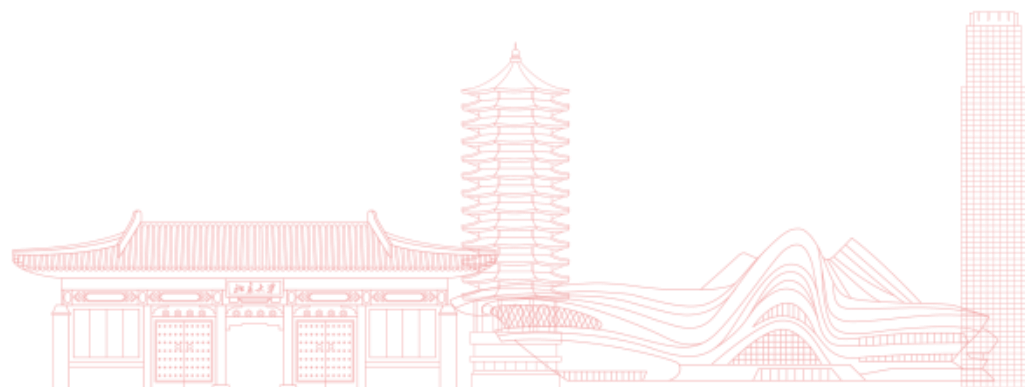


北京大学
长沙计算与数字经济研究院
PKU-Changsha Institute for Computing
and Digital Economy

《DeepSeek正是为王的理由》

本报告分享探讨近期备受瞩目的DeepSeek大模型家族的发展历程及其主要应用。DeepSeek家族属于大语言模型领域中的一个细分方向，即“推理模型”，指能够进行逻辑推理和复杂决策的大语言模型。本次汇报重点介绍推理模型的发展动力、应用场景和技术原理，并在最后分享一些关于DeepSeek的相关资料供参考学习。

汇报人：姚文翰



• 2024年9月6日

目录

一、deepseek 是什么？	1
二、大语言模型（Large Language Models , LLMs）前置知识.....	1
三、Deepseek 各个模型之间的关系	1
四、推理模型（Reasoning Model）的含义	2
五、何时使用推理模型？	2
六、Deepseek 模型家族简介	2
6.1 基于人类反馈的强化学习(RLHF).....	2
6.2 DeepSeek-R1-zero.....	2
6.2.1 DeepSeek-R1-zero： GRPO 决策.....	2
6.2.2 DeepSeek-R1-zero :自我进化.....	2
6.3 DeepSeek-R1.....	2
6.4 DeepSeek-R1-zero 与 R1 的实验结果.....	2
6.5 DeepSeek-V3	2
6.5.1 GPT 注意力	2
6.5.2 DeepSeek-V3 的注意力模块 Multi-head Latent Attention	2
6.5.3 DeepSeek-V3 的混合专家模块 MoE.....	2
七、构建“推理时拓展”（Inferencing Time Scaling ）能力	2
7.1 什么是推理时拓展？	2
7.2 Deepseek 的推理时扩展的方法： 蒸馏.....	2
7.3 Deepseek 蒸馏模型的效果.....	2
7.4 Deepseek 实际使用测试.....	2
八、结语	3

一、deepseek是什么？



杭州深度求索人工智能基础技术研究有限公司存续

科创分:

报告

笔记

关注

77万+ | 1小时38分钟前更新

统一社会信用代码：91330105MACPN4X08Y

法定代表人：裴浩 关联企业 3

注册资本：1000万元

成立日期：2023-07-17

电话：0571-85377238

邮箱：service@deepseek.com

官网：https://www.deepseek.com

地址：浙江省杭州市拱墅区环城北路169号汇金国际大厦西1幢1201室 附近企业


企查查行业：

企业规模：

员工人数：

简介：DeepSeek成立于2023年，是一家通用人工智能模型AGI研发商，专注于开发先进的大语言模型（LLM）和相关技术，以满足人工智能领域的各种需求。目前DeepSeek的模型...[展开](#)

核心人员1

序号	姓名	职位	简介
1	 梁文锋	创始人	DeepSeek创始人、幻方量化创始人。本硕就读于浙江大学，攻读人工智能。

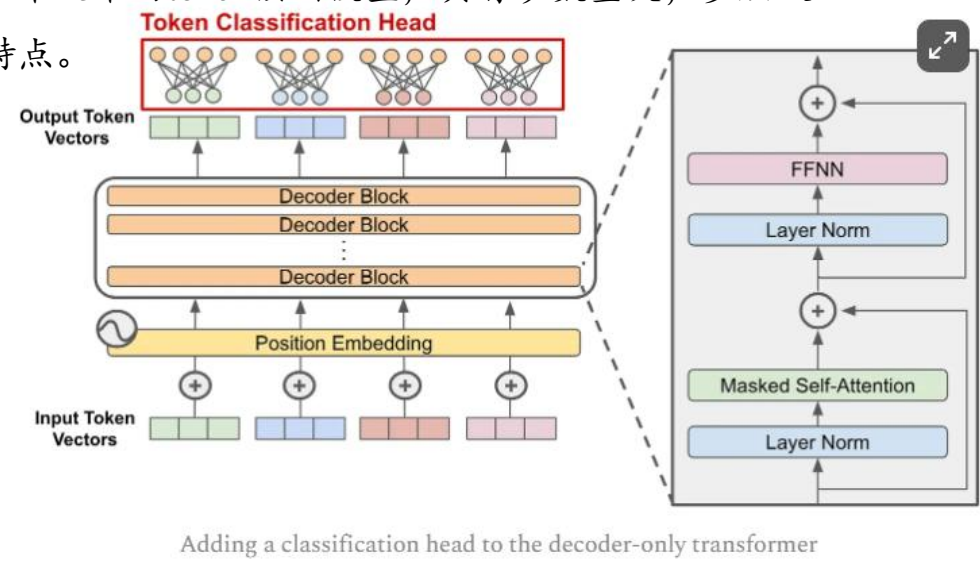
DeepSeek的论文时间线与技术演进对比

时间	论文研究内容	研究路线对应阶段
2024年1月	DeepSeek LLM	提出MoE架构，优化计算效率，提高推理能力
2024年2月	DeepSeekMath	采用GRPO进行RL训练，提升数学推理能力
2024年5月	DeepSeek-V2	采用MLA机制减少KV存储，提升长文本推理能力
2024年6月	DeepSeek-Coder-V2	采用MoE结构和RL训练优化代码推理能力
2024年12月	DeepSeek-V3	进一步优化MoE结构，结合RL和蒸馏
2025年1月	DeepSeek-R1-Zero	完全基于强化学习（RL），不依赖SFT训练
2025年2月	DeepSeek-R1	结合SFT和RL训练，优化可读性和任务泛化能力

二、大语言模型 (Large Language Models , LLMs) 前置知识

(一) LLM

是一个文本的token预测模型，具有参数量大，多层级的特点。



- (a) 无监督训练：仅需从文件里随机采样一个句子，也就是x。称为无监督训练。y可由x得到。
- (b) 有监督训练：需从文件里采样2个句子，x和y互相独立。为有监督训练。

(二)、训练与推理。

无监督训练的训练阶段：

输入x：“我爱中华人民共和国”
标签y：“爱中华人民共和国\end\” 或者 “那你爱哪个国家？”
最大化概率 $P(y | \text{Model}(x))$ 。

串行推理阶段：

输入：“帮我计算1+2等于几。”

输出：
帮我计算1+2等于几。1
帮我计算1+2等于几。1+
帮我计算1+2等于几。1+1
帮我计算1+2等于几。1+1等
帮我计算1+2等于几。1+1等于
帮我计算1+2等于几。1+1等于2
帮我计算1+2等于几。1+1等于2。|end|

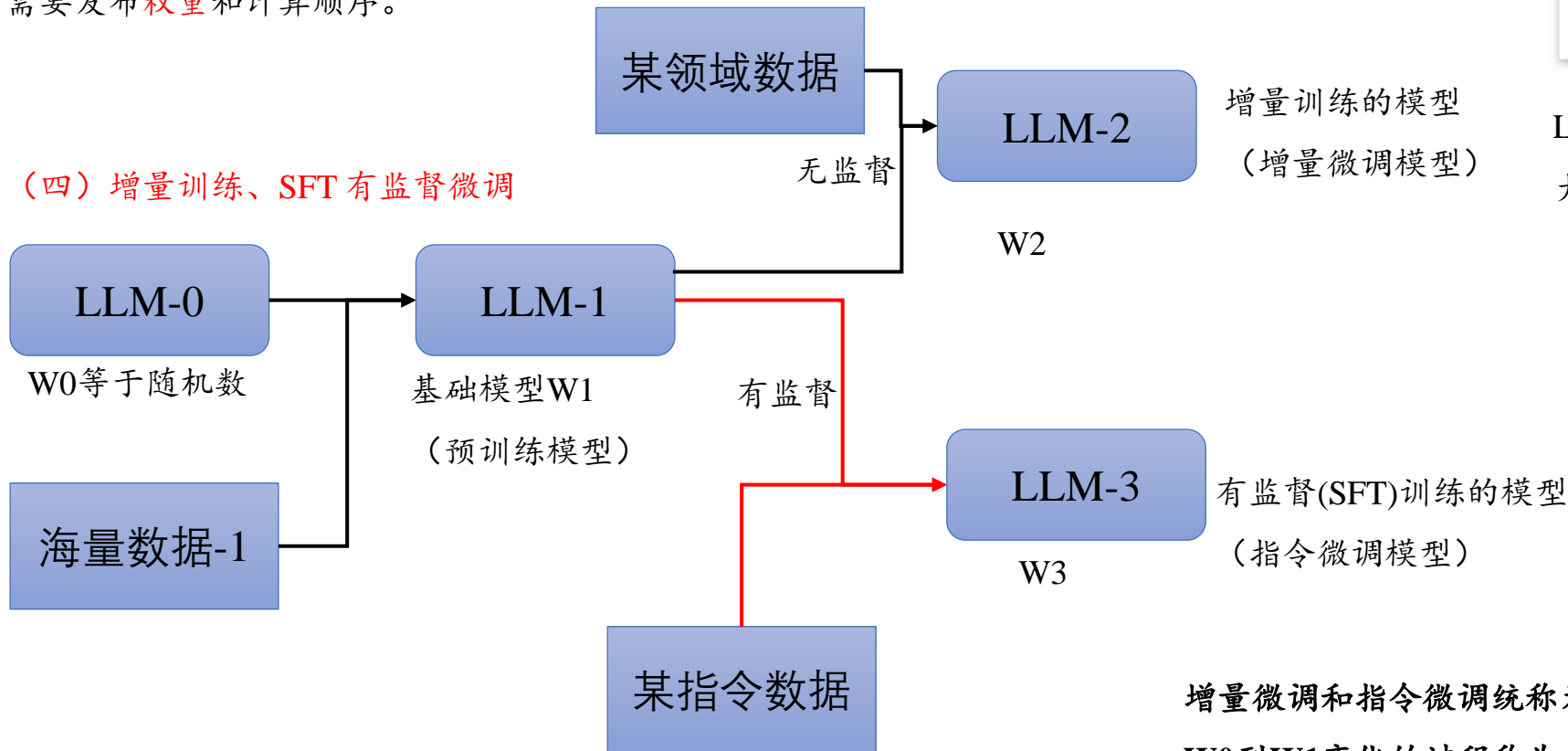
每次计算新token的时候，可以复用前面已存储的token序列。因此存在所谓“KV cache”
降低KV cache、计算复杂度、同时保证生成效果，是LLM的核心发展道路。Deepseek也不例外。

