



# 大模型入门之 监督微调与强化学习

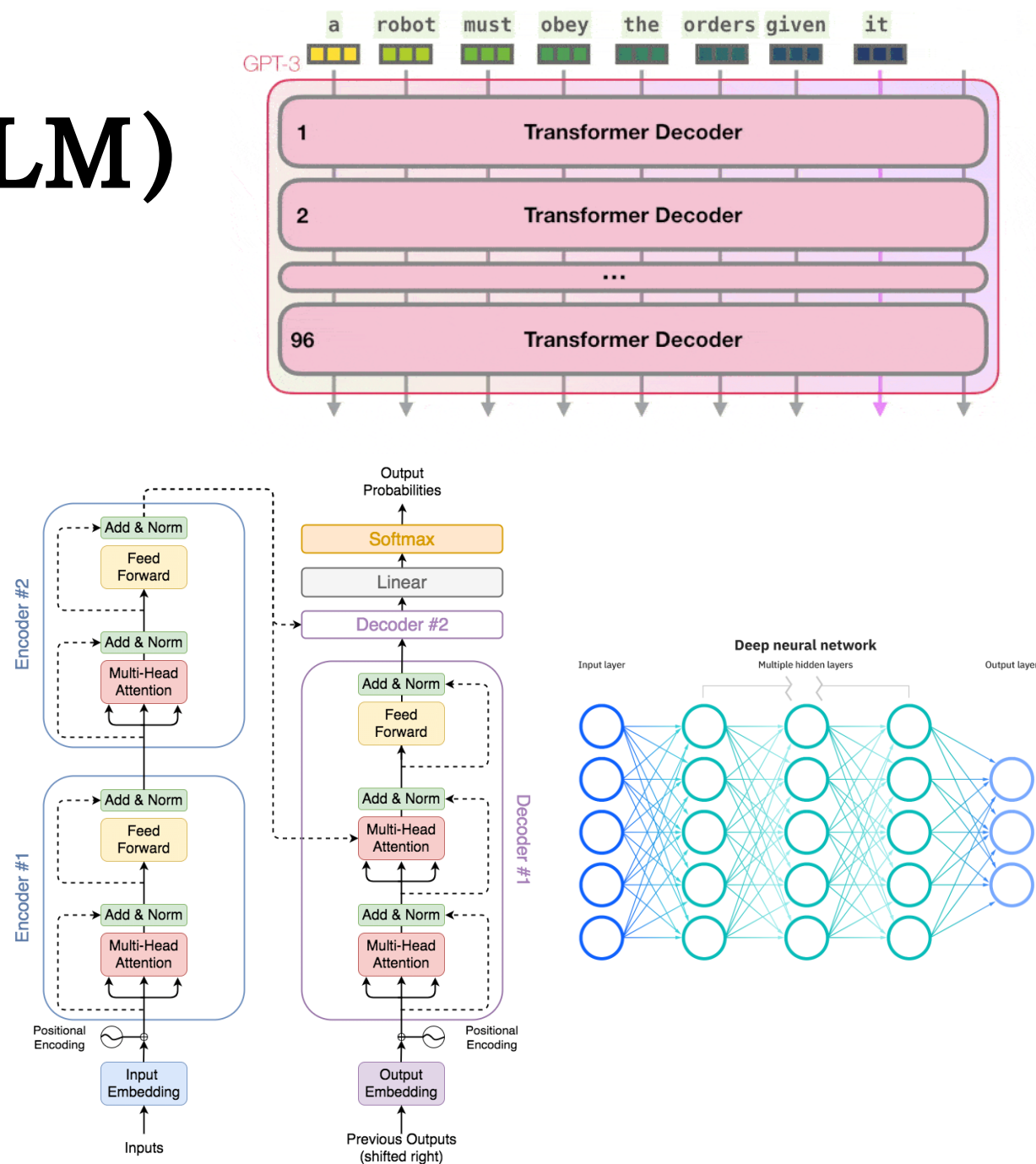
---

张以宁

北京大学 计算机学院

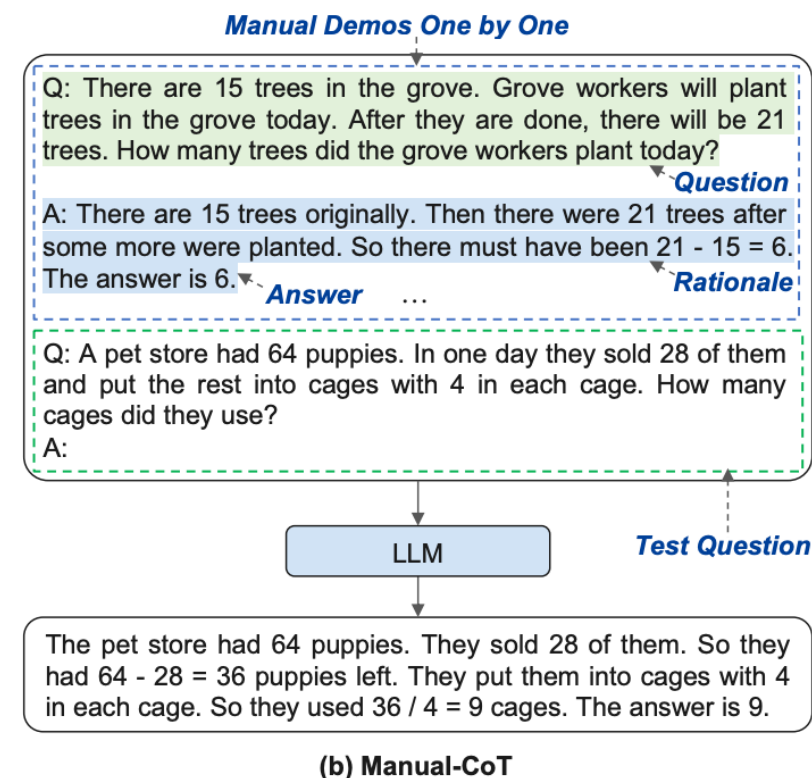
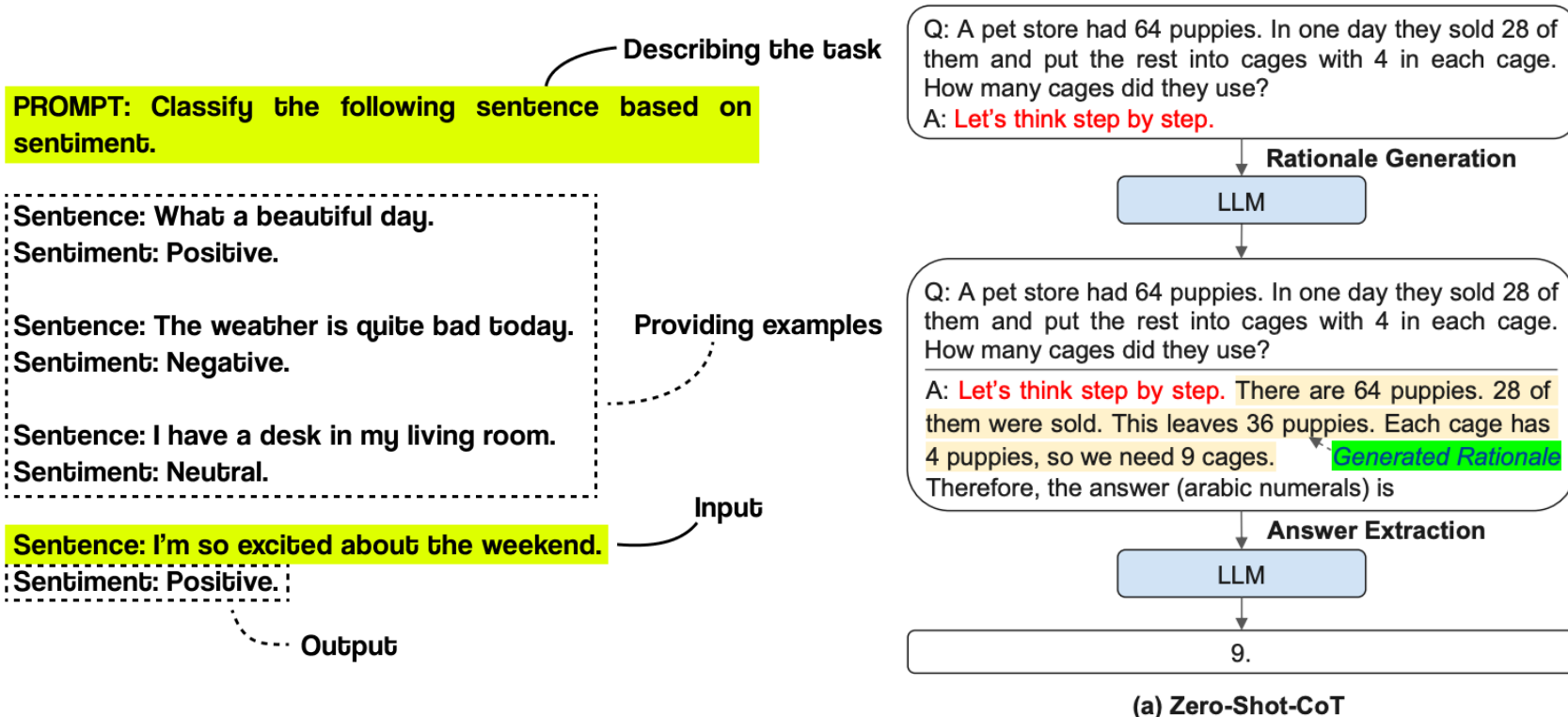
# 回顾：大语言模型（LLM）

