penAI

o1-preview

Model

GPT-4o
Great for most tasks

o1-preview
Uses advanced reasoning

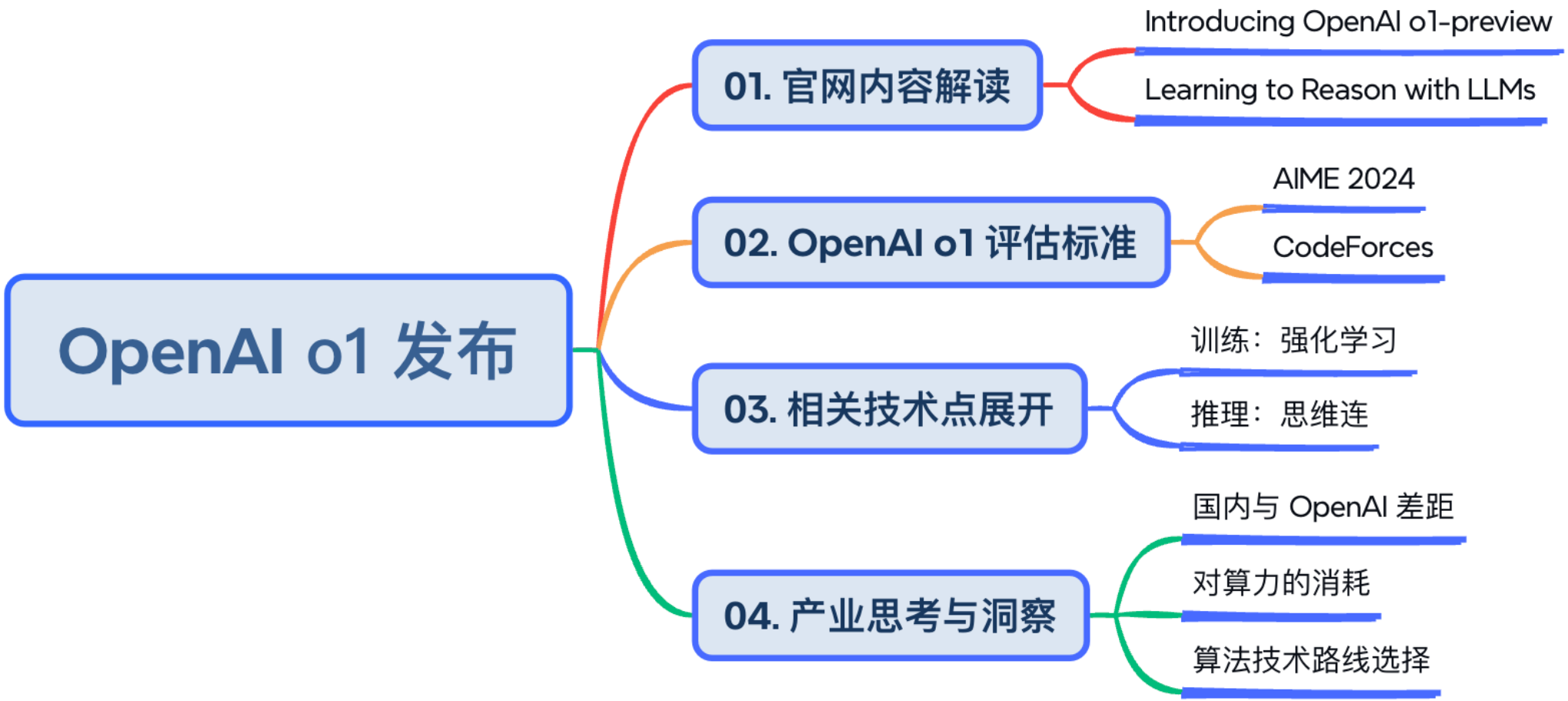
o1-mini
Faster at reasoning

More models

Temporary chat

1

视频内容


OpenAI o1 发布
01. 官网内容解读
Introducing OpenAI o1-preview
Learning to Reason with LLMs
02. OpenAI o1 评估标准
AIME 2024
CodeForces
03. 相关技术点展开
训练：强化学习
推理：思维连
04. 产业思考与洞察
国内与 OpenAI 差距
对算力的消耗
算法技术路线选择

01

OpenAI o1 官网解读

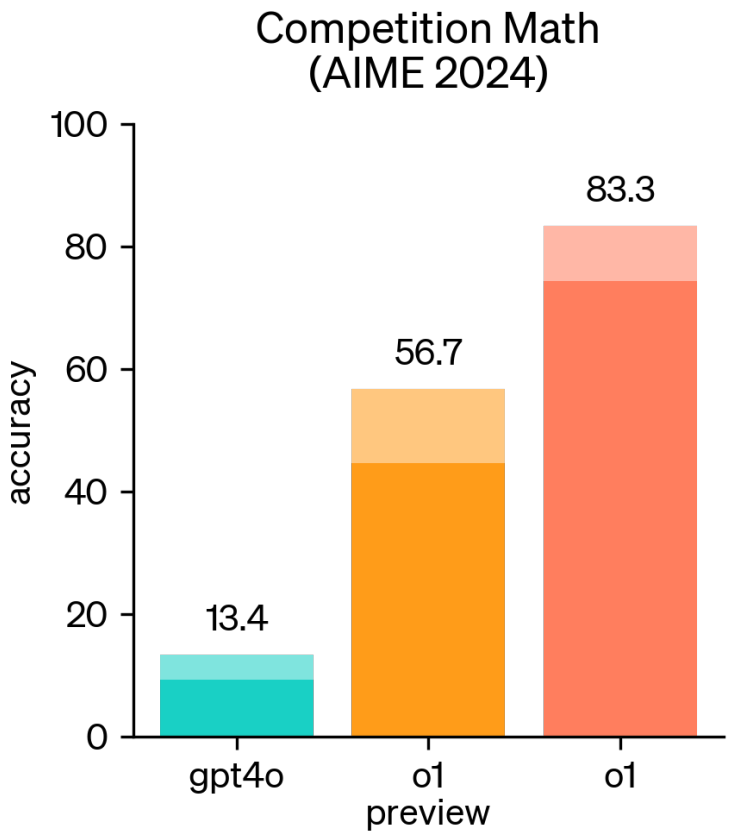
官网链接

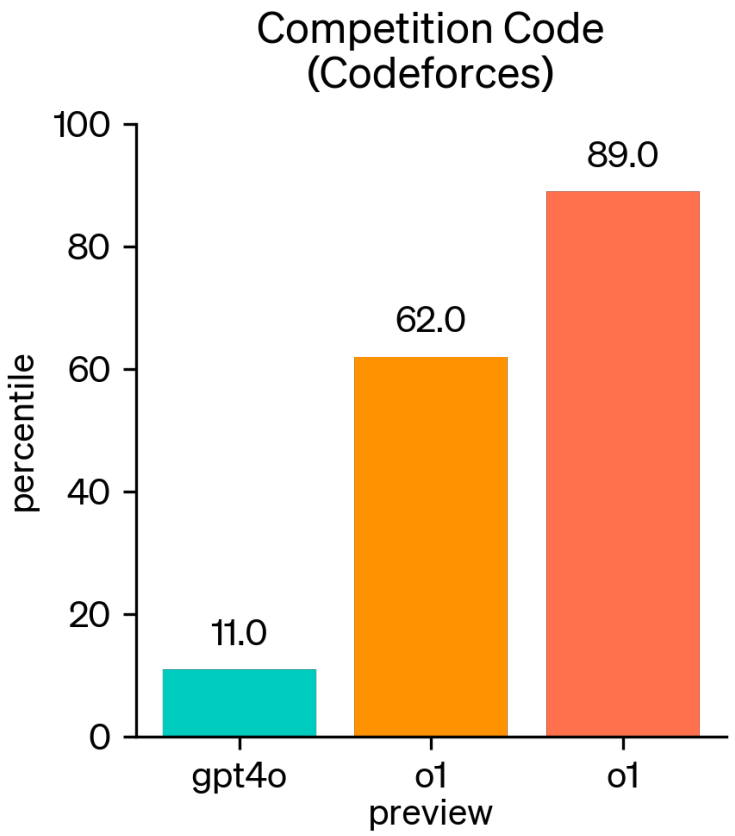
1. https://openai.com/o1/
2. https://openai.com/index/introducing-openai-o1-preview/
3. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/
4. https://openai.com/index/learning-to-reason-with-llms/