二、大语言模型 (Large Language Models , LLMs) 前置知识

(三) 权重 (weight)

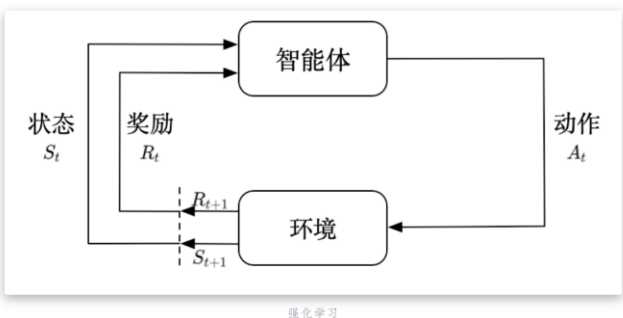
LLM从输入x到计算y的过程本质上是巨量的 $y = W * x + b$ 的矩阵计算和非线性函数计算。每个W都称为“权重”。整个模型的权重以一定的计算顺序组合在一起构成了LLM。因此，存储、开源LLM的时候，需要发布权重和计算顺序。

(四) 增量训练、SFT 有监督微调



增量微调和指令微调统称为“微调”
 W_0 到 W_1 变化的过程称为“训练”
如果使用W的时候，W一直不变化，称为“推理”

(五) 强化学习

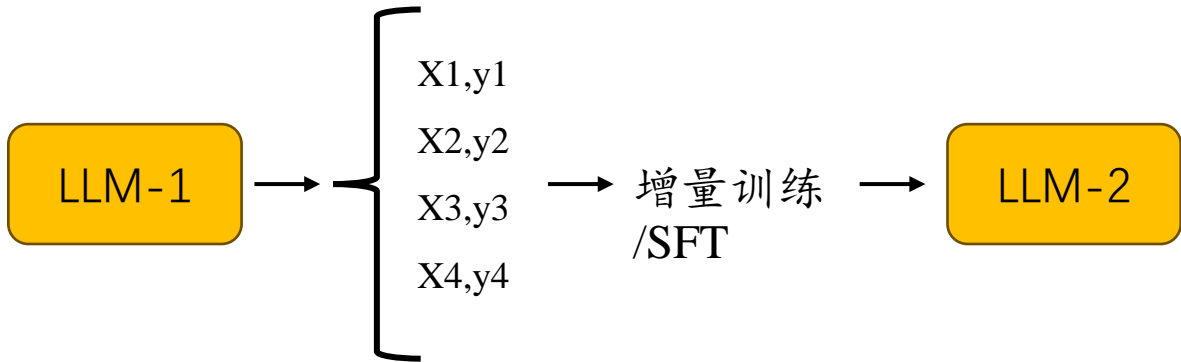


LLM领域里，强化学习的智能体是大模型，动作就是输入、输出文本。

二、大语言模型 (Large Language Models , LLMs) 前置知识

(六) 蒸馏 (Distill)

蒸馏是让一个预训练的LLM-1产生一批数据，然后拿这批数据去 训练/SFT 一个新模型的过程。相当于老师教学生。



(六) 常见大模型名称:

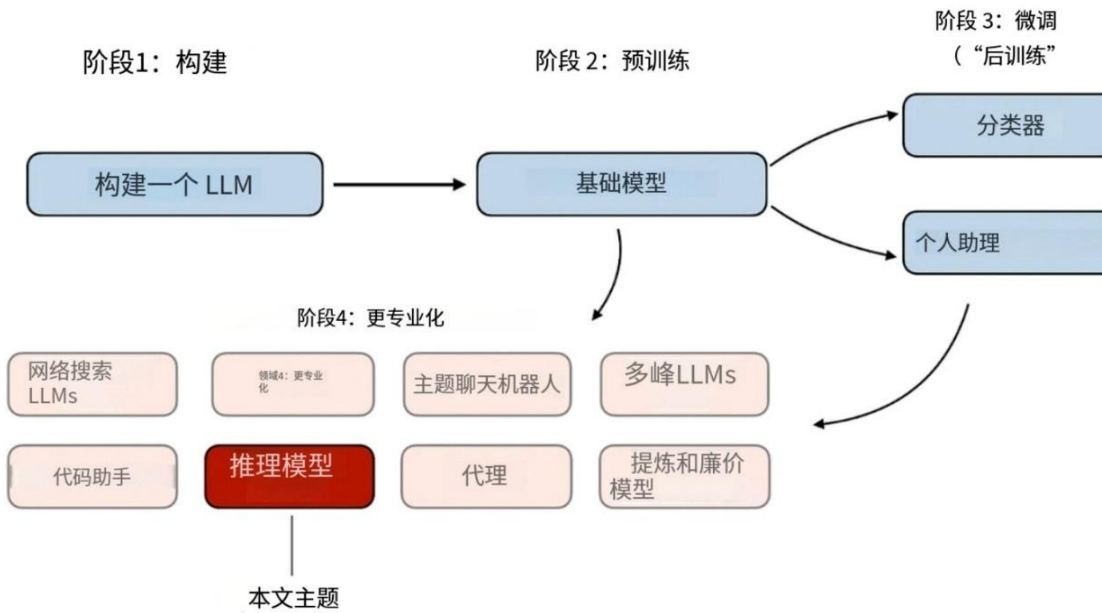
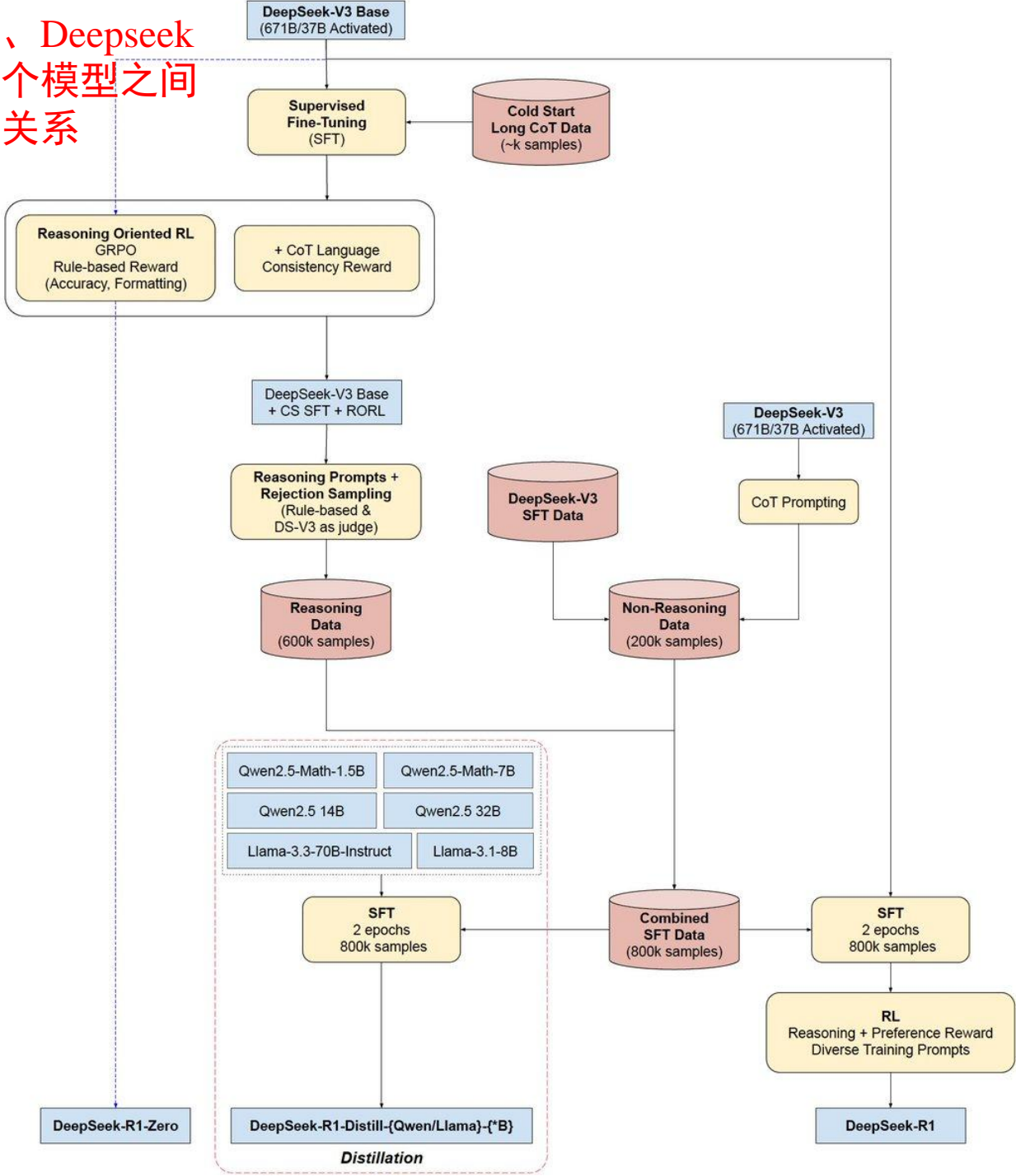
模型名称: qwen2.5-*B。这种为在海量文本数据上预训练模型。*B 表示其模型权重的中浮点数的数量。 通称“基础模型”。 具备基础的语言能力和基础知识。