- 给定一段输入序列，能够不断预测下一个词（Token）的概率分布的生成机器
- 基于 Transformer 结构的神经网络模型实现
- 模型拥有巨大的参数量（千亿级），经海量互联网文本（十万亿级Token）的预训练获得语言能力



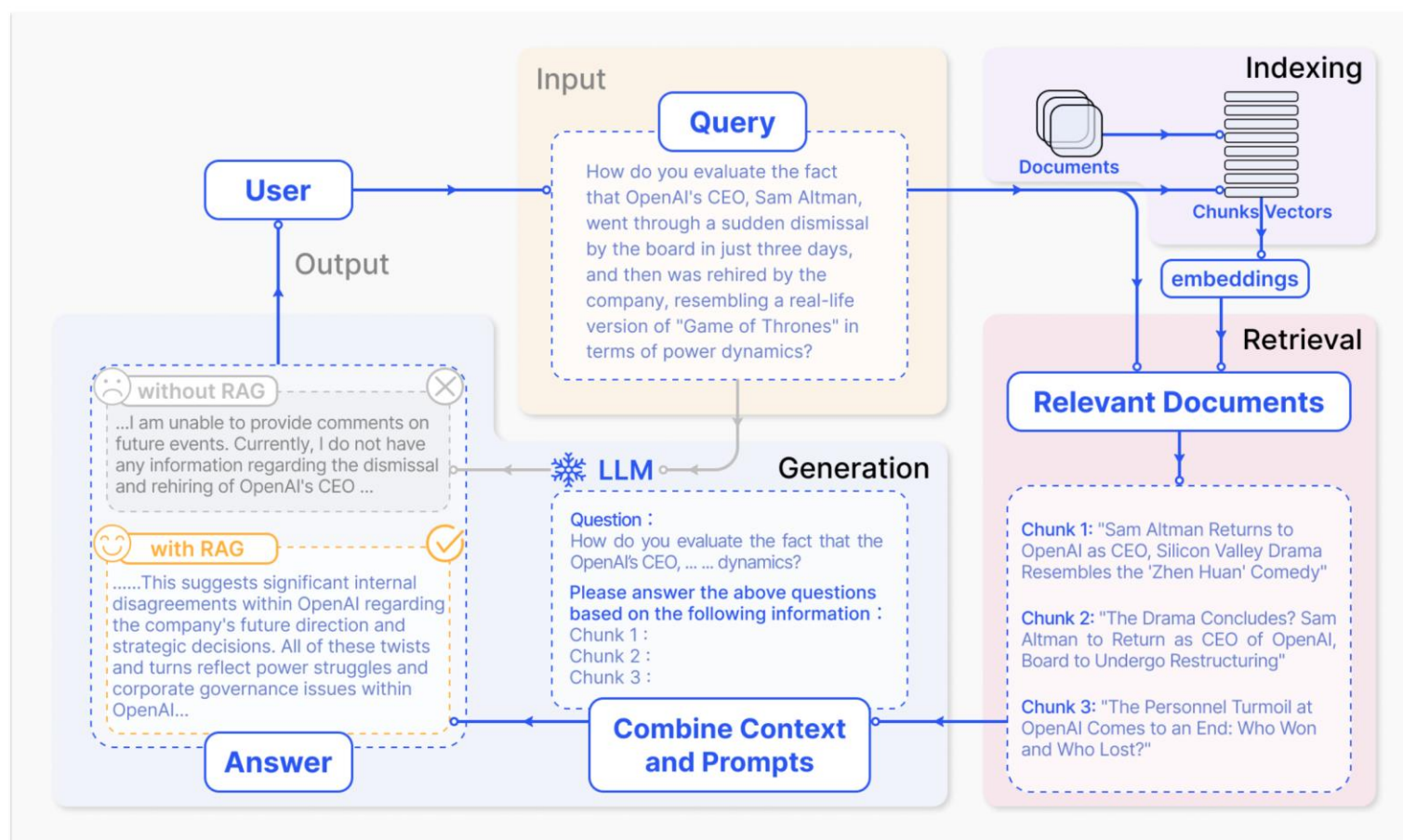
# 回顾：提示词工程（Prompt Engineering）

- 通过提示词（输入）设计的技巧引导模型产生质量更高的输出
  - Few-shot, 思维链（Chain-of-Thought, CoT）



# 回顾：检索增强生成（RAG）

- 将检索到与问题相关的材料加入提示词（输入），让模型“开卷考试”