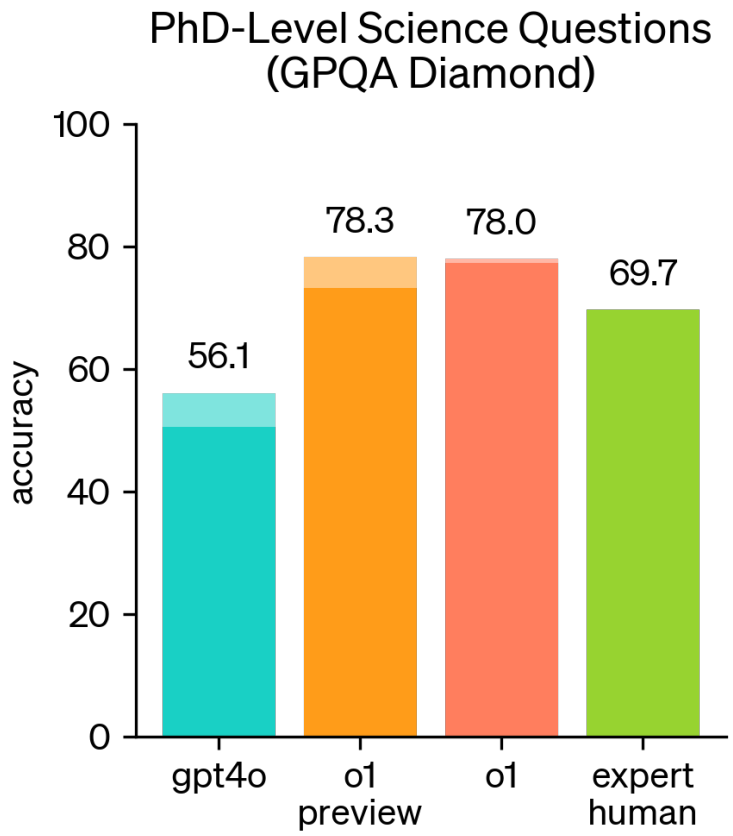
02

OpenAI o1 评估标准

评估标准


Competition Math
(AIME 2024)
accuracy
100
80
60
40
20
0
13.4
56.7
83.3
gpt4o
o1
preview
o1


Competition Code
(Codeforces)
percentile
100
80
60
40
20
0
11.0
62.0
89.0
gpt4o
o1
preview
o1


PhD-Level Science Questions
(GPQA Diamond)
accuracy
100
80
60
40
20
0
56.1
78.3
78.0
69.7
gpt4o
o1
preview
o1
expert
human

AIME 2024

1. 美国数学邀请赛（American Invitational Mathematics Examination）简称为AIME，是介于AMC10、AMC12及美国数学奥林匹克学术活动（USAMO）之间的一个数学学术活动。