模型+数据集名称: DeepSeekMath Qwen2.5-Coder 。这种为基础模型在特殊数据集上进行微调的模型。 具备基础的语言能力和特定领域的知识。

模型+instruct名称: Qwen-72B-Instruct 等。 用指令数据（同时有x和y的数据）微调基础模型后产生的模型。具备完成指令任务的能力和基础知识。（如摘要、翻译、解释、记忆、推理等。）

后浪模型: DeepSeek-R1-Zero DeepSeek-R1 DeepSeek-R1-Distill

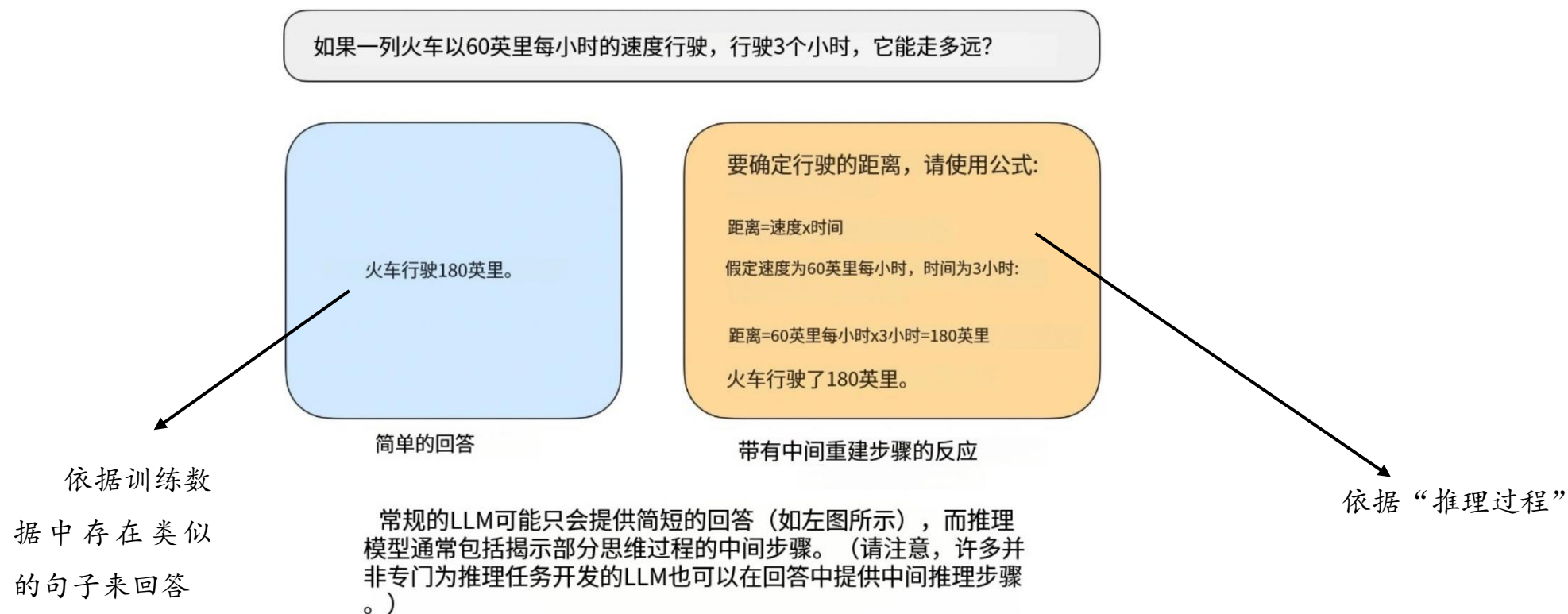
三、Deepseek
各个模型之间的
关系



第1-3阶段是开发 LLMs 的常见步骤。第4阶段专门针对特定用例开发 LLM。
。 针对逻辑推演任务的推理模型

四、推理模型（Reasoning Model）的含义

“推理”定义为回答需要复杂、多步骤生成并涉及中间步骤的问题的过程。例如，像“法国的首都是什么？”这样的问题不涉及推理。相比之下，问题“如果一列火车以每小时60英里的速度行驶并持续3小时，它会行驶多远？”则需要一些简单的推理。例如，它需要在得出答案之前，先识别距离、速度和时间之间的关系。



四、推理模型 (Reasoning Model) 的含义

1) Intermediate reasoning chain that is not explicitly shown to the user
非显式出现的推理过程

2) Intermediate reasoning steps that are shown as part of the answer
部分显式出现的推理过程

The main answer

ChatGPT o1

What is the last digit of 7^{999}

Reasoned about last digit of powers for a couple of seconds

To find the last digit of 7^{999} , note that the last digits of powers of 7 repeat in a cycle of length 4:

$7^1 = 7$	(last digit 7)
$7^2 = 49$	(last digit 9)
$7^3 = 343$	(last digit 3)
$7^4 = 2401$	(last digit 1)

Then it repeats: 7, 9, 3, 1, 7, 9, 3, 1, ...

To determine which of these four digits appears for 7^{999} , we look at $999 \bmod 4$:

$$999 \div 4 = 249 \text{ remainder } 3.$$

Since the remainder is 3, we take the third digit in the cycle (7, 9, 3, 1), which is 3.

Hence, the last digit of 7^{999} is **3**.

"Reasoning" is used at two different levels: 1) processing the input and generating via multiple intermediate steps and 2) providing some sort of reasoning as part of the response to the user.

五、何时使用推理模型？

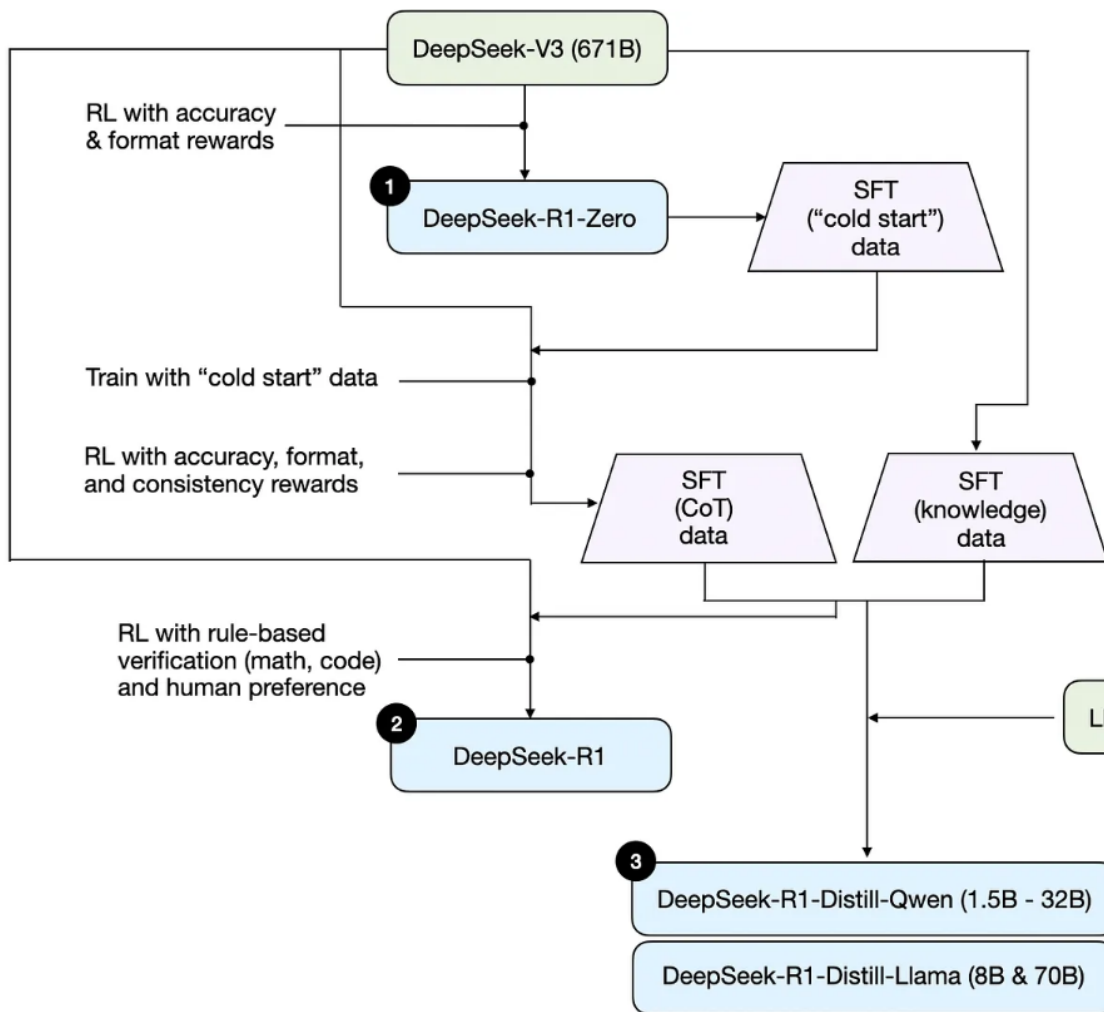
Good at	Bad at
<div><div>+ Deductive or inductive reasoning (e.g., riddles, math proofs)</div><div>+ Chain-of-thought reasoning (breaking down multi-step problems)</div><div>+ Complex decision-making tasks</div><div>+ Better generalization to novel problems</div></div> <div>擅长：</div> <div><div>1、演绎或者归纳。</div><div>2、思维链推理（分解多步骤问题）。</div><div>3、复杂决策任务。</div><div>4、概括新问题。</div></div>	<div><div>- Fast and cheap responses (more inference time)</div><div>- Knowledge-based tasks (hallucination)</div><div>- Simple tasks (“overthinking”)</div></div> <div>不擅长：</div> <div><div>1、短时、快速的回复。（思考过程带来时间开销）</div><div>2、基于知识的回答。（模型小，含的知识可能少。）</div><div>3、简单任务。（容易过度思考）</div></div>

