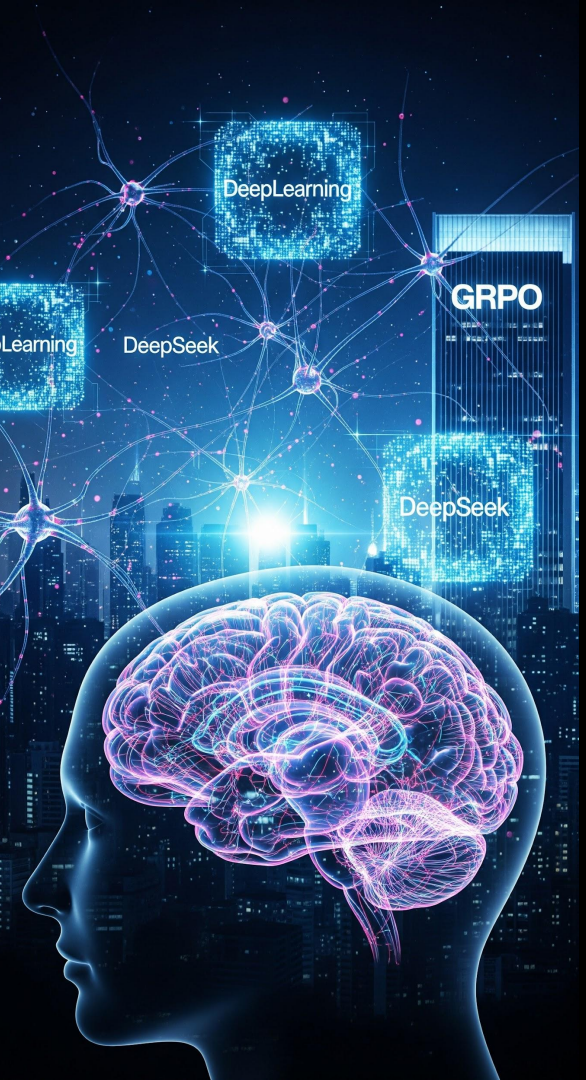


DeepSeek & GRPO

Ran Tavory 2025

TII

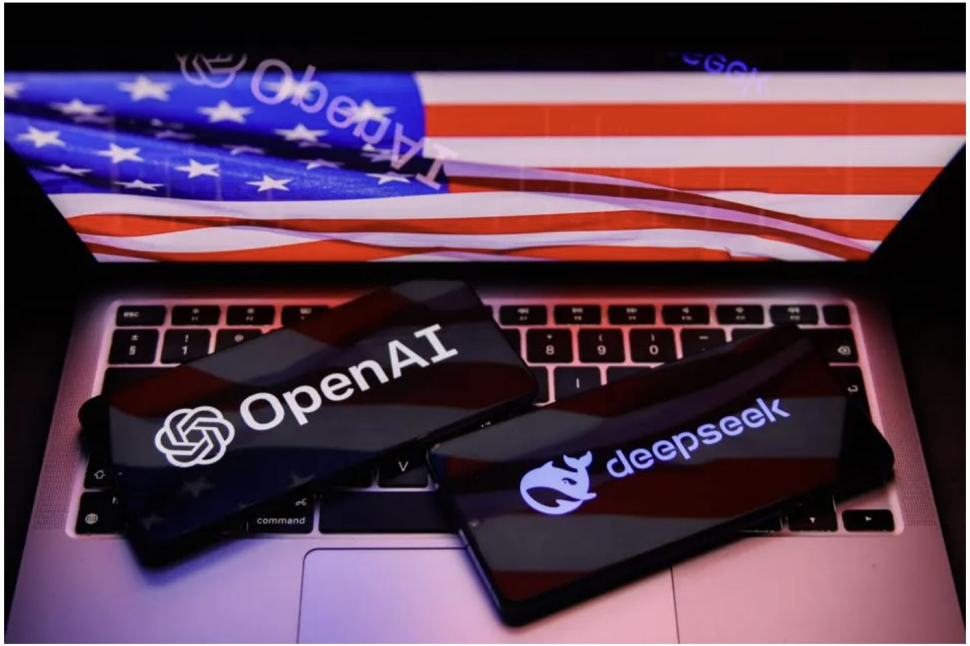