https://artofproblemsolving.com/wiki/index.php/2024_AIME_I

CodeForces

1. CodeForces 一个提供在线编程评测系统的俄罗斯网站。该网站由一群来自ITMO大学的程序员创建并维护。

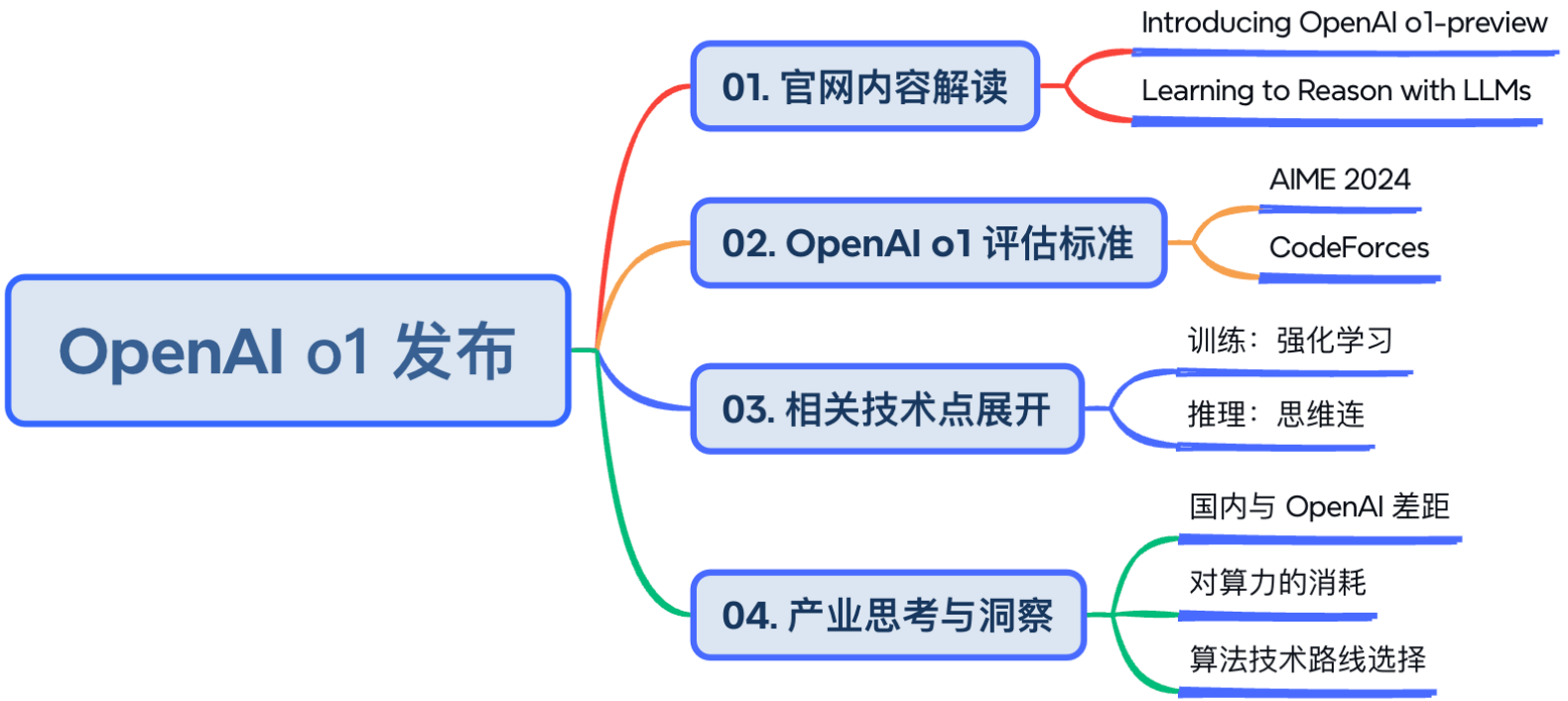
https://codeforces.com/

03

强化学习 RL
思维链 CoT

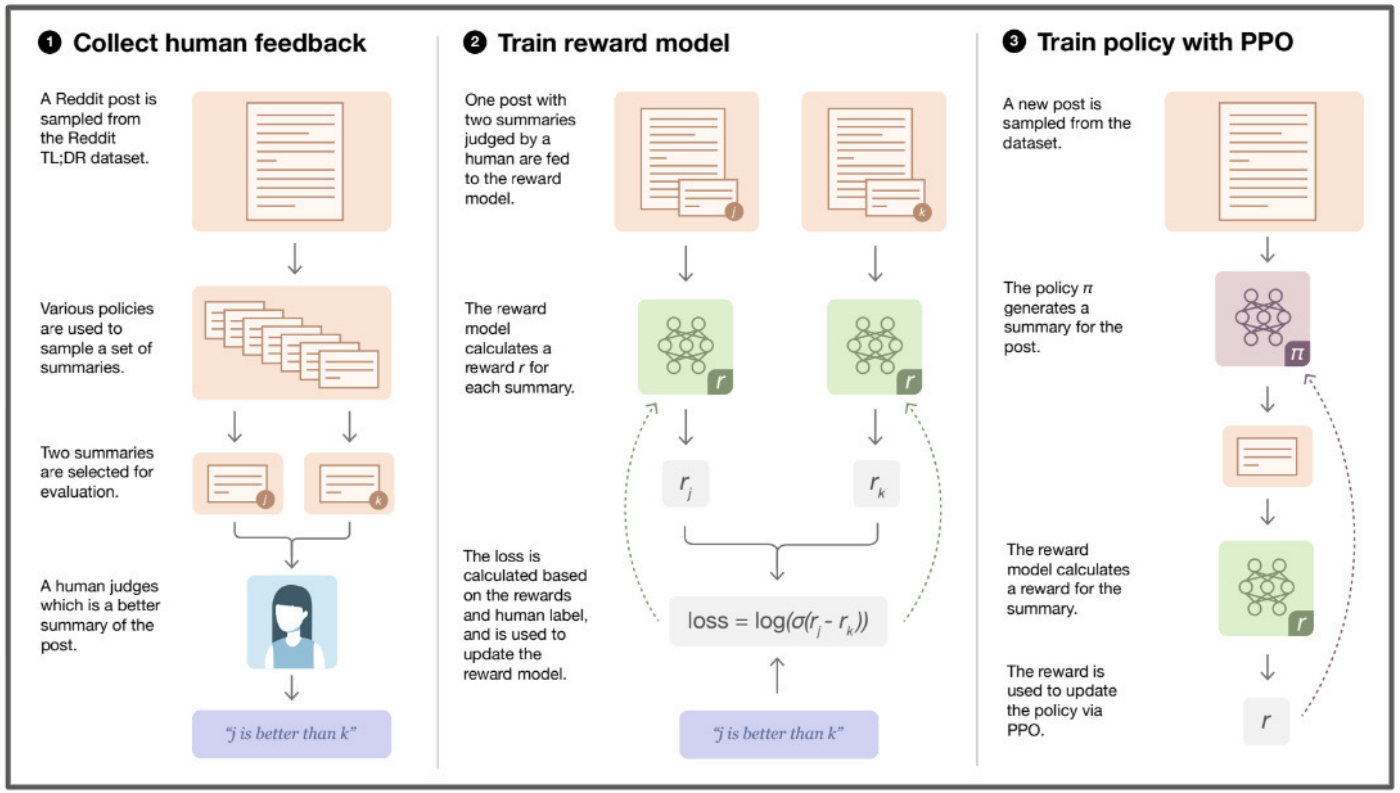
强化学习 Reinforcement Learning

1. ChatGPT： RLHF
2. LLAMA3.1： DPO
3. Quiet-STaR

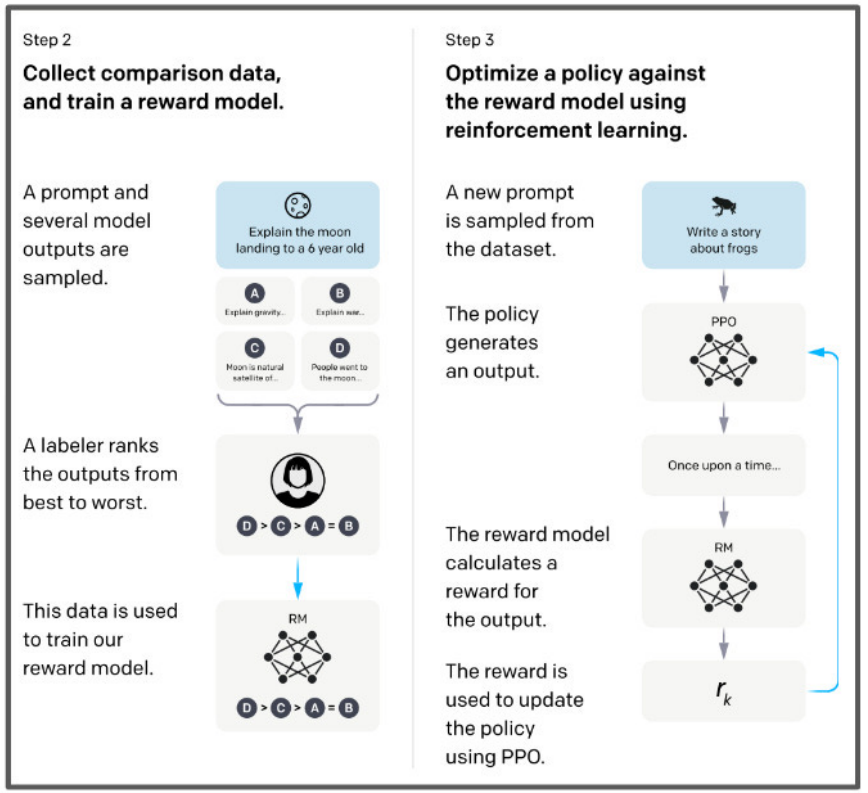


1、ChatGPT

RLHF in [1]

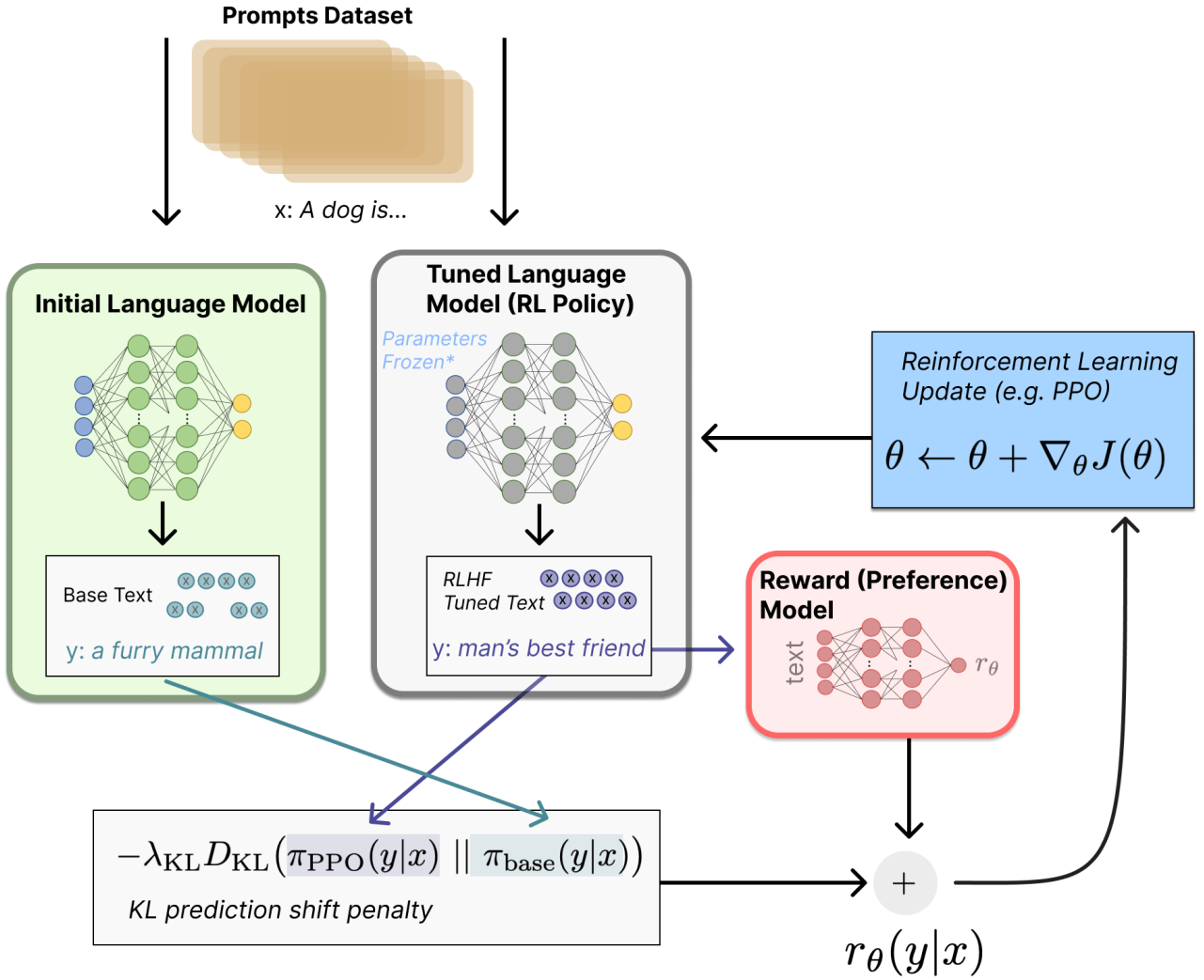


RLHF in [2]