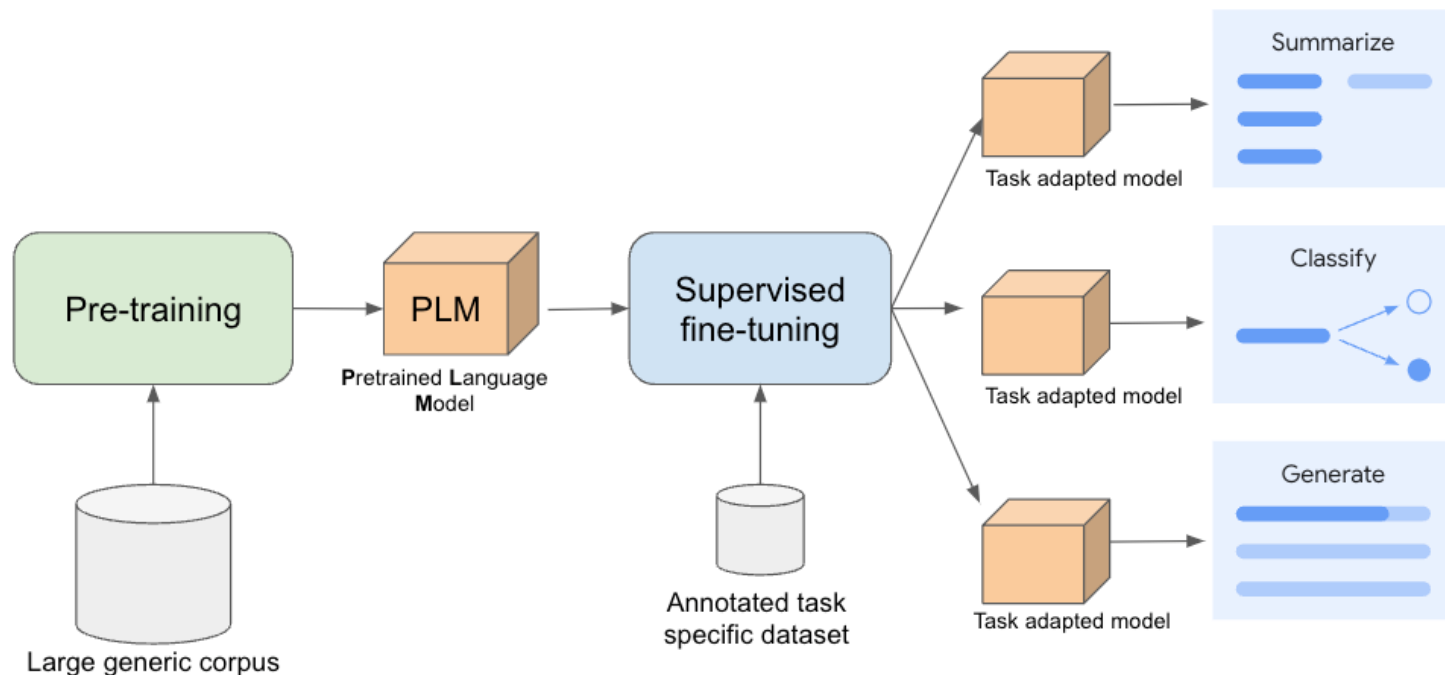


# 模型表现的提升

- 以上方法都是通过改变模型输入来改善模型生成内容的质量
  - 输入不变的情况下如何提升模型的表现？调整模型参数（微调 / 后训练）。
- 思考：期末考试临近，你是如何复习的？
  - 方法1：背书。给一堆习题和标准解答过程，照着标答背，让自己写的东西和标答尽可能相似 —— **监督微调 (Supervised Fine-Tuning, SFT)**
  - 方法2：做题。给一堆习题和答案，每道题自己尝试写多种解，对答案看哪个解正确就加强记忆 —— **强化学习 (Reinforcement Learning, RL)**
- 大模型的完整训练过程：预训练 → 监督微调 → 强化学习

# 监督微调 (SFT)

- 数据：任务的输入-输出对
  - 想提升什么任务上的表现，就用对应任务的数据（大约几百-几千条）
- 训练过程：
  - 对于每条数据，模型根据输入产生输出
  - 计算模型输出和标准输出的交叉熵（差异度）作为 Loss
  - 根据 Loss 计算对模型参数的梯度，更新参数



*Explain the moon landing to a 6 year old in a few sentences.*

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

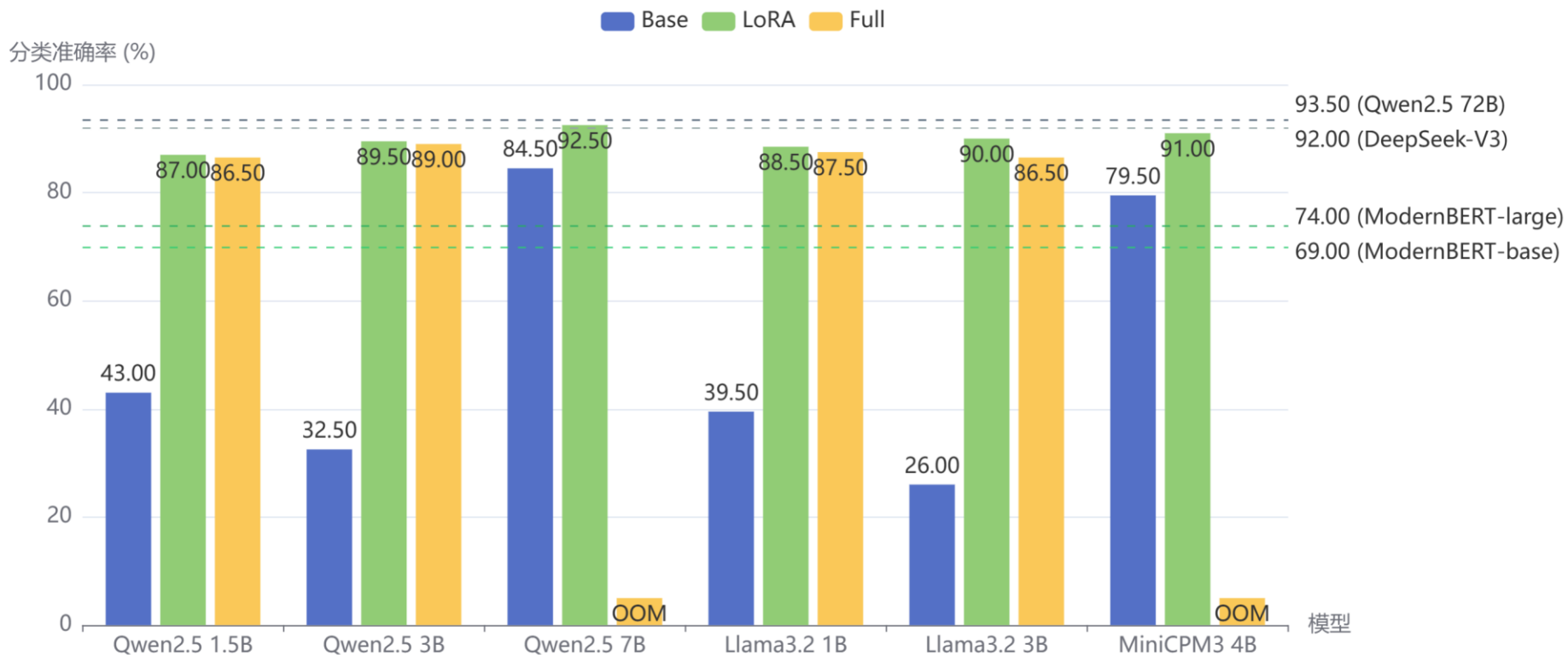
Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