提示词示例

例如“根据数据库，我希望查询工资表中信息部门的员工，谁的工资在5K元到8K元之间，请你返回sql查询语句”

例如“请告诉我从表中筛选特定部门的数据记录的sql语句”

六、Deepseek模型家族简介



(1) **DeepSeek-R1-Zero**: 该模型基于2024年12月发布的671B预训练DeepSeek-V3基础模型。研究团队使用强化学习（RL）进行训练，并引入了两种类型的奖励。这种方法被称为“冷启动”训练，因为它没有包括监督微调（SFT）步骤，而后者通常是人类反馈强化学习（RLHF）的一部分。

(2) **DeepSeek-R1**: 这是DeepSeek的旗舰推理模型，基于DeepSeek-R1-Zero模型构建。团队通过额外的SFT阶段和进一步的RL训练进一步优化了“冷启动”R1-Zero模型。

(3) **DeepSeek-R1-Distill***: 使用前述步骤中生成的SFT数据，DeepSeek团队对Qwen和Llama模型进行了微调，以增强它们的推理能力。虽然这不是传统意义上的蒸馏过程，但该过程涉及将较小的模型（Llama 8B和70B，以及Qwen 1.5B-30B）在更大的DeepSeek-R1 671B模型输出上进行训练。

6.1 基于人类反馈的强化学习(RLHF)

R1的意思是强化学习。先了解经典的RLHF。Zero的意思是免除SFT

OpenAI:基于人类反馈的强化学习：Reinforcement Learning from Human Feedback(RLHF)

Step 1

Collect demonstration data and train a supervised policy.

1.1 人工创建提示词文本库。

A prompt is sample from our prompt dataset.

1.2 采样一个提示词

A labeler demonstrates the desired output behavior.

1.3 人工写回答

This data is used to fine-tune GPT-3.5 with supervised learning.

1.4 重复大量1.2-1.3工作，

构建人工问答数据集。得

人工微调GPT-3模型。



Step 2

Collect comparison data and train a reward model.

2.1 准备四个能对话的模型:

A: 人类。B: 微调的GPT

C: 原始GPT。D: 其他

A prompt and several model outputs are sampled.

2.2 输入一个提示词，得到

ABCD的四个回答

A labeler ranks the outputs from best to worst.

2.3 人工排序ABCD的好坏。

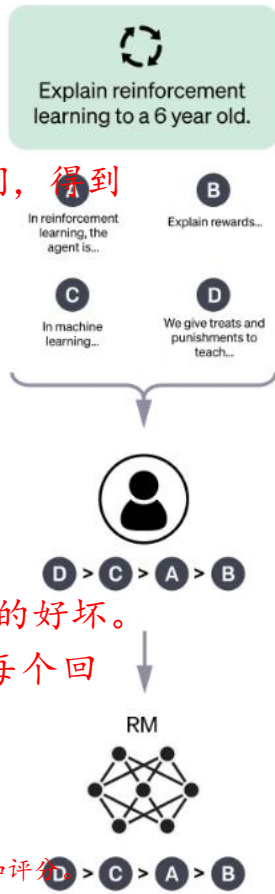
使用ELO机制，为每个回

答评分。

This data is used to train our reward model.

2.4 形成多个【提示词、回答】和评分

用于训练奖励模型。



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

3 训练策略(policy)模型 (即1中的模型)。Proximal Policy Optimization算法。

A new prompt is sampled from the dataset.

3.1 采样一个提示词

The PPO model is initialized from the supervised policy.

3.2 PPO模型初始化 (也是一个文本对话模型)

The policy generates an output.

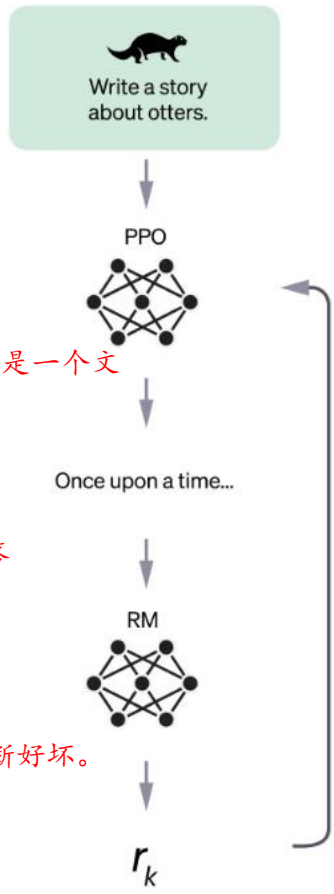
3.3 PPO模型产生 文本回答

The reward model calculates a reward for the output.

3.4 回答送给奖励模型判断好坏。

The reward is used to update the policy using PPO.

3.4 根据奖励，更新PPO模型 (LLM)



6.1 基于人类反馈的强化学习(RLHF)

RLHF中的PPO算法

1、获得人工调教过的LLM模型G（这一步灌入价值观正面的数据，以及符合人类习惯的对话模式），再准备一个未被调教过的模型B

2、获得奖励模型Rθ，能标注任意“问答”文本对的好坏。

3 训练过程

被训练模型是G。

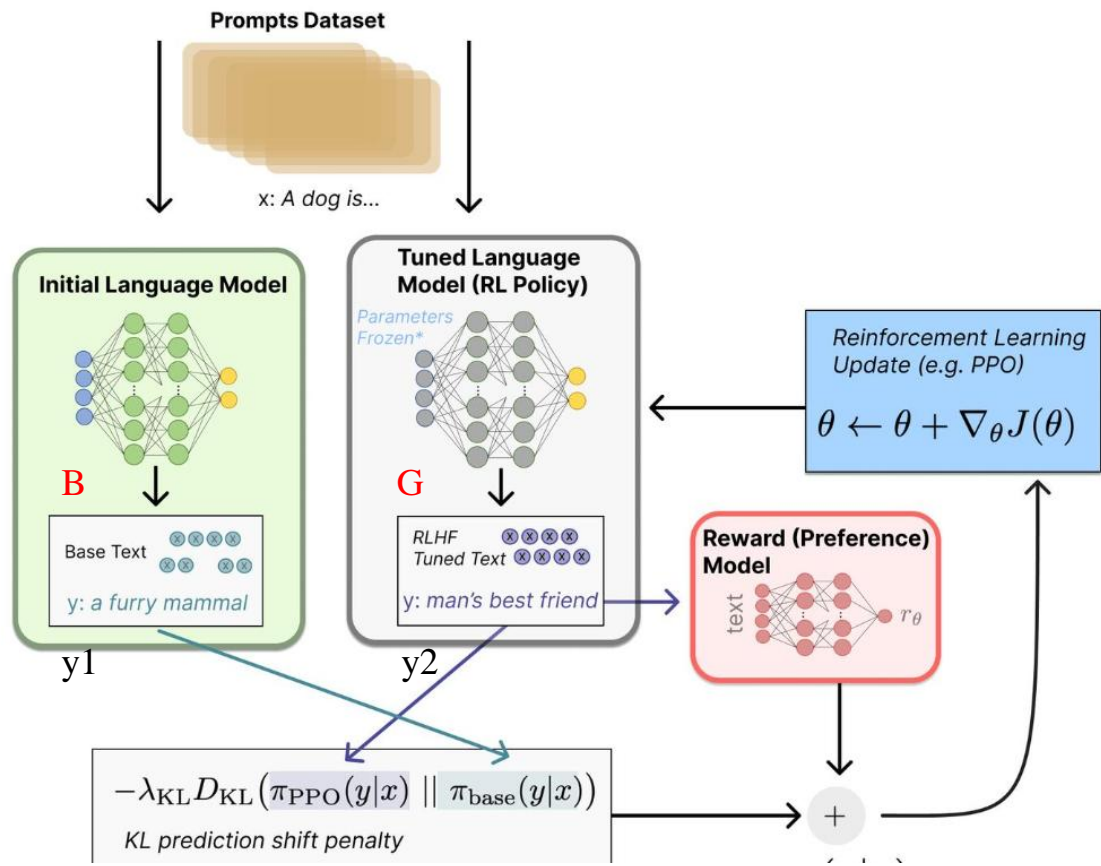
3.1 将某提示词输入B和G，得到回答y1,y2。y1相当于“未知好坏的回答”，y2相当于“好回答”。

3.2 y2进入奖励模型得到一个分数r。

3.3 以r为目标更新G的参数。让G的输出越来越“好”。

但是，为了不让G生成奇怪的文本以获得“高奖励”，加上惩罚项。
惩罚项的作用是使得G生成的文本仍然与起始LLM的输出文本接近。

总得来说，输入问题后，为了高分奖励，
，模型越来越倾向于生成“具有人类对话风格”的文本回答。
实际上模型不是硬性学习人类反馈，而是学习“哪种人类反馈”更优秀。



6.2 DeepSeek-R1-zero

为何使用RLHF？而不是拿step1中的人工问答对直接SFT 大模型？

- 1.省数据，有RL之后可以降低对SFT阶段的数据量要求，RLHF阶段每次新增的数据不多，标注压力不大；
- 2.省时间，训练RM和策略模型可以反复迭代，少量多次；
- 3.能引导模型产生正面价值观对话而非有害文本。
- 4.强化学习比较契合对话的过程；

R1-zero模型的RL有什么特点？

GPT-3.5 是创建一个“符合人类对话偏好”的文本对话模型。

R1-zero 则是为了从基础模型（deep-V3）开始获得一个推理模型。

6.2.1 DeepSeek-R1-zero: GRPO决策

DeepSeek-R1-zero中的决策模型

typically the same size as the policy model, and estimates the baseline from group scores instead. Specifically, for each question q , GRPO samples a group of outputs $\{o_1, o_2, \dots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_{θ} by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)] \\ \frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right), \quad (1)$$

$$\mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1, \quad (2)$$

where ϵ and β are hyper-parameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \dots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (3)$$

DeepSeek-R1-zero中的特殊奖励模型

2.2.2. Reward Modeling

The reward is the source of the training signal, which decides the optimization direction of RL. To train DeepSeek-R1-Zero, we adopt a rule-based reward system that mainly consists of two types of rewards:

- **Accuracy rewards:** The accuracy reward model evaluates whether the response is correct. For example, in the case of math problems with deterministic results, the model is required to provide the final answer in a specified format (e.g., within a box), enabling reliable rule-based verification of correctness. Similarly, for LeetCode problems, a compiler can be used to generate feedback based on predefined test cases.
- **Format rewards:** In addition to the accuracy reward model, we employ a format reward model that enforces the model to put its thinking process between '`<think>`' and '`</think>`' tags.