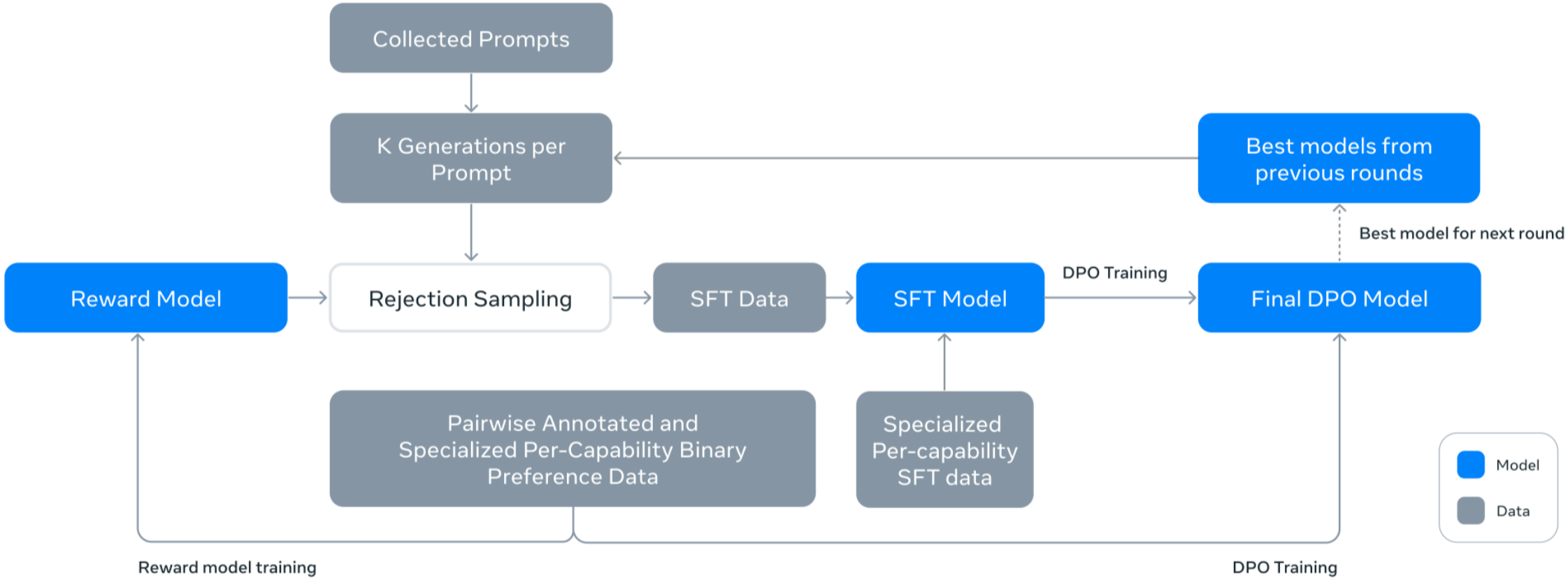
1、ChatGPT

1. **第一阶段：**冷启动阶段的监督策略模型
2. **第二阶段：**训练奖励模型（Reward Model，RM），通过人工标注数据训练回报模型
3. **第三阶段：**采用 PPO（Proximal Policy Optimization，近端策略优化）强化学习来优化策略。