SFT 魔法之一：将预训练大模型从“续写”变为“对话”模型

- SFT 魔法之二：迅速提升大模型在特定任务上的表现

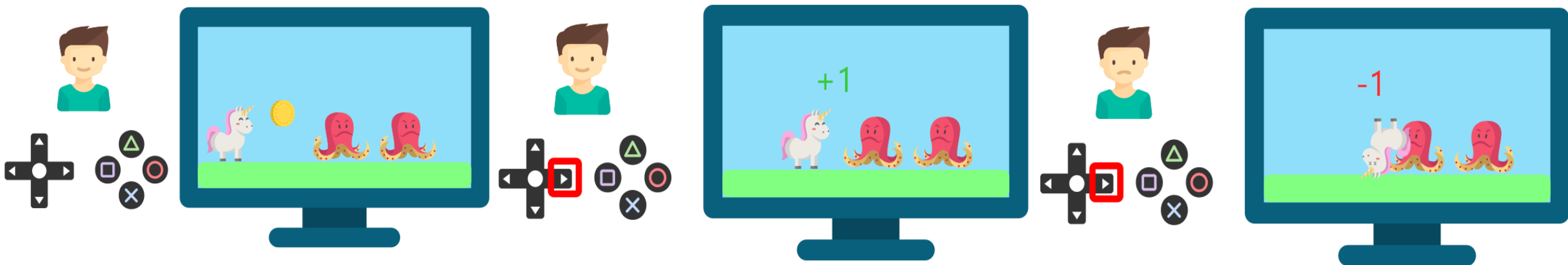


- 优点：提升效果好，训练成本低，适合输入输出要求明确的任务
- 缺点：每条训练数据需要有标准输出，如果使用人工标注成本较高
  - “模型蒸馏”：使用大模型生成的数据来训练小模型



# 强化学习

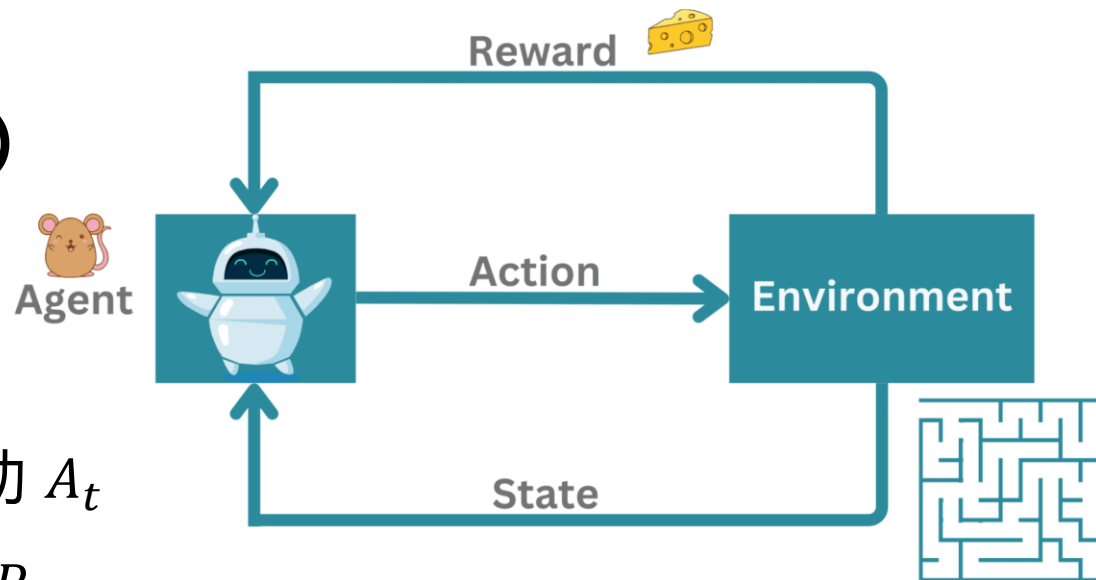
- 思考：你是如何学会玩一个游戏的？



- 在“观察环境，做出行动，获取反馈，观察新环境”的交互中改善自身的行动策略

- **强化学习 (Reinforcement Learning, RL)**

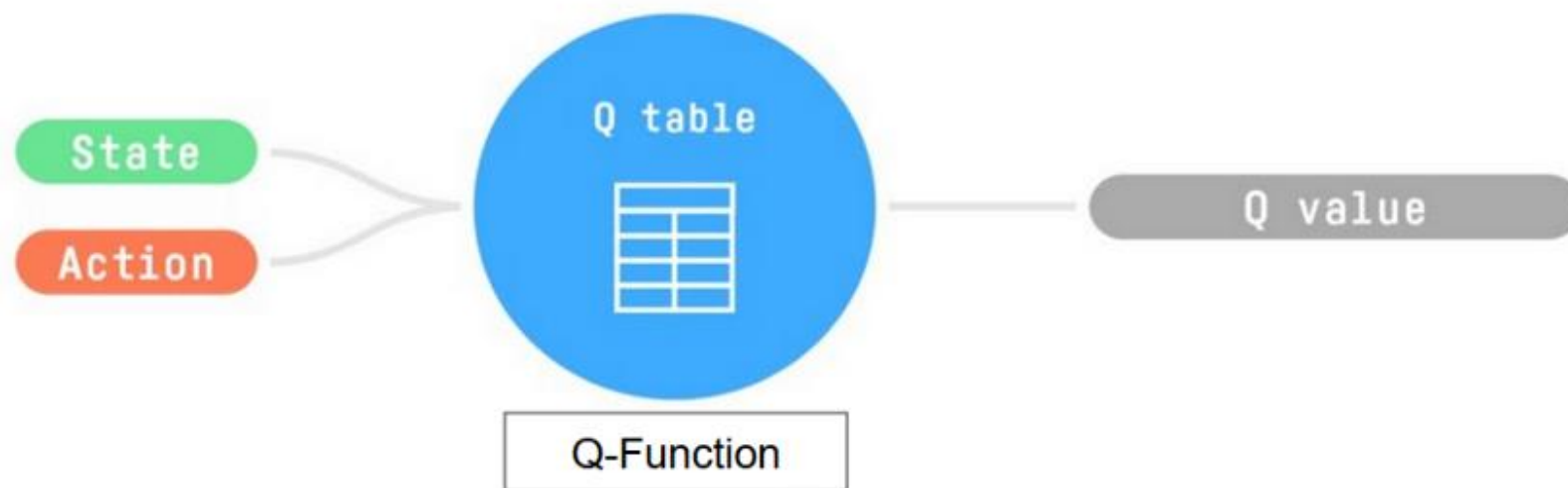
- 在某个时间步  $t$ :
  - 环境处于状态  $S_t$
  - 智能体 (Agent) 观察到状态, 做出一个行动  $A_t$
  - 环境根据状态和行动反馈给智能体一个奖励  $R_t$
  - 环境 (受智能体动作影响) 进入新状态  $S_{t+1}$
- 环境的状态转换和奖励的产生是客观属性 (“游戏规则”)
  - 我们的目标是优化智能体的 “策略”, 使其获得最大的累计奖励
    - 策略 (Policy) : 各个状态下采取各种行动的概率分布













$$\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1], \pi(s, a) = \Pr(A_t = a \mid S_t = s)$$