Forbes

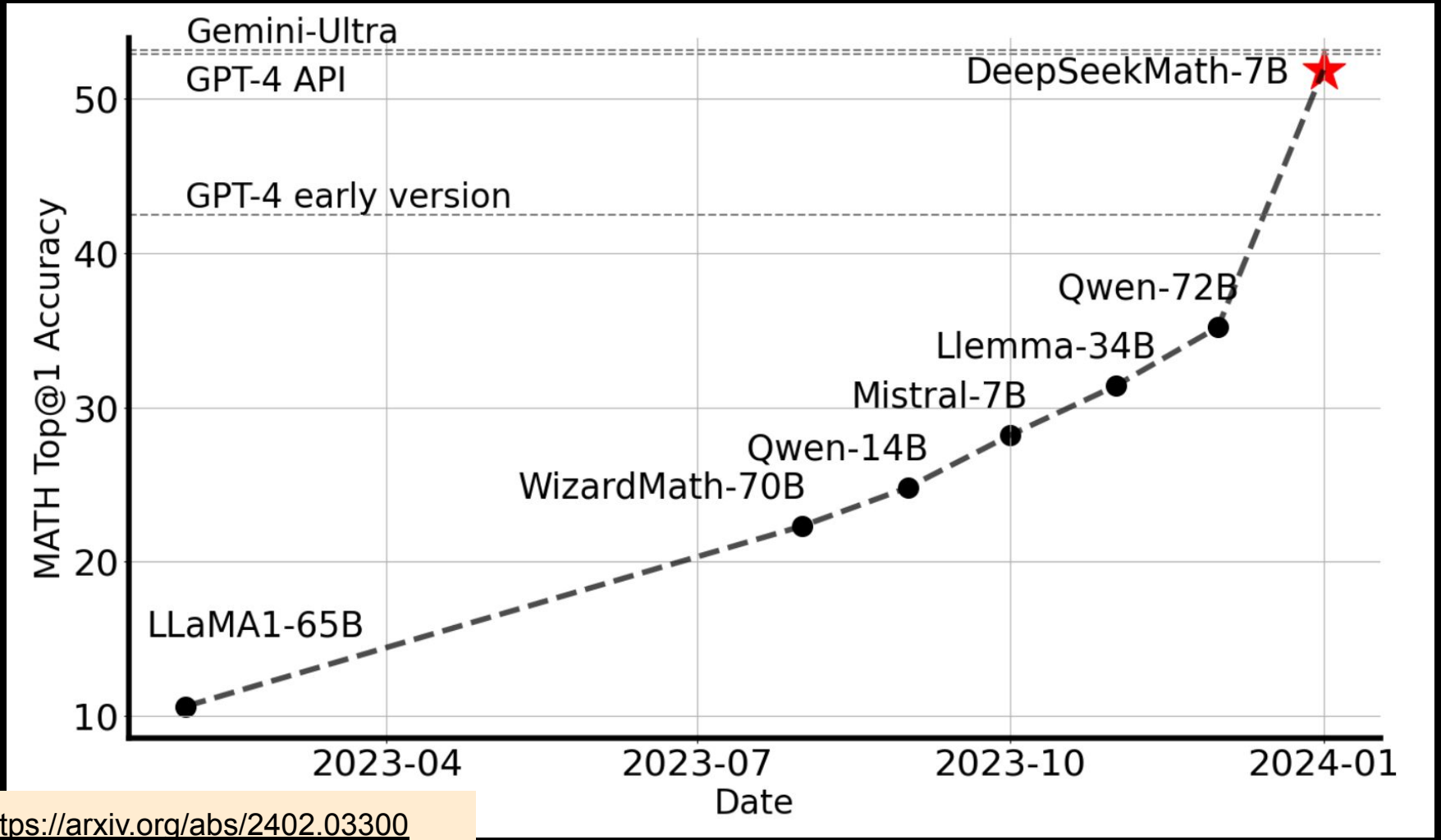
The $6 Million AI Bombshell: How DeepSeek Shook Wall Street And AI Leadership