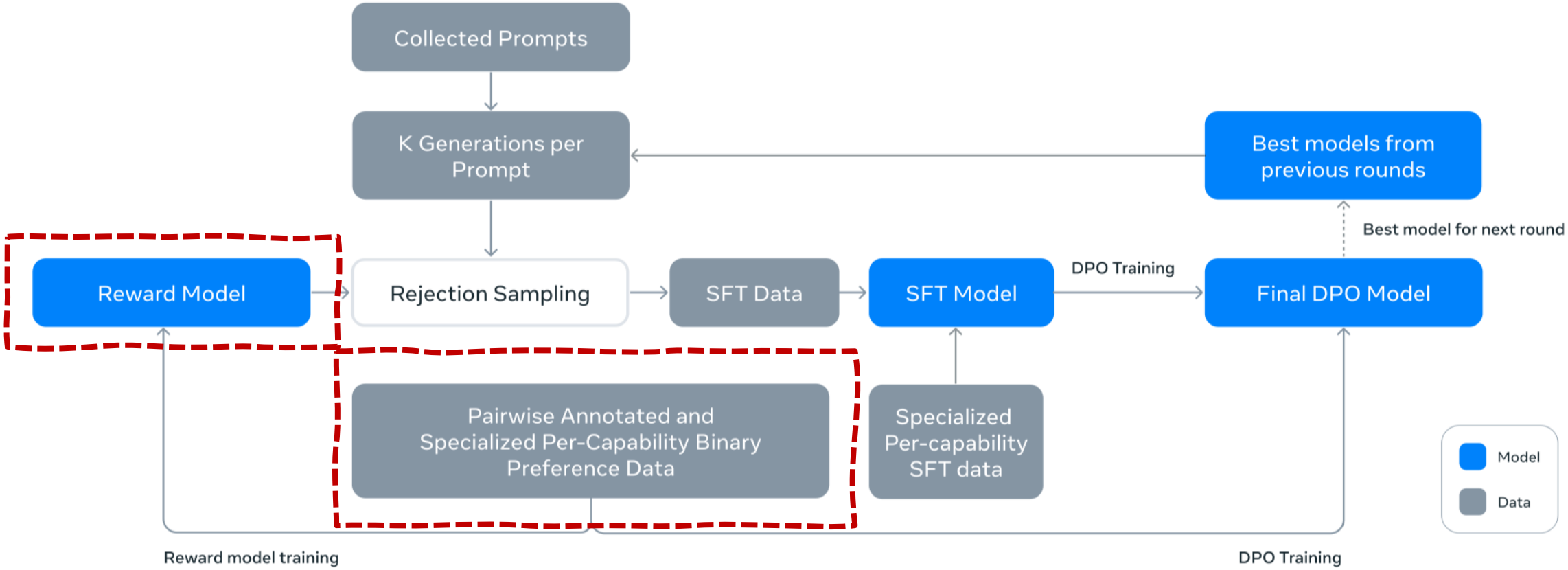
2、LLAMA3.1 后训练过程

- 通过多轮对齐来完善 Chat 模型，基于监督微调（SFT）、拒绝采样（RS）和通过 DOP 直接优化偏好。



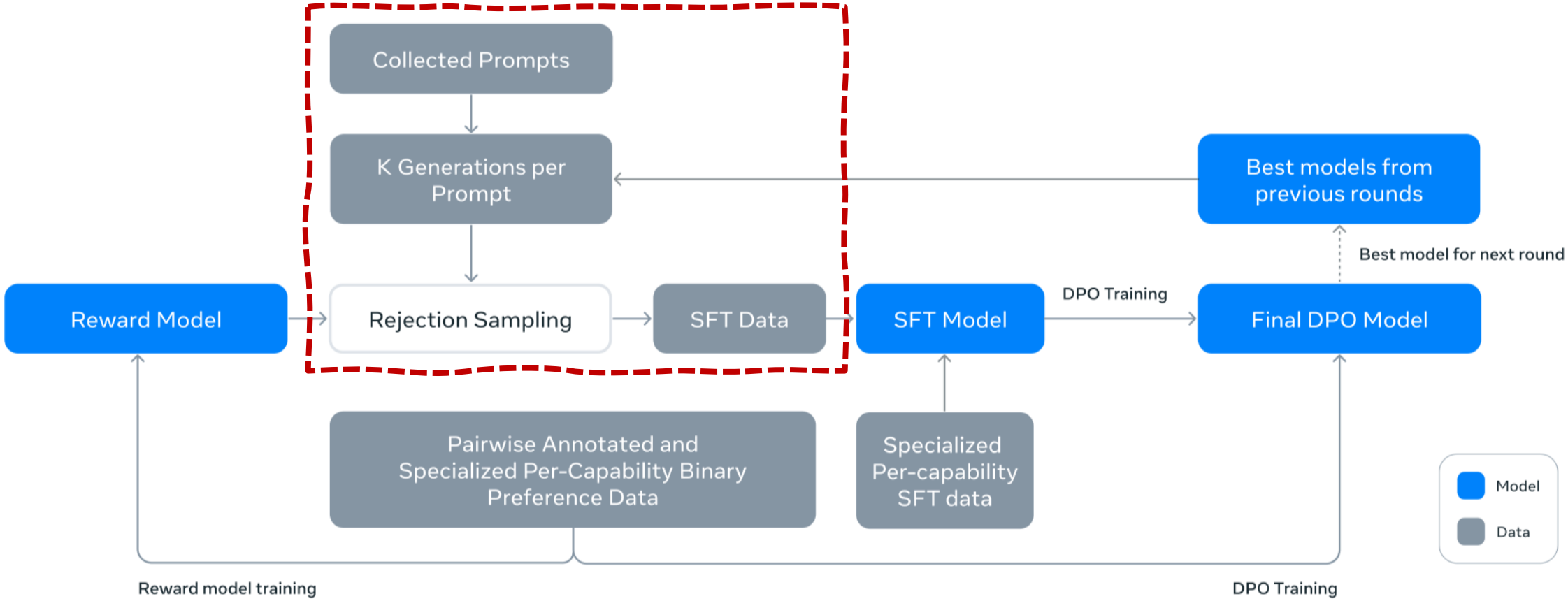
2、LLAMA3.1 后训练过程

1. **训练 RM 模型：**人工标注数据训练 RM 模型，用来评价 <prompt, answer> 数据对质量；



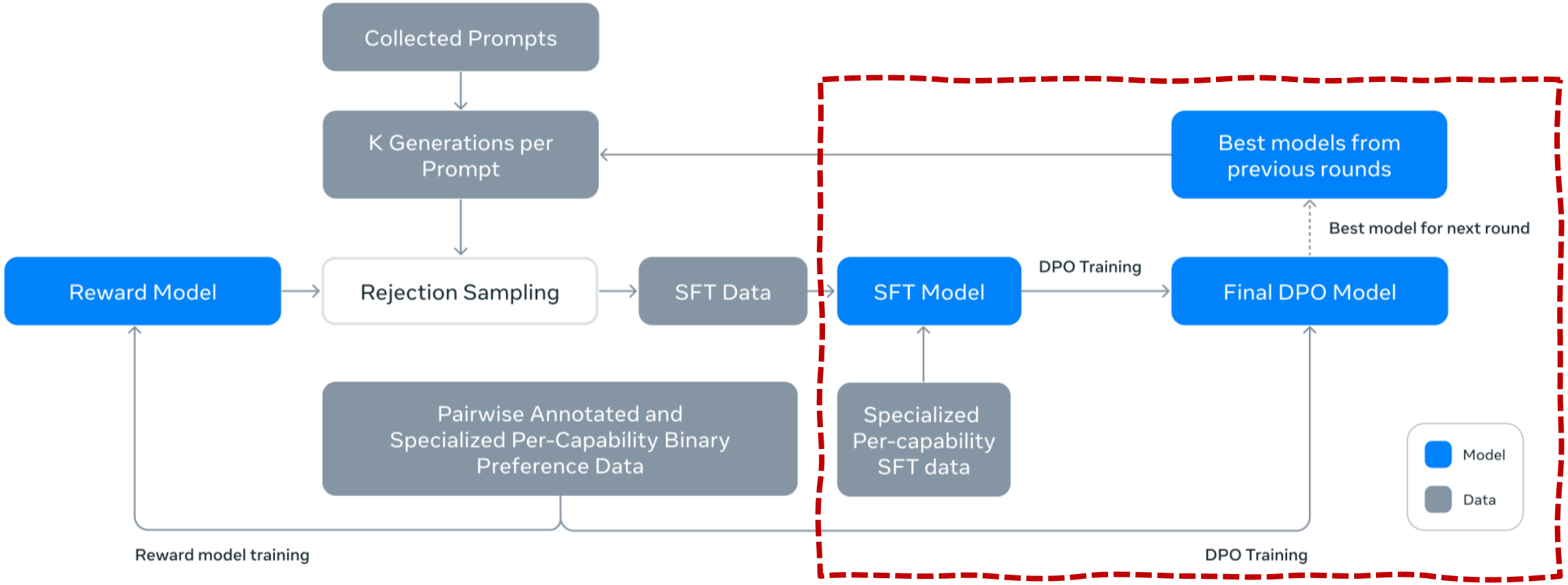
2、LLAMA3.1 后训练过程

2. **拒绝采样（Rejection Sampling）**：对输入 Prompt，模型生成若干个回答，RM 给予质量打分，选择得分最高保留作为 SFT 数据，其它抛掉；



2、LLAMA3.1 后训练过程

3. **DPO 训练：**人工标注数据给 DPO 模型调整 LLM 参数。DPO本质为二分类，从人工标注 <Prompt, Good Answer, Bad Answer> 三元数据里学习，调整模型参数并鼓励输出 Good Answer。



2、LLAMA3.1 后训练过程

- 上述过程会反复迭代，流程相同，区别 Rejection Sampling 阶段用来对给定 Prompt 产生回答 LLM 模型，从上一轮流程最后产生若干不同 DPO 模型，选择最好的那个在下一轮拒绝采样阶段给Prompt生成答案。随迭代增加 DPO 模型越来越好，拒绝采样能选出的答案质量越高，SFT 模型就越好，形成正反馈循环。

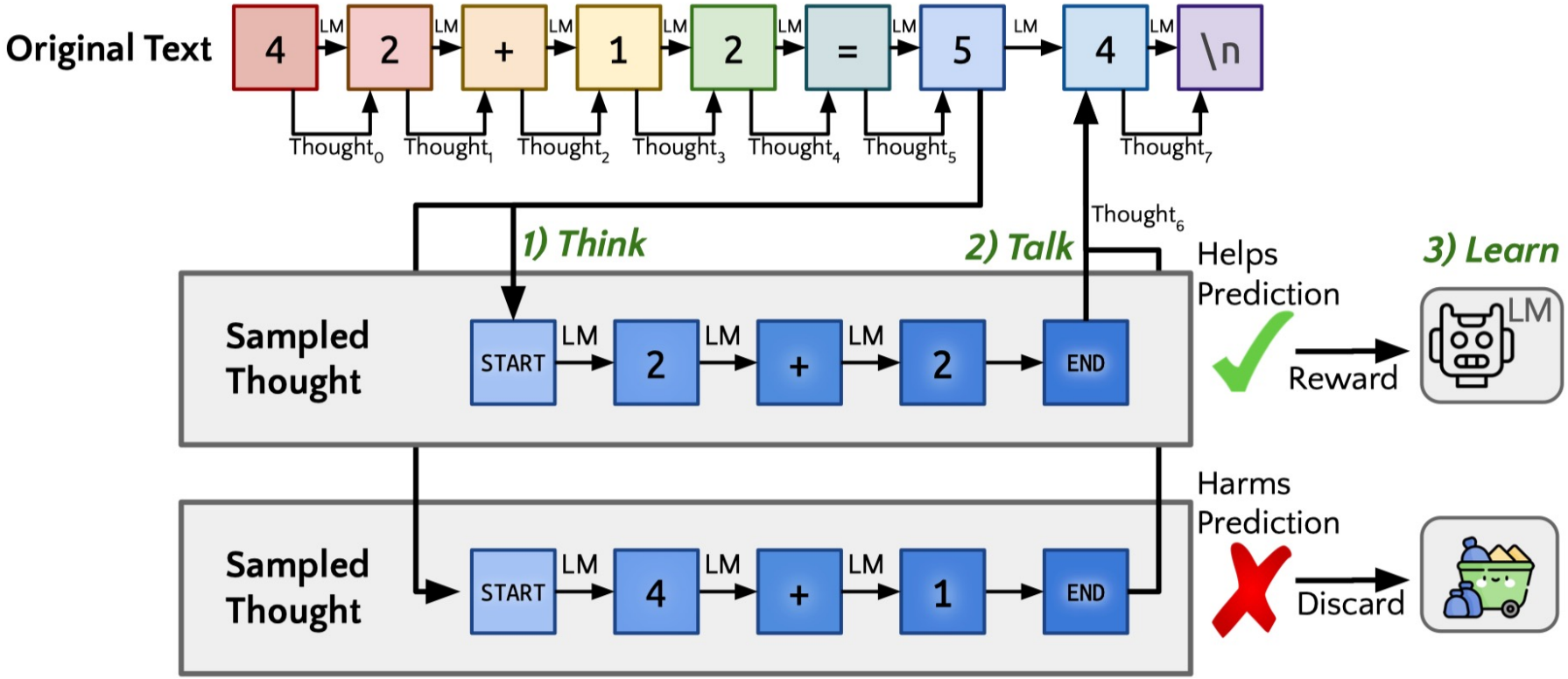
- **RM 在后训练作用区别：**RLHF 是把 RM 打分用在 PPO 强化学习阶段；Llama3.1 用 RM 筛选高质量 SFT 数据。
- **SFT 数据合成：**因为拒绝采样过程的回答由 LLM 产生，因此采用合成数据来训练 SFT 模型。