We do not apply the outcome or process neural reward model in developing DeepSeek-R1-Zero, because we find that the neural reward model may suffer from reward hacking in the large-scale reinforcement learning process, and retraining the reward model needs additional training resources and it complicates the whole training pipeline.

对于一个输入问题 q ，决策模型每次产生 G 个输出。随后用式1-2获得惩罚项。再把 G 个输出给奖励模型获得 G 个分数 r_1 到 r_g 。

GRPO 是怎么干的?

1. 输入问题，采样输出

首先，给模型一个输入问题，然后让它生成一堆输出。每个输出都包含一个推理过程和答案。

2. 用规则打分

GRPO 会根据一些预定义的规则，给每个输出打分，也就是计算奖励。这些规则主要看两点：

- **准确度：**比如数学题，答案对不对可以直接检查；编程题的话，可以用测试用例来验证。
- **格式：**模型得按照规定的格式输出。比如论文里要求推理过程放在 `<think>` 标签里，答案放在 `<answer>` 标签里，格式奖励就是确保模型不乱来。

注意到此时奖励模型不是一个神经网络模型。

6.2.2 DeepSeek-R1-zero :自我进化

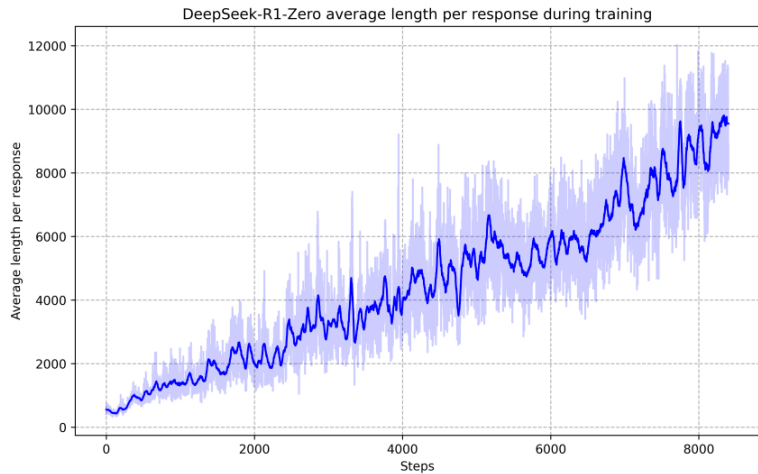


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

特点1：越训练，模型输出的回答的长度越长。表示模型逐渐加深思考的长度。

DeepSeek-R1-zero的创新之处：

- 1、无需SFT（zero含义）
- 2、基于规则的强化学习
- 3、有了很强的推理能力。

缺陷：

- 1、对话可读性差，语言混乱。这方面不如RLHF的模型。缺乏良好的体验。

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

Table 3 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

特点2：模型获得思考能力后，出现“啊哈”时刻。

“顿悟”时刻意味着模型在遇到逻辑不通时可自行修正。

冷启动（第 1 阶段）

- 1、利用 DeepSeek-R1-Zero 模型，生成一批 SFT 数据，且这些结果经过验证，确保高质量和可读性。
- 2、利用 1 的数据，监督微调从预训练模型 DeepSeek-V3-Base。
- 3、步骤 2 有助于 DeepSeek-R1 缓解在初始模型中观察到的可读性问题。

推理强化学习（第 2 阶段）

- 1、采用大规模强化学习来增强模型的推理能力。
- 2、重点提升编码、数学、科学和逻辑推理等任务的能力。
- 3、这些任务的明确解决方案可用于定义强化学习过程中的奖励规则。

拒绝采样和监督微调（第 3 阶段）

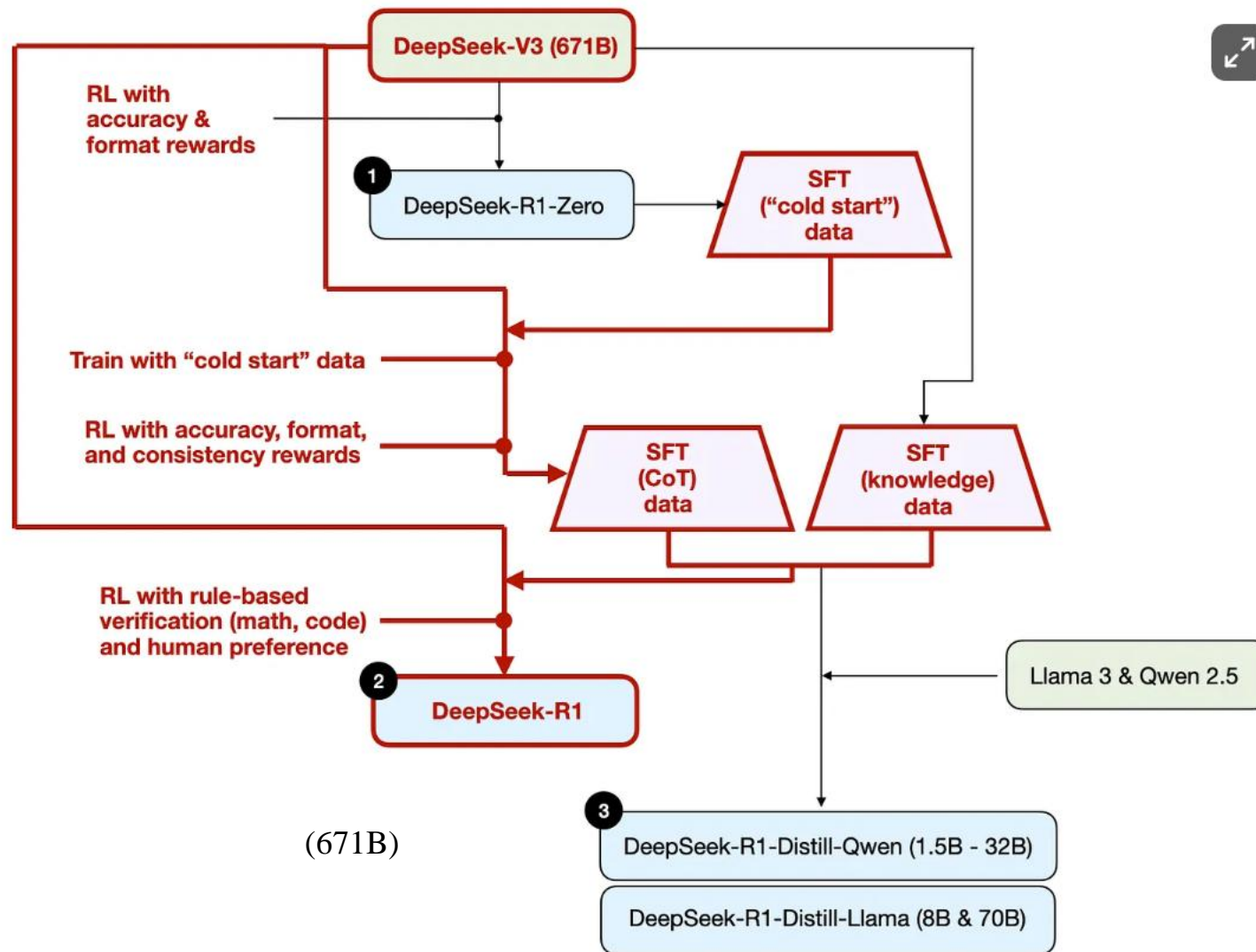
- 1、该阶段使用第 2 阶段的模型检查点来生成大量样本。
- 2、采用拒绝采样，仅保留正确且可读的样本。
- 3、训练奖励模型 DeepSeek-V3，用于决定应保留哪些样本。
- 4、此阶段还包括 DeepSeek-V3 的部分训练数据。
- 5、监督微调用于训练模型，使其在推理导向问题之外，增强在更多领域的能力。

多样化强化学习阶段（第 4 阶段）

该阶段包含多样化的任务：

- 1、对于允许的任务（例如数学），使用基于规则的奖励。
- 2、对于其他任务，LLM 提供反馈，使模型与人类偏好保持一致。
- 3、使用第 3 阶段的数据集提炼了各种较小的开源模型，提供了具有高推理能力的较小替代方案。

6.3 DeepSeek-R1



6.4 DeepSeek-R1-zero与R1的实验结果

The final model, DeepSeek-R1 has a noticeable performance boost over DeepSeek-R1-Zero thanks to the additional SFT and RL stages, as shown in the table below.