# 强化学习的实现

- 第一类方法：基于价值 (Value-Based)
  - 直觉：如果能知道在某个状态下采取各种行动分别带来的“价值”是多少，就可以贪心选取最大价值的动作
  - “价值”不等于当前步获得的奖励，还取决于**后续**的情况



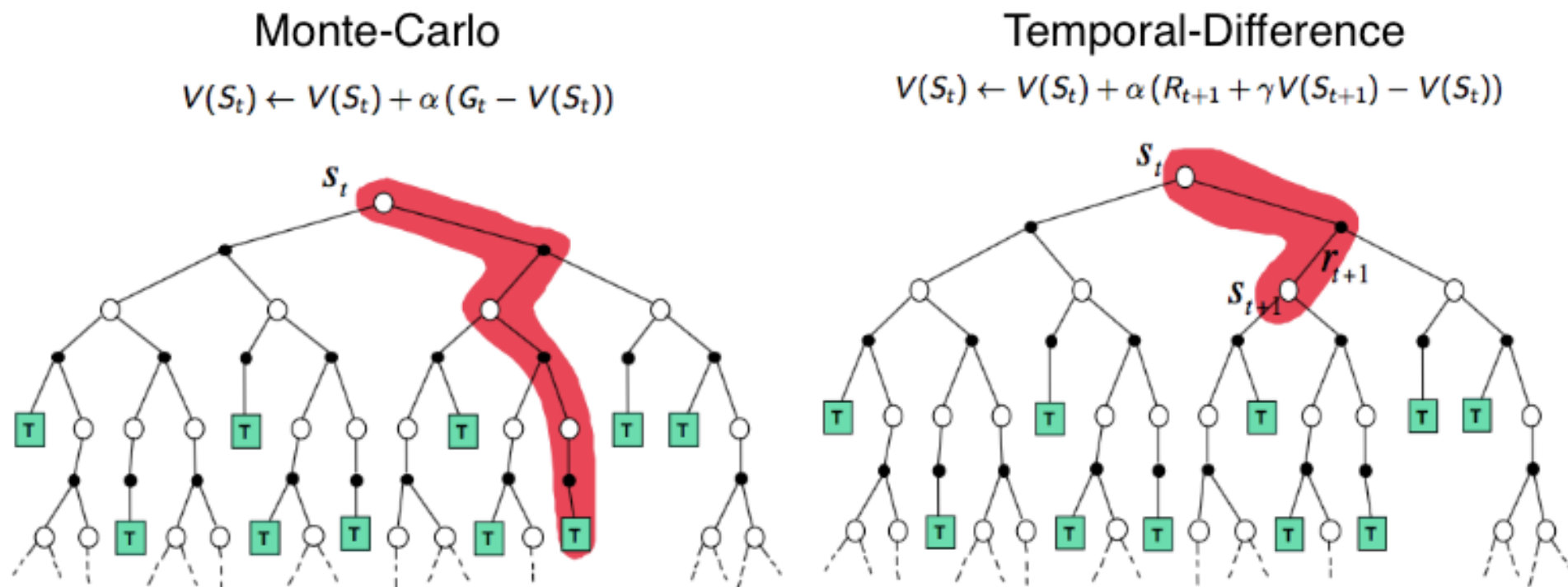


|   |  |  |  |  |
|---|---|---|---|---|
|    | 0   | 0   | 0   | 0   |
|    | 0   | 0   | 0   | 0   |
|  | 0   | 0   | 0   | 0   |
|  | 0   | 0   | 0   | 0   |
|  | 0   | 0   | 0   | 0   |
|  | 0   | 0   | 0   | 0   |

→  
Training

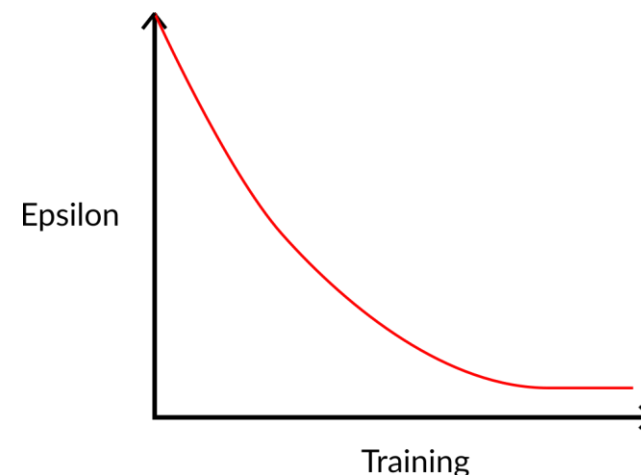
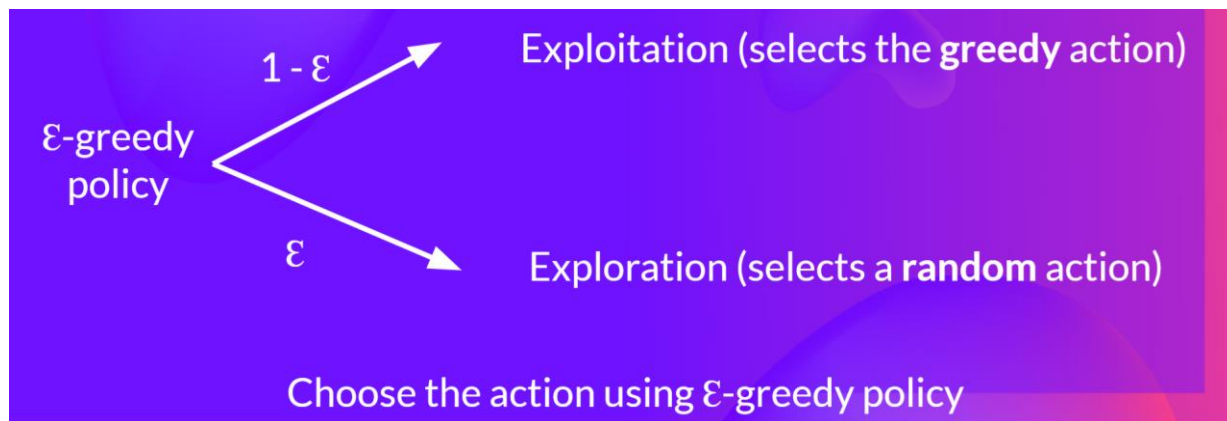
|   |  |  |  |  |
|---|---|---|---|---|
|    | 0   | 10.8  | 0   | 0   |
|    | 0   | 9.9   | 0   | -10   |
|  | 0   | 0   | 0   | 10  |
|  | 0   | -10   | 0   | 0   |
|  | 0   | 0   | 0   | 0   |
|  | 0   | 0   | 0   | 0   |

- 如何评估价值？先考虑状态的价值  $V(s)$ 
  - $V(s)$  由后续状态转换（直至终态）中产生的累计奖励决定
  - 两种估计方法：蒙特卡洛（MC）和时序差分（TD）



- 如何学习？多玩多练

- 每个步骤决策行动时，以一定概率选择“探索”（随机动作） 还是“利用”（最佳动作）



- 根据观察到的  $S_t, A_t, R_t, S_{t+1}$ ，基于 TD 更新  $Q(S_t, A_t)$
- 多轮交互迭代，Q 函数的估计逐渐收敛到准确值

$$\underbrace{Q(S_t, A_t)}_{\text{New Q-value estimation}} \leftarrow \underbrace{Q(S_t, A_t)}_{\text{Former Q-value estimation}} + \underbrace{\alpha}_{\text{Learning Rate}} [\underbrace{R_{t+1}}_{\text{Immediate Reward}} + \underbrace{\gamma \max_a Q(S_{t+1}, a)}_{\text{Discounted Estimate optimal Q-value of next state}} - \underbrace{Q(S_t, A_t)}_{\text{Former Q-value estimation}}]$$