2、LLAMA3.1 提高特定下游任务性能

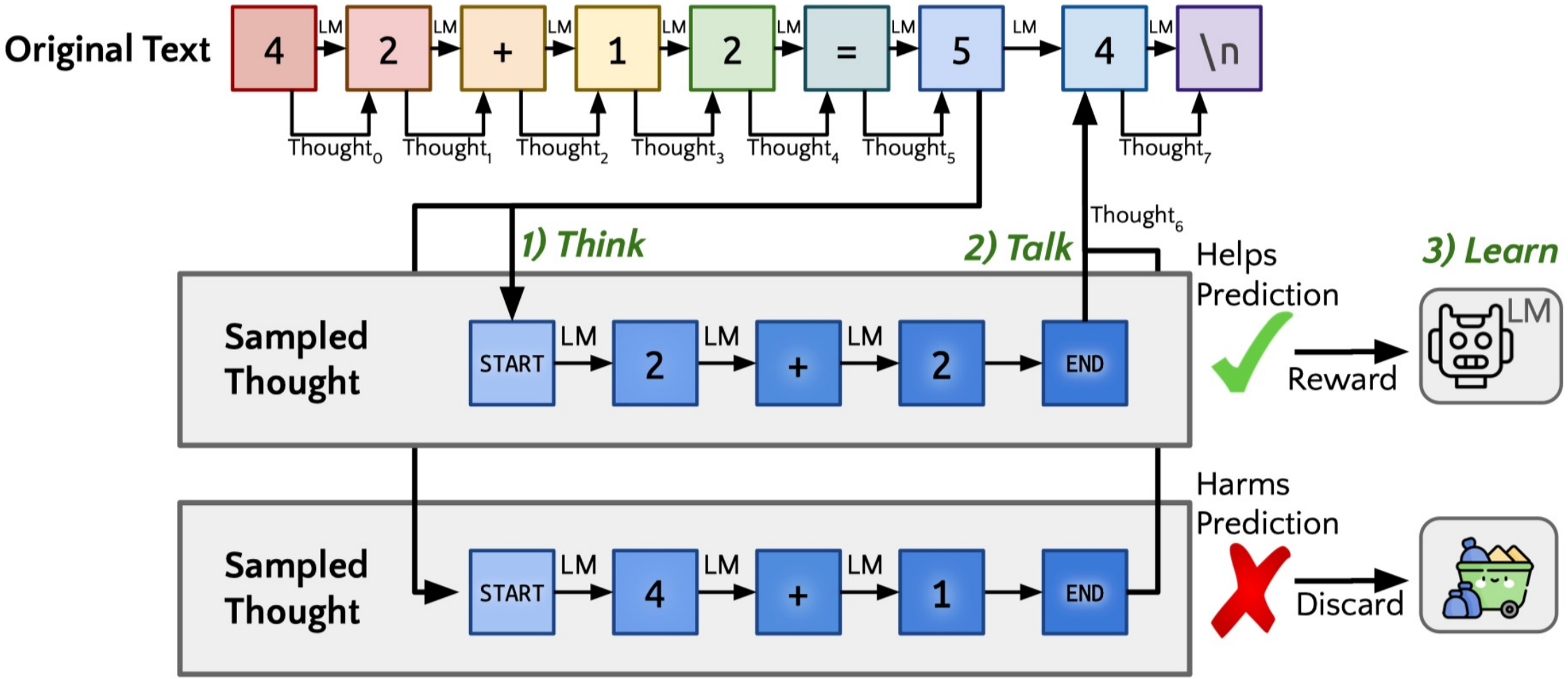
1. **DPO 原因：**采用 DPO 的 RLHF 而非更复杂 RL 强化学习算法， 因为后者稳定性不确定且制约AI 集群规模扩展 Scaling Law；
2. **提高模型编码能力：**采用训练代码专家、生成 SFT 合成数据、通过系统提示引导改进格式，以及创建质量过滤器（从训练数据中删除不良样本）等方法。
3. **Float8 量化推理：**将权重/输入量化为 fp8，然后乘以缩放因子 scale，fp8 x fp8 输出 bf16。使得推理速度更快，显存占用更少（推理在 H100 云资源集群）。

3、Quiet-STaR 强化学习新范式

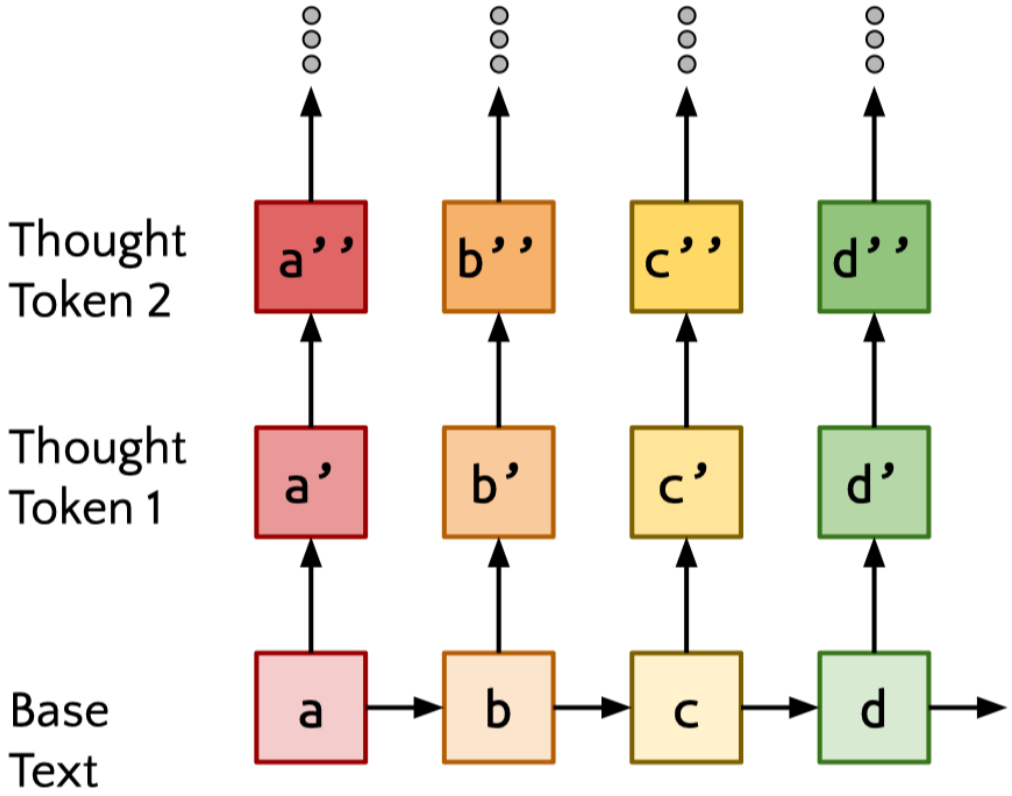
- LLM 刷榜依赖于特定任务微调。Quiet-STaR 采用全新范式：输入 token 后插入“思考”步骤，让打磨新内部推理。系统评估推理是否有助于预测后续文本，并相应地调整模型参数。



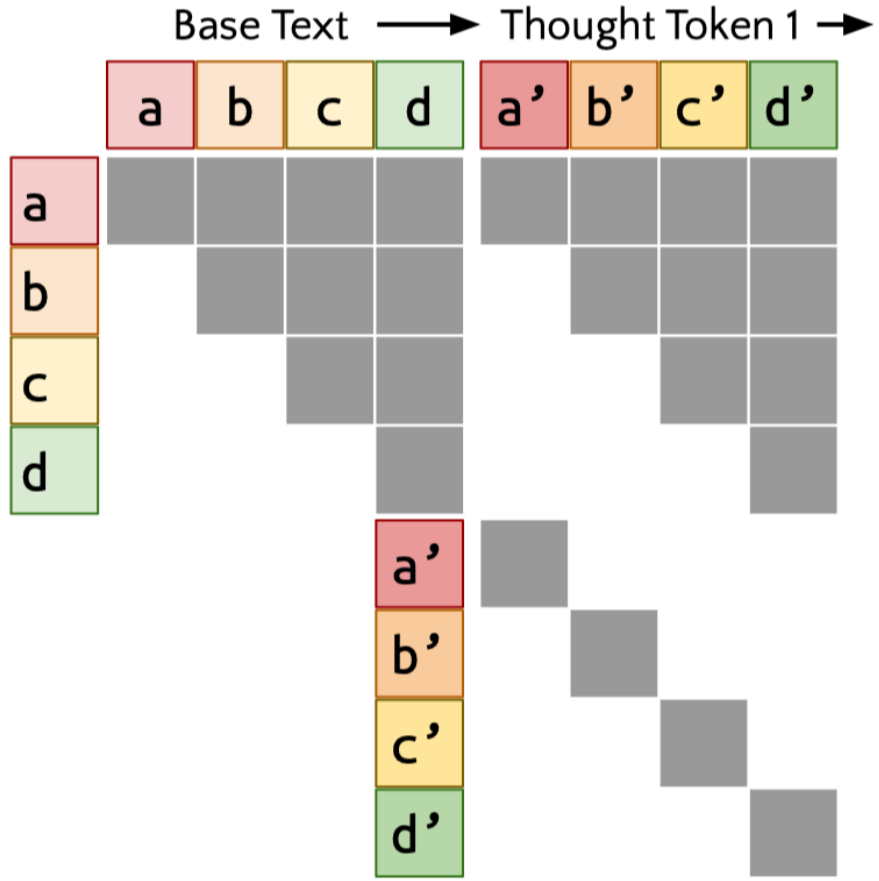
3、Quiet-STaR


Original Text
4
LM
2
LM
+
LM
1
LM
2
LM
=
LM
5
LM
4
LM
\n
Thought0
Thought1
Thought2
Thought3
Thought4
Thought5
Thought7
Thought6
1) Think
2) Talk
3) Learn
Sampled Thought
START
LM
2
LM
+
LM
2
END
Helps Prediction
Reward
LM
Sampled Thought
START
LM
4
LM
+
LM
1
END
Harms Prediction
Discard