Model	Math benchmarks			Bio, physics & chemistry	Code benchmarks	
	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
RL only → DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444
SFT + RL → DeepSeek-R1	79.8		97.3	71.5	65.9	2029

Higher is better

数学题、化学物理题、代码题表现均超越 openAI-o1-mini

6.5 DeepSeek-V3

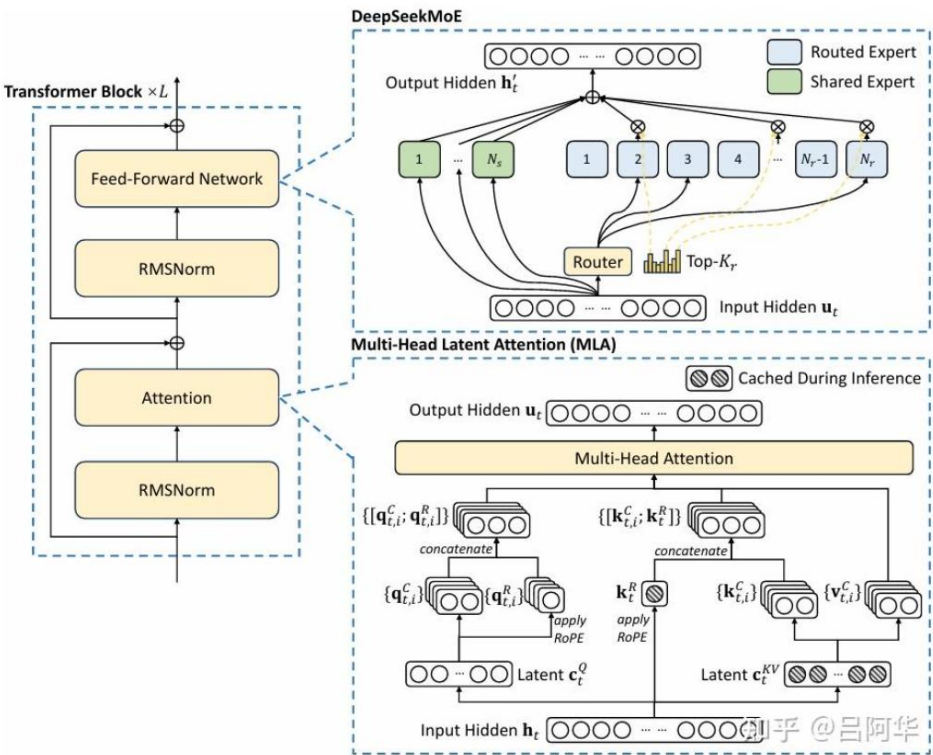


图2: DeepSeek-V3 基本架构示意图。基于 DeepSeek-V2, 团队采用了多头潜在注意力 (MLA) 和 DeepSeekMoE 架构, 以实现高效推理和经济的训练。

V3训练中的技术:

一、模型架构技术

- (1) MLA 多头潜在注意力, 显著减少KV cache。
- (2) 带共享专家的MoE, 可自由扩展模型容量。
- (3) 无额外损耗的负载均衡, 这用来帮助各个专家被合理分配。

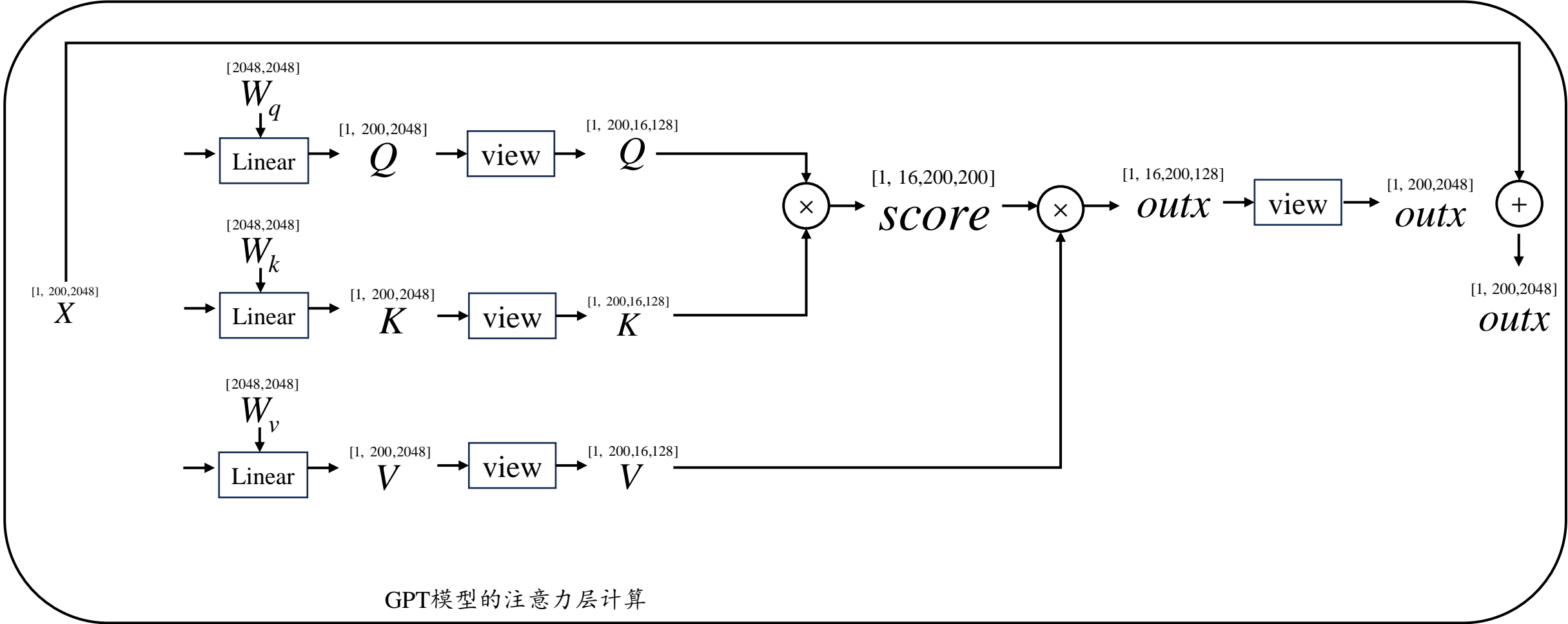
二、工程技术

- (1) DualPipe 流水线并行。前向传播、反向传播产生的变量尽量占满通信道路。
- (2) 选择性FP8精度训练 (1符号位, 4位指数, 3位尾数)。对于输出精度影响大的模块, 用bf16; 而影响小的, 用fp8。
- (3) 跨节点通信优化。专为MoE设计。由于不同专家可能在不同节点、显卡, token进入它们的时候, 互相通信需要IO时间和带宽。进行了优化。