New  
Q-value  
estimation

Former  
Q-value  
estimation

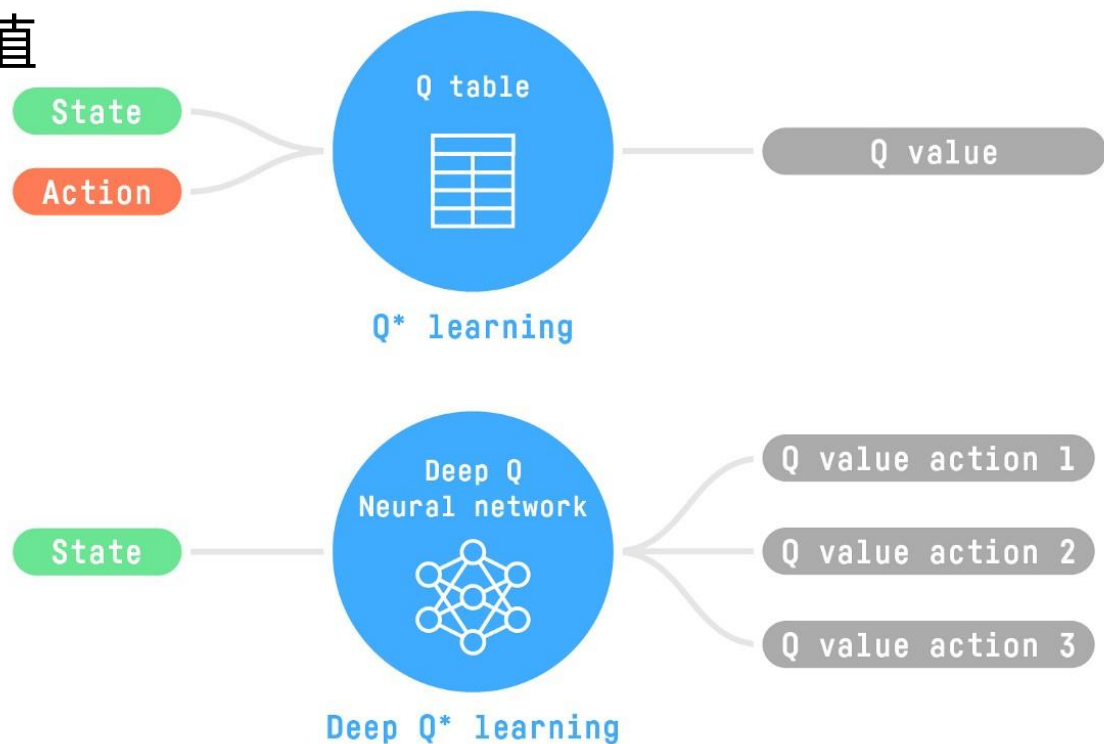
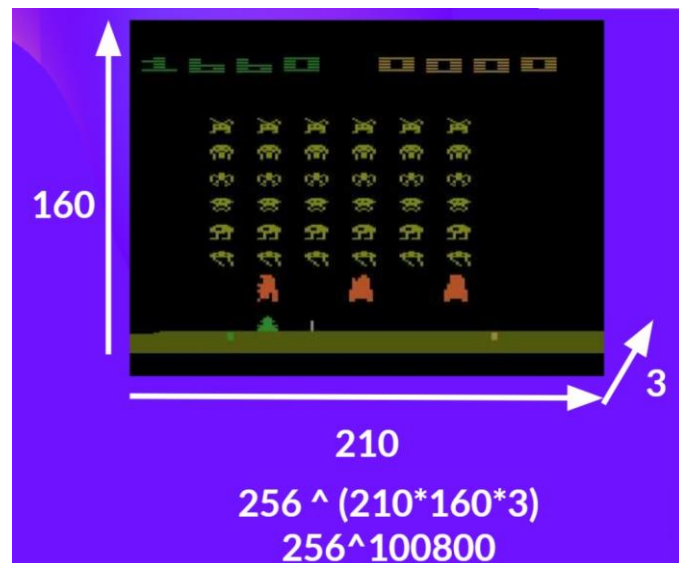
Learning  
Rate

Immediate  
Reward

Discounted Estimate  
optimal Q-value  
of next state

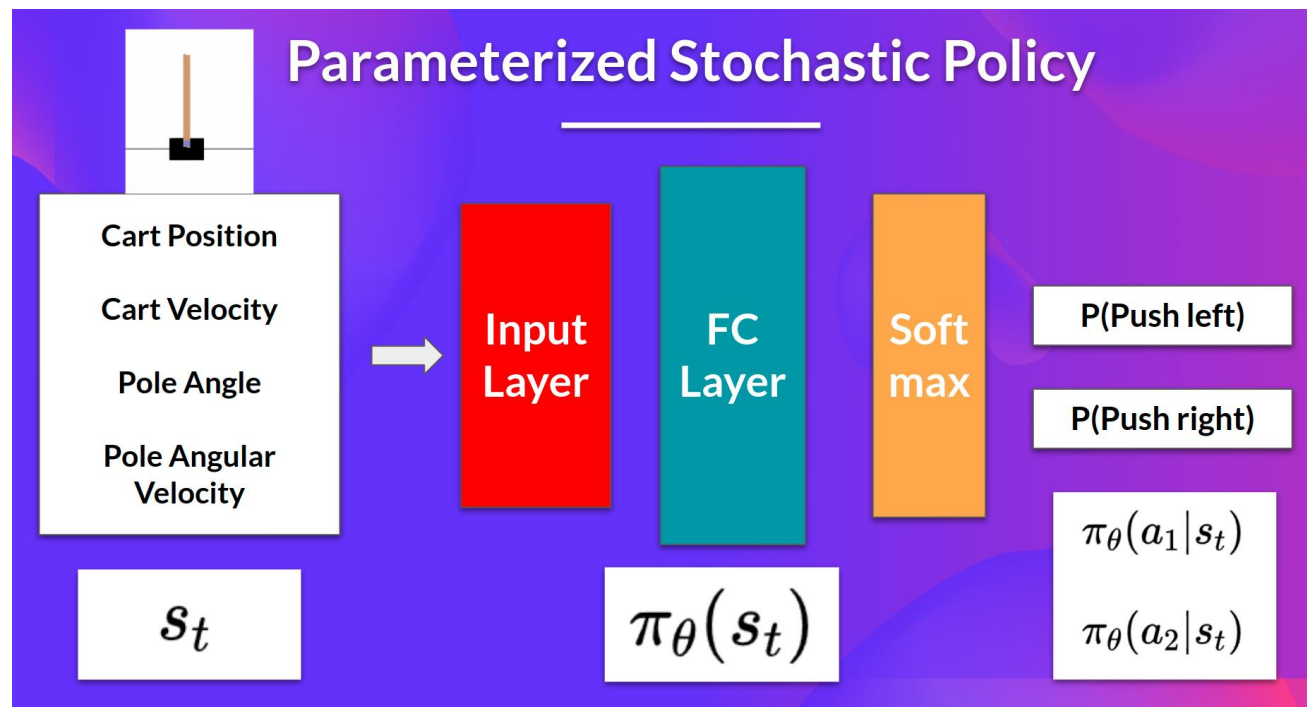
Former  
Q-value  
estimation

- 现实世界中的问题：
  - 状态和动作空间巨大或者连续
  - Q 函数表格巨大，难以充分训练
- 使用神经网络模型替代表格拟合 Q 函数
  - 输入为状态向量，输出为每个动作的价值
  - 仍然要求动作空间离散
- 深度强化学习 (Deep RL)