3、Quiet-STaR：并行加速思考 thought 过程

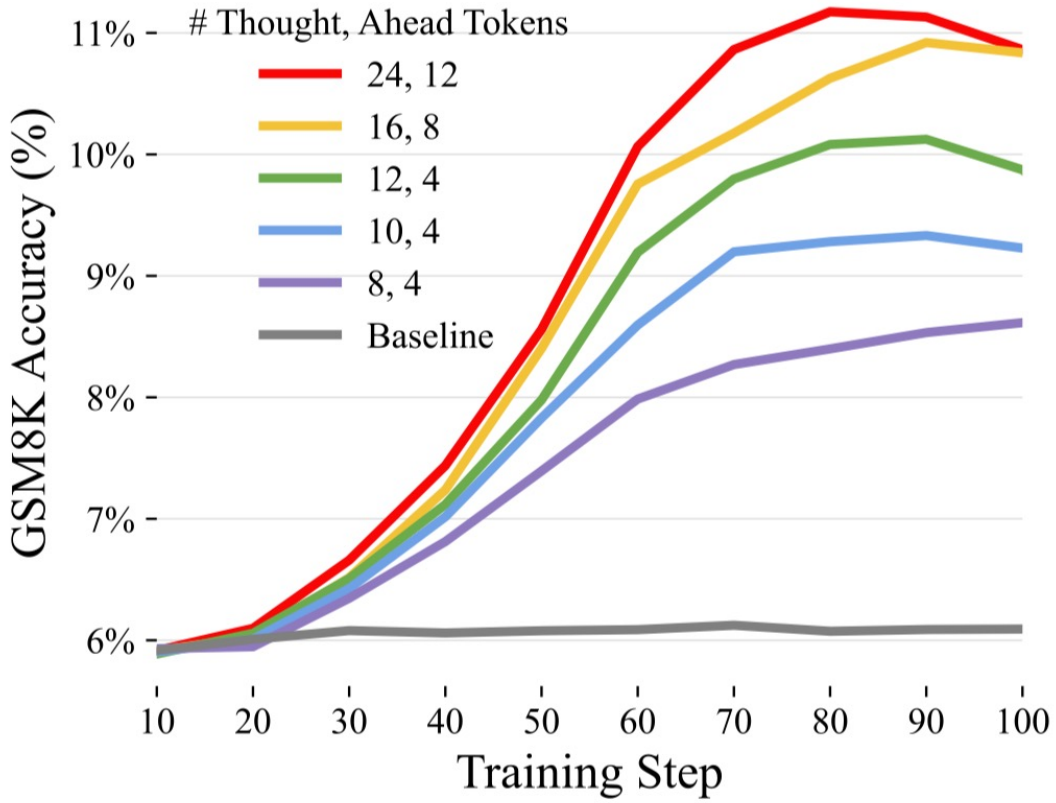


Generation Paths

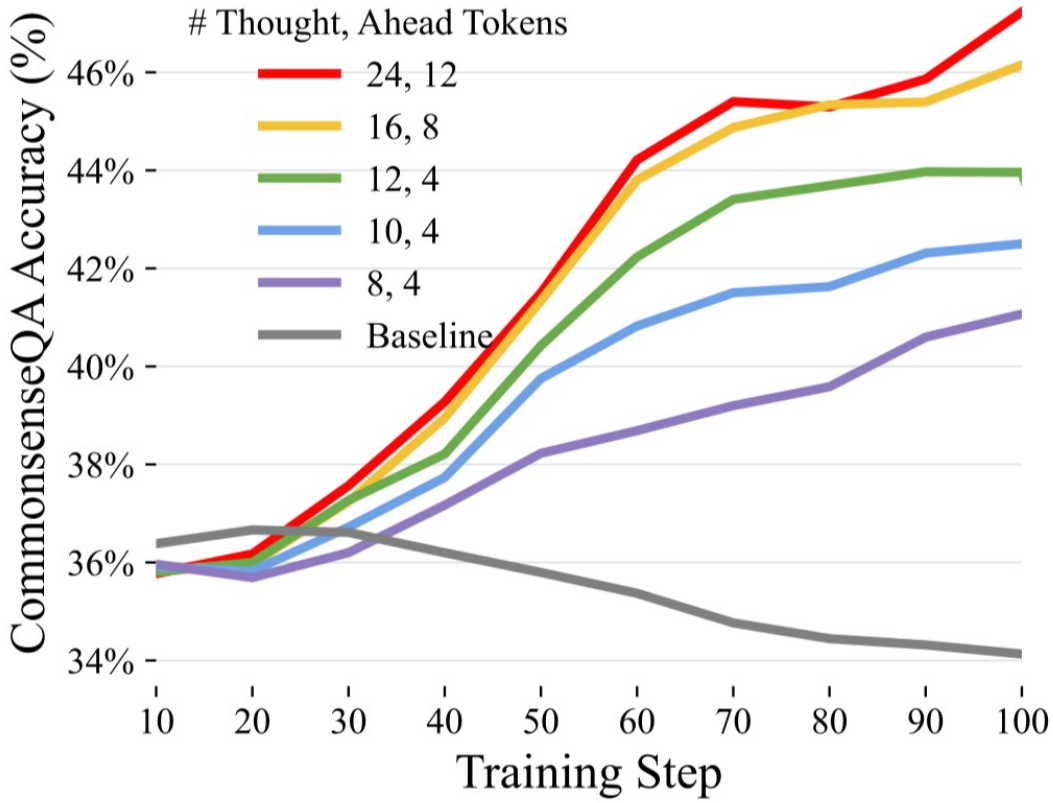


Parallel Inference Mask

3、Quiet-STaR：下游任务效果提升明显



(a) GSM8K



(b) CommonsenseQA

3、Quiet-STaR 与 CoT 结合

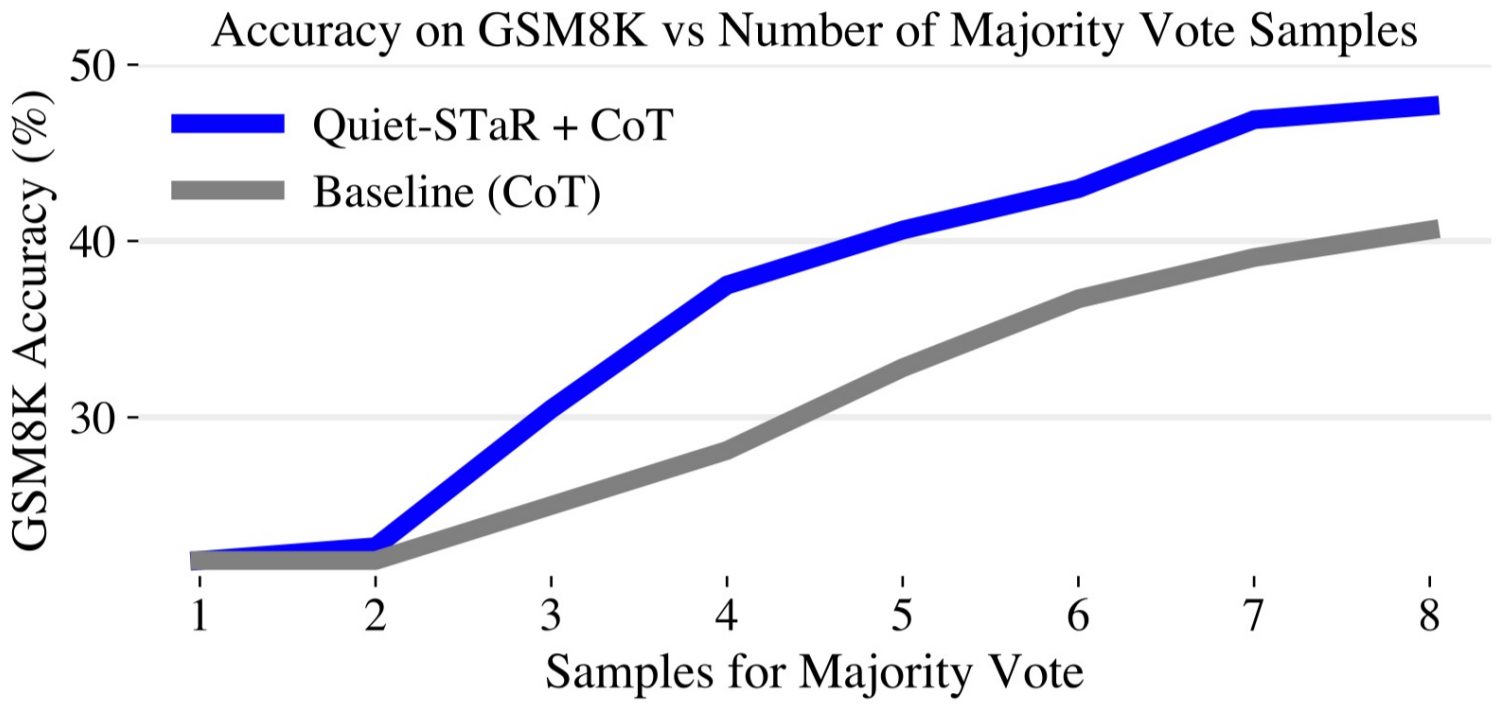
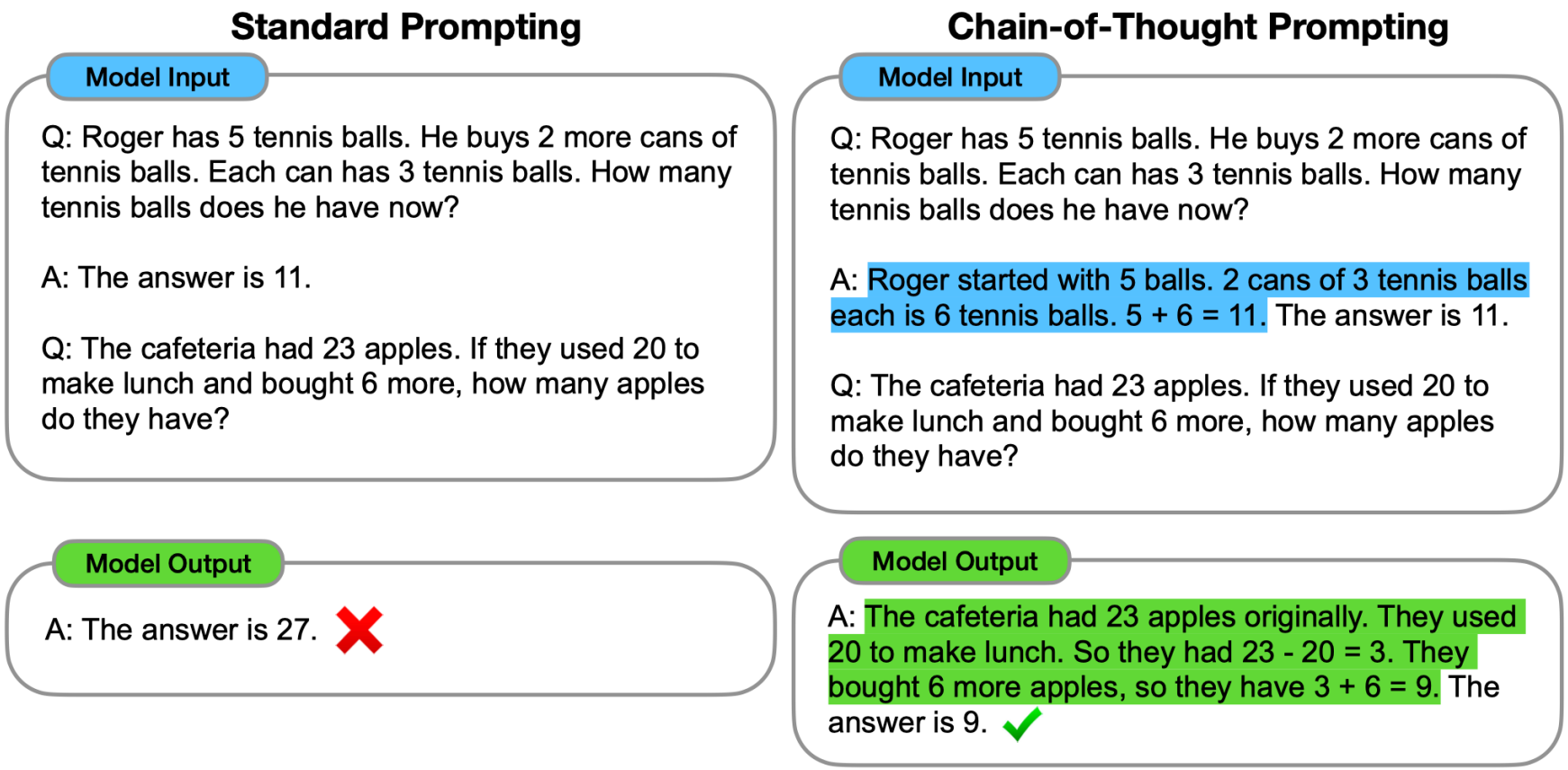


Figure 5: **Zero-shot performance on Quiet-STaR applied to chain-of-thought on GSM8K.** We visualize how using a Quiet-STaR trained Mistral model can improve chain-of-thought performance. We use an 8-thought-token-trained model and use its internal thoughts to improve the tokens in a zero-shot chain-of-thought (Kojima et al., 2022)