三、预训练

- (1) 14.8万亿token, 数学、编程数据占比高。
- (2) 最小化数据重复和冗余。
- (3) 多个文档拼接成一个。避免截断导致缺失上下文。

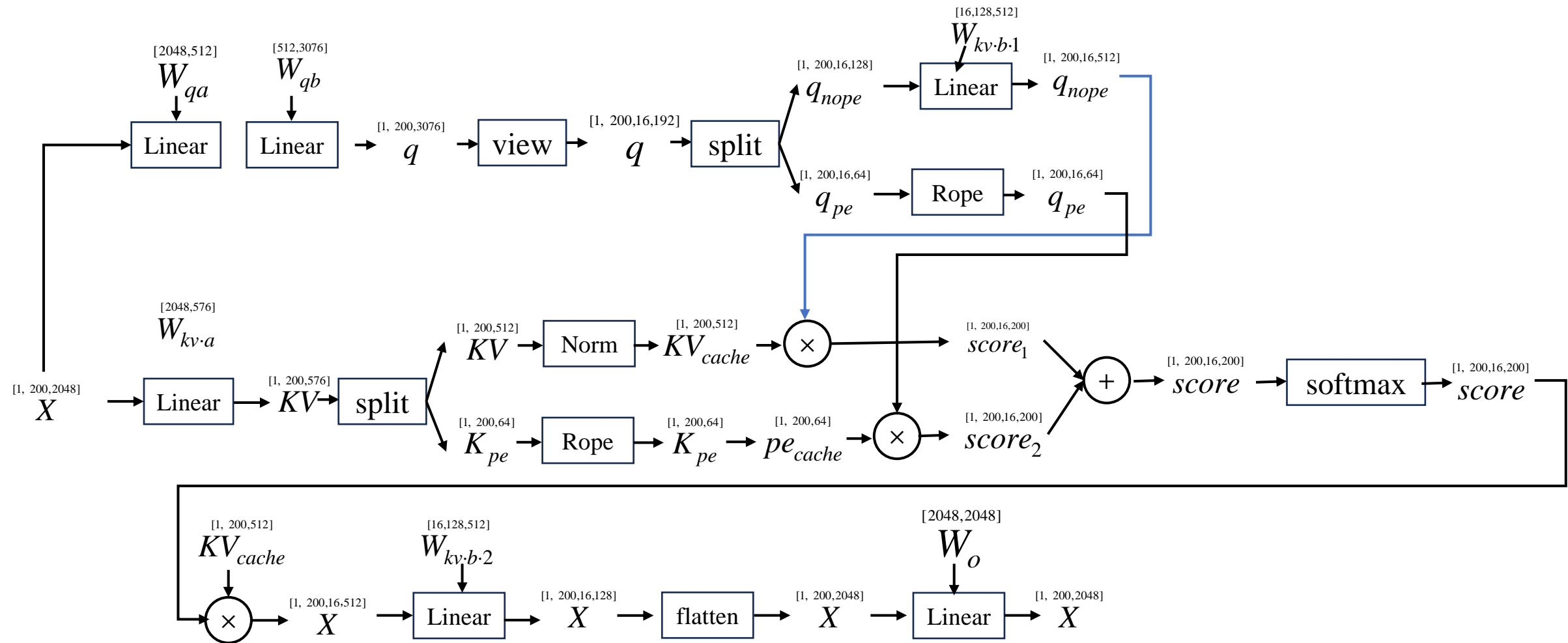
6.5.1 GPT注意力



计算复杂度存在 2048的平方

6.5.2 DeepSeek-V3的注意力模块Multi-head Latent Attention

$A[M,N] \otimes B[N,K]$,复杂度= $O(MNK)$ 推荐看: <https://spaces.ac.cn/archives/10091>

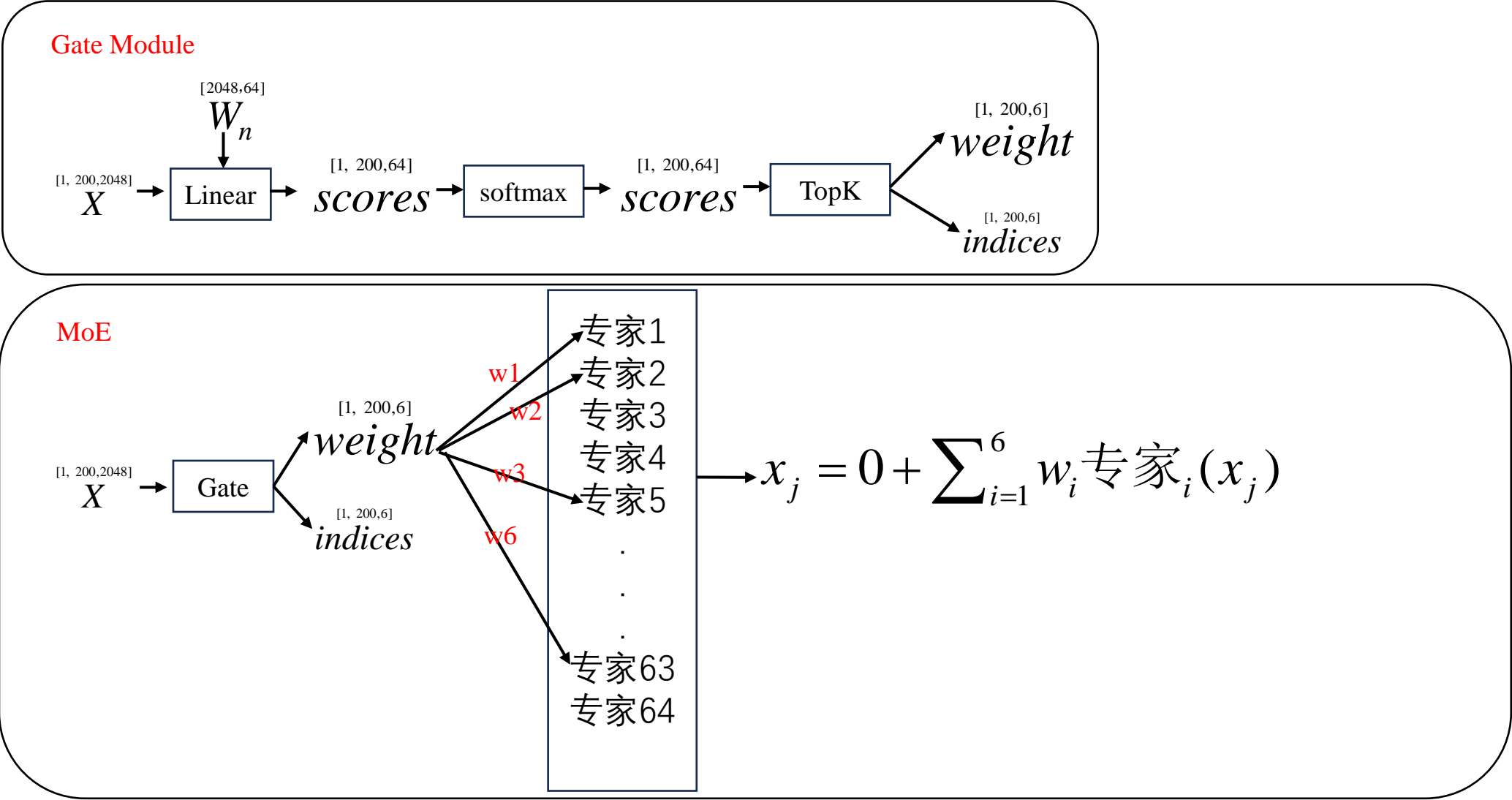


与GPT模型相比的特点:

- 1、利用“低秩矩阵”，显著减少kv cache
- 2、推理next token的时候，kv cache 不受 head num 影响。（这与之前的LLM设计不同）
- 3、总参数量和显存占用实际比相同规模GPT要高，但是由于KV cache小，推理长文本更具优势。

```
>>> (200*2048*2048) / (200*2048*512+200*512*2048)
2.0
>>>
```

6.5.3 DeepSeek-V3的混合专家模块MoE



X中每个token的向量由对应的专家“坐诊”进行线性矩阵乘法计算。

七、构建“推理时拓展”（Inferencing Time Scaling）能力

7.1 什么是推理时拓展？

1、什么是推理时拓展？

不增加训练工作，但在模型推理时，提升或者强迫LLM使用更多计算资源（更多token），来提升输出的质量。使得推理的结果更长、更好。

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model input

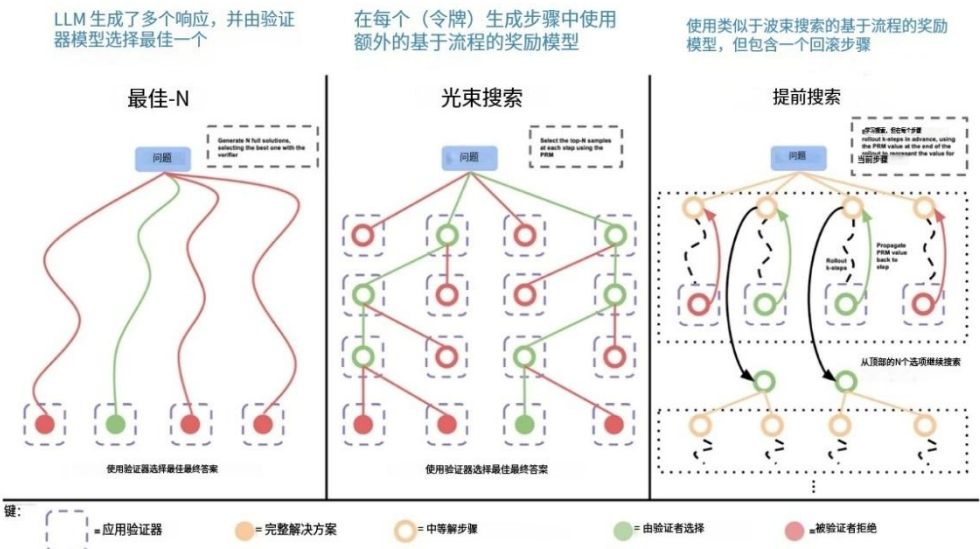
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅



不同的基于搜索的方法依赖于基于过程-奖励的模型来选择最佳答案。来自 LLM Test-Time Compute 论文的注释图，<https://arxiv.org/abs/2408.03314>

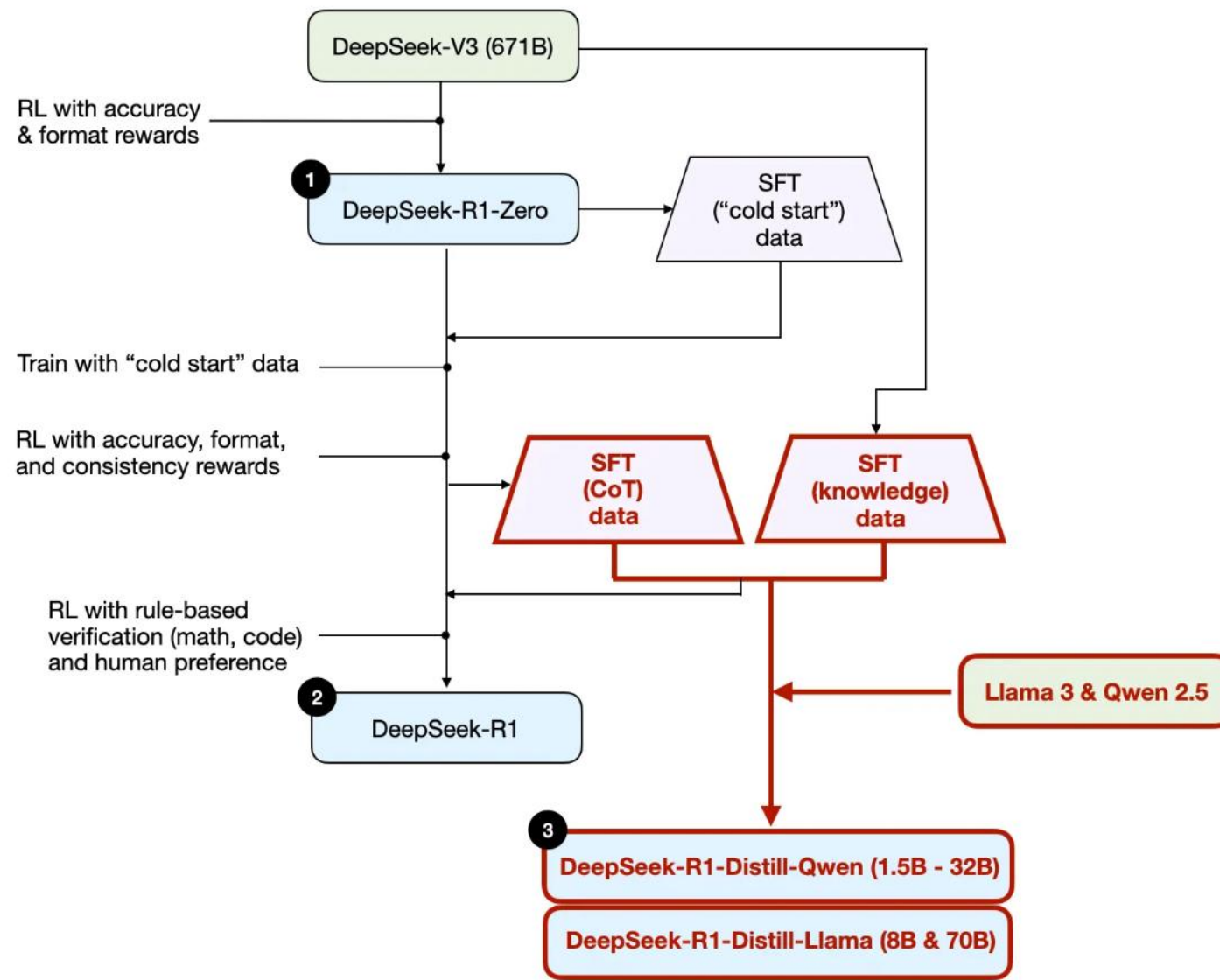
第一种方法手段1：推理时CoT（思维链）

第一种方法手段2：多回答投票

第二种方法：纯粹强化学习

第三种方法：SFT + 强化学习

7.2 Deepseek的推理时扩展的方法：蒸馏



The development process of DeepSeek-R1-Distill models.