# 强化学习的实现

- 第二类方法：基于策略 (Policy-Based)
  - 直觉：既然目的是获得游戏的最佳策略，那就直接学习最优的 状态  $\rightarrow$  动作 映射
  - 输入：状态向量，输出：采取各行动的概率分布，使用神经网络模型实现

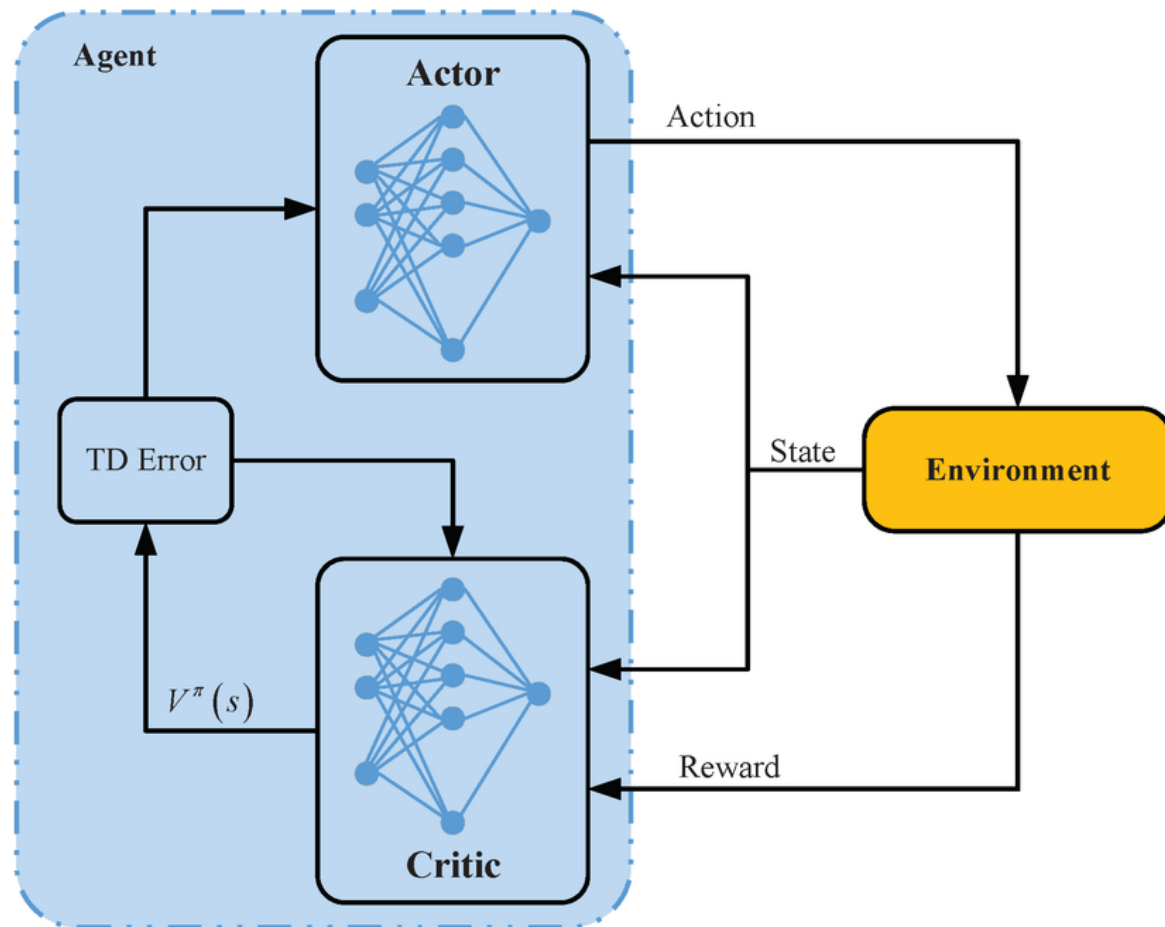




- 朴素想法：观察一条“轨迹”上奖励总和，为正则强化、为负则弱化所有动作概率
  - 问题：滥竽充数，鹤立鸡群？
  - 需要评估某动作相比其它动作的好坏
  - V 函数和 Q 函数！
- Actor-Critic 算法
  - 同时训练一个策略模型和一个评估模型

**Advantage Function**

$$A(s, a) = \underbrace{Q(s, a)}_{\substack{\text{q value for action } a \\ \text{in state } s}} - \underbrace{V(s)}_{\substack{\text{average} \\ \text{value} \\ \text{of that} \\ \text{state}}}$$



# 总结：强化学习

---

- 是机器学习领域下与监督学习并列的一个分支
- 不需要有标注的样本，而需要一个可交互，能给予反馈的环境
- 目标是学习最优的策略以获得最大奖励，一般通过神经网络模型实现
- 应用场景：自动驾驶，具身智能（机器人），游戏 AI，...

# 大模型与强化学习

- 监督微调：使用输入输出对数据，让模型学习标准答案
- 很多场景中，我们能评判模型的回答质量好不好，但不容易写出标准答案
  - 希望模型能够输出我们喜好的答案——人类偏好对齐（Alignment）
  - 模型生成回答 → 获得打分 → 模型改善回答以获取更高得分
- 大模型生成的强化学习建模
  - 环境状态：输入+已生成的 Token 序列
  - 行动：大模型选择继续生成下一个 Token
  - 奖励：人类的打分

## 与标准 RL 的不同：

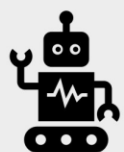
1. 大模型本身已经是很好的策略
2. 奖励信号可能只在最后一步出现（对于完整生成序列的打分）

- Q1：如何探索一个状态下的多种动作？
  - 大模型生成的是下一个 Token 的概率分布，可以采样 (Sampling)
  - 回想：探索 (Exploration) 与 利用 (Exploitation)
- Q2：模型生成的所有结果都让人类来打分，成本过高？
  - 先收集一些人类打分样本，训练一个和人类打分结果相似的奖励模型 (Reward Model)，帮我们干活
- Q3：让人类对模型生成内容进行绝对值打分，标准可能也比较模糊？
  - 让人类对不同内容根据质量做排序，通过对比学习训练能打分的奖励模型