思维连 Chain-of-Thought

1. COT通过要求模型在输出最终答案之前，显式输出中间逐步的推理步骤这一方法来增强大模型的算数、常识和推理能力。



04

产业思考与洞察

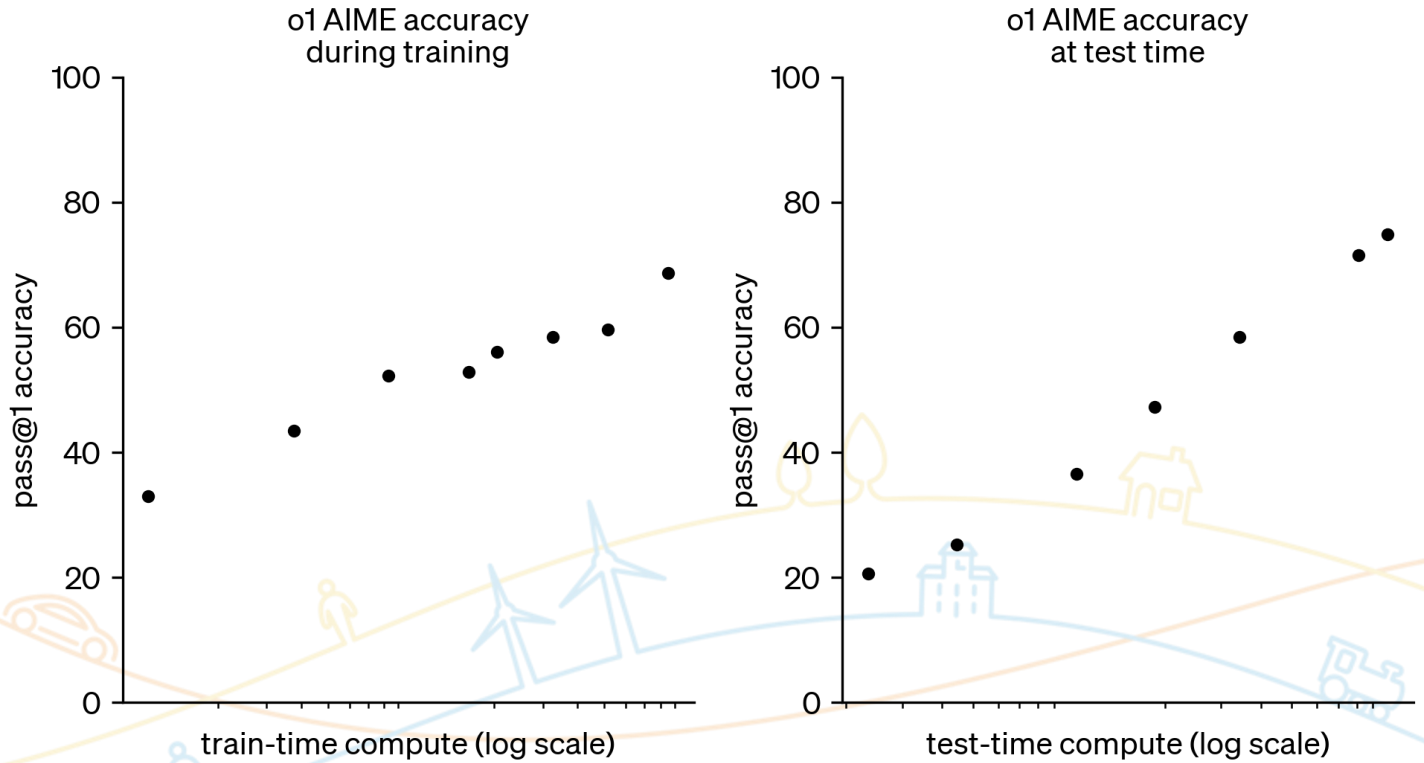
与 OpenAI 的差距

1. OpenAI 训练出 GPT4，全世界在挖 OpenAI 训练 LLM，希望追平 GPT4；
2. OpenAI 已在数据工程积累颇深发出 GPT-4o，全世界在思考如何做数据工程；
3. OpenAI 利用强化学习发明 o1 模型，全世界肯定开始挖掘 RL 算法潜力；

- OpenAI 永远在技术的制高点！

对算力的消耗的预测

1. Reinforcement Learning start Scaling Law !



算法技术路线的选择

1. 跟随派：Reinforcement Learning start Scaling Law ! MoE 都不重要!
2. 新路派：RL + CoT 不是 AGI 的未来！ mamba、 RKWV 结构才是未来!

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

Copyright © 2023 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/AIFoundation