7.3 Deepseek 蒸馏模型的效果

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
DeepSeek-R1-Zero	71.0		95.9	73.3	50.0	1444
DeepSeek-R1	79.8		97.3	71.5	65.9	2029

Benchmark comparison of distilled versus non-distilled models. Annotated figure from the DeepSeek-R1 technical report (<https://arxiv.org/abs/2501.12948>).

蒸馏32B模型的性能基本已经达到o1 mini水平，也强于R1-zero（671B），那么，如果纯粹的RL能否让小模型也智力提升呢？

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
pure RL → 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
pure RL → DeepSeek-R1-Zero	71.0		95.9	73.3	50.0
DeepSeek-R1	79.8		97.3	71.5	65.9

答案似乎是不能，蒸馏比RL更有效。

- 1、可能是因为模型小。
- 2、蒸馏适合小模型，而RL比较适合大模型（671B）

7.4 Deepseek 实际使用测试

DeepSeek-R1-Distill-Qwen-14B

把下面的文本内容翻译为中文，直接输出翻译结果，从第一个字开始翻译，不要有遗漏。不要输出任何不相关内容。
文本: {}。

We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.—That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed,—That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness. Prudence, indeed, will dictate that Governments long established should not be changed for light and transient causes; and accordingly all experience hath shewn, that mankind are more disposed to suffer, while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, pursuing invariably the same Object evinces a design to reduce them under absolute Despotism, it is their right, it is their duty, to throw off such Government, and to provide new Guards for their future security.

对于翻译、摘要这样的任务而言，think是非必要的。

第一次回答：LLM Response:
<think>
好，我现在要处理用户的这个翻译请求。首先，用户给了一段英文文本，看起来像是《独立宣言》的一部分，内容涉及人生而平等、不可剥夺的权利等。用户的要求很明确，直接翻译成中文，从第一个字开始，不能遗漏，而且不要输出任何不相关内容。

我需要先通读一遍这段英文，确保理解其含义和结构。这段话比较正式，有很多法律和政治术语，翻译时要准确传达原意。比如，“self-evident truths”翻译成“不言而喻的真理”，“unalienable Rights”是“不可转让的权利”，这些词汇在中文中有固定的译法，不能随意改动。

接下来，我会逐句翻译，确保每个部分都准确无误。要注意句子的连贯性和逻辑性，比如连接词和标点符号的使用。例如，“That to secure these rights, Governments are instituted among Men”可以翻译为“为了保障这些权利，人们设立政府”，这样既准确又通顺。

在翻译过程中，还要注意保持原文的语气和风格，比如正式、庄严的感觉。同时，要避免直译导致的生硬，必要时进行意译，使中文读起来更自然。例如，“pursuing invariably the same Object”可以翻译为“一再从事同样的暴行”，这样更符合中文的表达习惯。

完成翻译后，我会通读一遍，检查是否有遗漏或错误，确保每个部分都准确传达了原文的意思。同时，还要确保没有添加任何额外的内容，完全按照用户的要求，只输出翻译结果。

最后，我会将翻译好的中文内容返回给用户，确保格式正确，没有多余的信息。这样用户就能得到准确、完整的翻译结果，满足他们的需求。
</think>

我们认为下面这些真理是不言而喻的：人生而平等，造物主赋予他们若干不可转让的权利，其中包括生存权、自由权和追求幸福的权利。为了保障这些权利，人类才在他们中间建立政府，而政府的正当权力，是经人民同意而产生的。任何形式的政府，一旦破坏这些目的，人民就有权废除它，并建立新的政府，其原则和组织形式，应是他们认为最能保障他们的安全和幸福的。确实，审慎将指出，不应为轻而易举、一时的动机而改变长期存在的政府。因此，所有经验表明，人类更愿意在邪恶尚可忍受时继续忍受，而不愿通过废除他们习惯于服从的形式来纠正自己的权利。然而，当一个政府长期从事压迫和暴行，显露出企图使他们处于专制统治下的明确用心时，人民有权、有义务抛弃这种政府，并为他们未来的安全建立新的保障。

第二次回答：LLM Response:
<think>

</think>

我们认为下面这些真理是不言而喻的：人生而平等，造物者赋予他们若干不可转让的权利，其中包括生命权、自由权和追求幸福之权。为了保障这些权利，人类在他们中间建立政府，政府的正当权力，是经人民同意而产生的。任何形式的政府，一旦破坏这些目的，人民就有权废除它，并建立一个新政府，其组织原则和权力形式，应使他们认为最能保障他们的安全和幸福。确实，审慎将指出，不应为轻率和一时的原因而改变长期存在的政府；而经验表明，人类更愿意在邪恶尚可忍受时逆来顺受，而不愿废除他们习惯于服从的政府形式。但是，当一种政府以长期的压迫和篡夺，一再显示出企图将人民置于专制统治下的时候，人民有权，更有义务，摆脱这种政府，并为他们未来的安全建立新的保障。

7.4 Deepseek 实际使用测试

提示词： "有个人写了10篇论文， 投给会议ACL， ACL的稿件接受率是20%， 请问他恰好有一篇论文被接受的概率是？ "

<think>
嗯，我现在遇到了一个概率问题，想仔细思考一下怎么解决。题目是说，有个人写了10篇论文，投给了ACL会议，ACL的稿件接受率是20%。那么，他恰好有一篇论文被接受的概率是多少呢？

首先，我需要明确这是一个概率问题，应该用概率论中的相关知识来解决。我记得，这种类型的问题通常涉及到二项分布，因为每次投递论文的结果都是独立的，而且只有两种可能的结果：接受或者不接受。

好的，那我先回忆一下二项分布的公式。二项分布的概率质量函数是：

$$P(k) = C(n, k) \cdot p^k \cdot (1-p)^{(n-k)}$$

其中，n是试验的总次数，k是成功次数，p是每次试验成功的概率，C(n, k)是组合数，也就是从n次试验中选出k次成功的组合数。

在这个问题中，n就是10篇论文，k是恰好被接受的1篇，p是20%，也就是0.2。那我需要计算的就是P(1)。

首先，计算组合数C(10, 1)。组合数C(n, k)的计算公式是 $n! / (k!(n - k)!)$ 。所以，C(10, 1)就是 $10! / (1! \cdot 9!)$ 。计算一下，10!是10×9×8×...×1，而9!是9×8×...×1，所以10! / 9! = 10，然后除以1!也就是1，所以C(10, 1) = 10。

接下来，计算 p^k ，也就是 0.2^1 ，这很简单，就是0.2。

然后，计算 $(1-p)^{(n - k)}$ ，也就是 $0.8^{(10 - 1)} = 0.8^9$ 。这里我需要计算一下0.8的9次方是多少。可能需要一步步来算：

$0.8^1 = 0.8$

$0.8^2 = 0.64$

$0.8^3 = 0.512$

$0.8^4 = 0.4096$

$0.8^5 = 0.32768$

$0.8^6 = 0.262144$

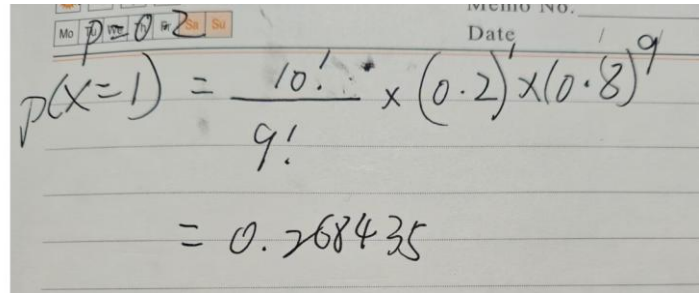
$0.8^7 = 0.2097152$

$0.8^8 = 0.16777216$

$0.8^9 = 0.134217728$

。。。。。。。（此处我省略了很多字）

</think>



就适合推理任务！

为了计算某人恰好有一篇论文被接受的概率，我们可以使用二项分布。二项分布的概率质量函数为：

$$P(k) = C(n, k) \cdot p^k \cdot (1-p)^{(n-k)}$$

其中， $n = 10$ （论文数量）， $k = 1$ （被接受的论文数量）， $p = 0.2$ （接受率）。

- 计算组合数 $C(10, 1)$ ：
$$C(10, 1) = \frac{10!}{1!(10-1)!} = 10$$
- 计算 p^k ：
$$0.2^1 = 0.2$$
- 计算 $(1-p)^{(n-k)}$ ：
$$0.8^9 \approx 0.134217728$$
- 代入二项分布公式：
$$P(1) = 10 \cdot 0.2 \cdot 0.134217728 = 2 \cdot 0.134217728 \approx 0.268435456$$

因此，恰好有一篇论文被接受的概率是：

$$\boxed{0.2684}$$

八、结语

一、构建推理模型的四种方法：

1、推理时扩展（CoT、多回答投票等）。2、纯强化学习。3、RL + SFT。4、蒸馏

二、DPSK家族的关系

(0) 训练deepseek-V3。

(1)DeepSeek-R1-Zero：在671B预训练DeepSeek-V3基础上直接RL得到。“zero”的意思是“无SFT”。R1的意思是“第一次强化学习”。

(2) DeepSeek-R1：基于DeepSeek-R1-Zero模型构建，通过额外的SFT阶段和进一步的RL训练进一步优化。名字里没有zero有R1，因此是有SFT和RL的。

(3) DeepSeek-R1-Distill*：使用前述步骤中生成的SFT数据，DeepSeek团队对Qwen和Llama等小型模型进行了微调。名字里有R1和Distill，意为在R1模型的输出数据的基础上，进行蒸馏。

三、DPSK模型的使用

适合推理任务。

感谢聆听！

欢迎批评指正