### Step 1 Supervised Fine-Tuning

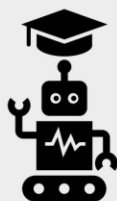


Collect human  
demonstration data



Base LLM

Supervised  
Fine-tune



SFT

### Step 2 Training a Reward Model



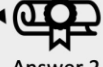
Prompt



SFT



Answer 1



Answer 2



Answer 1



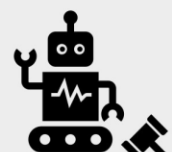
Answer 2

Collect human preference data



Base LLM

Supervised  
Fine-tune



RM

### Step 3 Optimize Policy



Prompt



Policy



Answer



RM

Reward

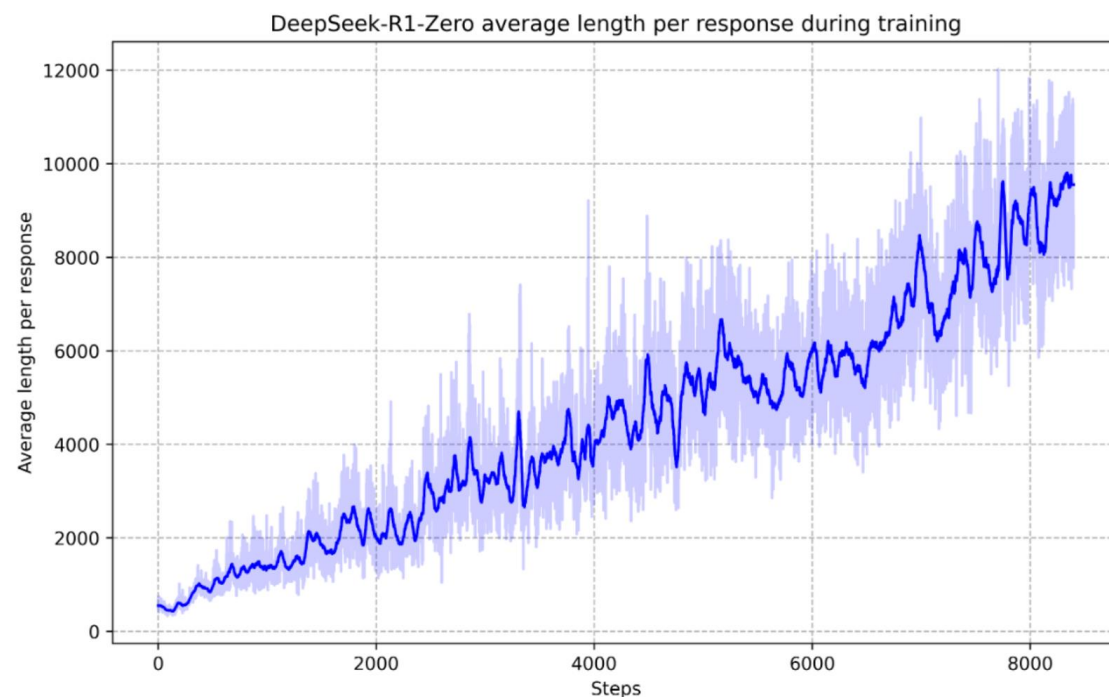
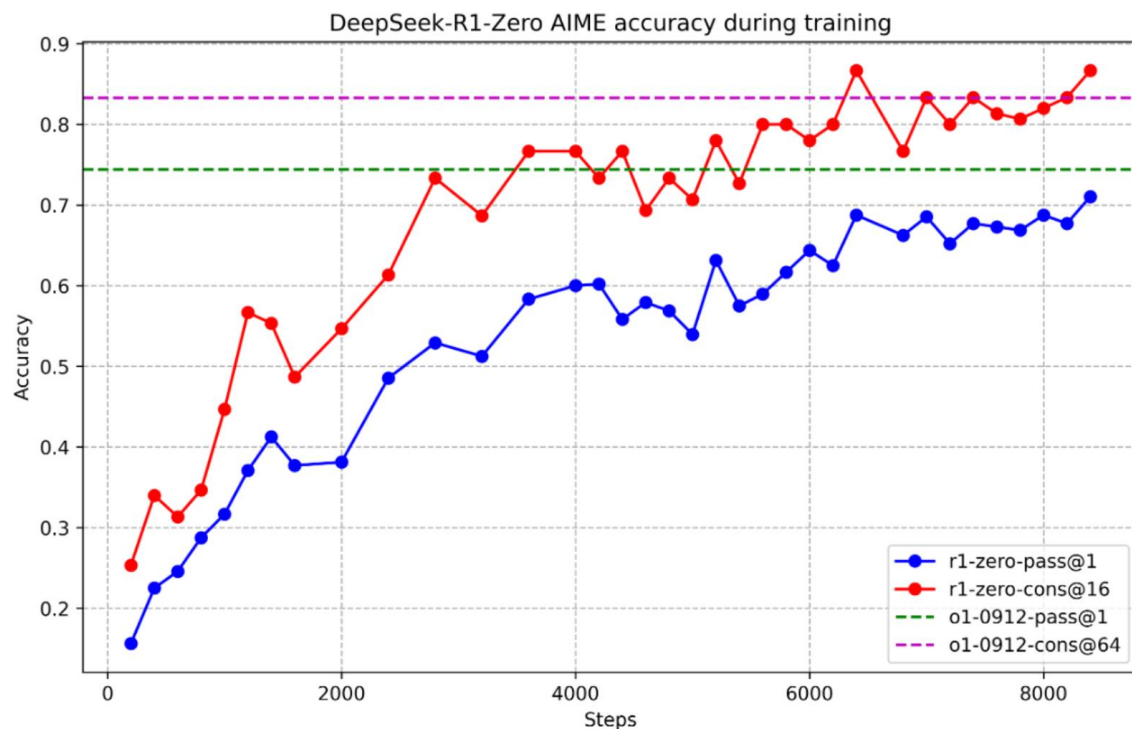
PPO

# "长思考"：强化学习的意外产物

- 24 年下半年：大模型基础能力接近瓶颈，困难任务上的表现难以提升
  - “困难”任务：AIME（美国数学邀请赛题目），SWE-Bench（Github 真实 issue）
- OpenAI-o1, DeepSeek-R1：“深度思考”
- DeepSeek-R1 的强化学习训练
  - 数据：竞赛级数学题和编程题
  - 奖励：正确性奖励 + 格式奖励
    - 正确性：**只看结果对不对，不看过程**（数学题看答案，编程题看 AC）
    - 格式：输出了思考过程且放在 `<think> </think>` 中

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:

## R1 强化学习提示词设计：要求输出思考过程（但奖励中不会对长思考加分）



在训练过程中，观察到模型自发输出了越来越长的思考过程，也获得了越来越高的奖励（在 AIME 数据集上的表现不断提升）

---

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  $\dots$

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

Rearrange to isolate the inner square root term:

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

$\dots$

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  $\dots$

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:  $\dots$

$\dots$

---

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.

“奇点时刻”：以做对题为目标的大模型突然学会了拟人的反思口吻



# 大模型强化学习的反思

- 强化学习有效的前提？
  - 训练前的模型必须有足够强的能力，能够探索到正确答案，否则奖励信号永远为 0
  - 用小尺寸模型尝试复现 R1 时，Llama 系列效果很差，Qwen 系列用 RL 就能训起来
- 为什么要先做监督微调再做强化学习？
  - “先背书、再做新题”——提高模型 RL 时探索到正确答案的可能性，加快训练
  - 奖励只看结果，但希望模型的思考过程也尽量规范，减少“野路子”
    - R1 的第一版训练尝试了不做 SFT 直接 RL，结果准确率也很高，但做题的思考过程中英文混杂，可读性很差



# 感谢聆听

---

Any Question?