By Mark Minevich, Contributor. Mark Minevich is a NY-based strategist foc... Follow Author

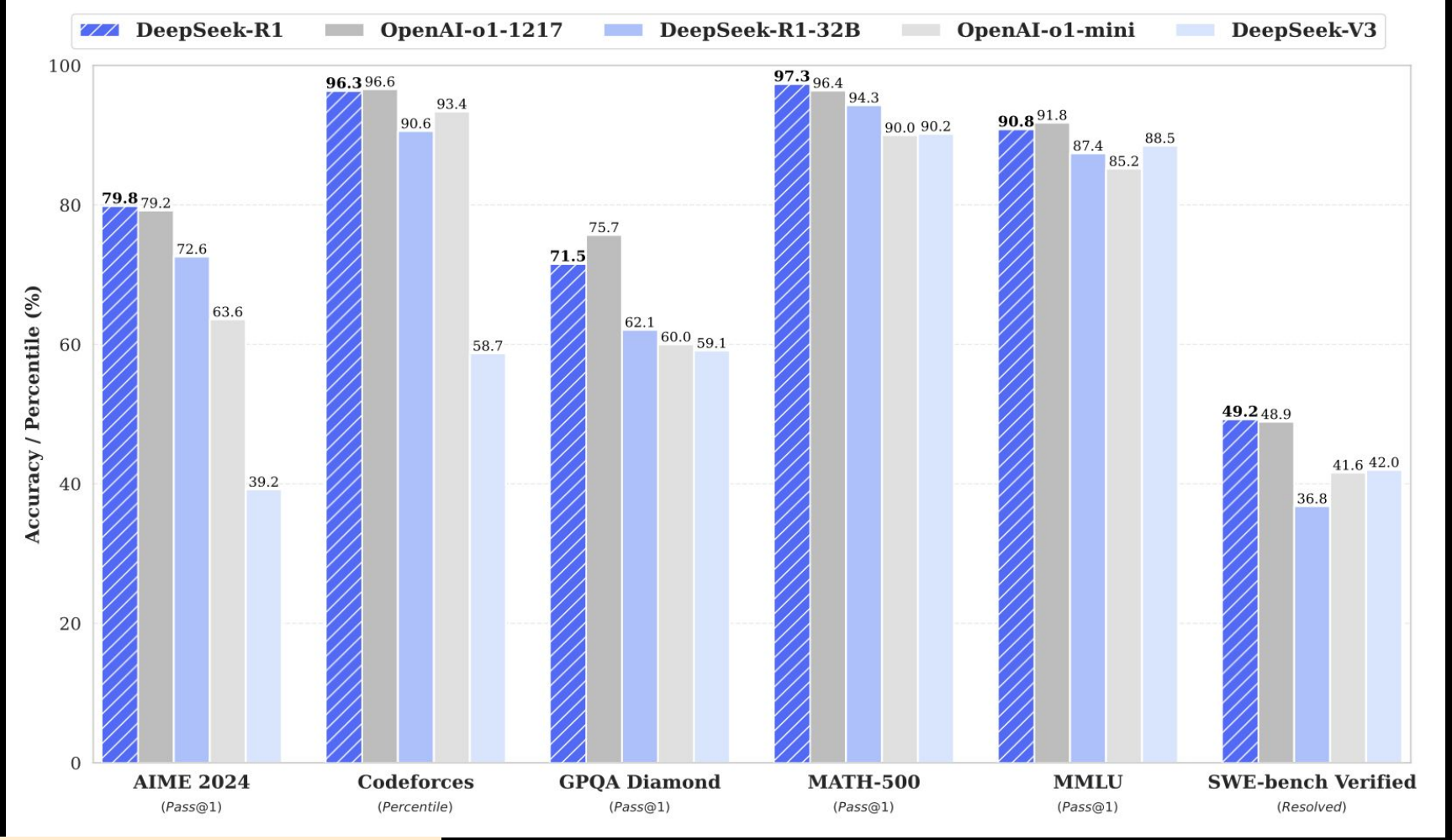
Feb 06, 2025, 01:41pm EST

Share Save Comment 0





https://arxiv.org/abs/2402.03300

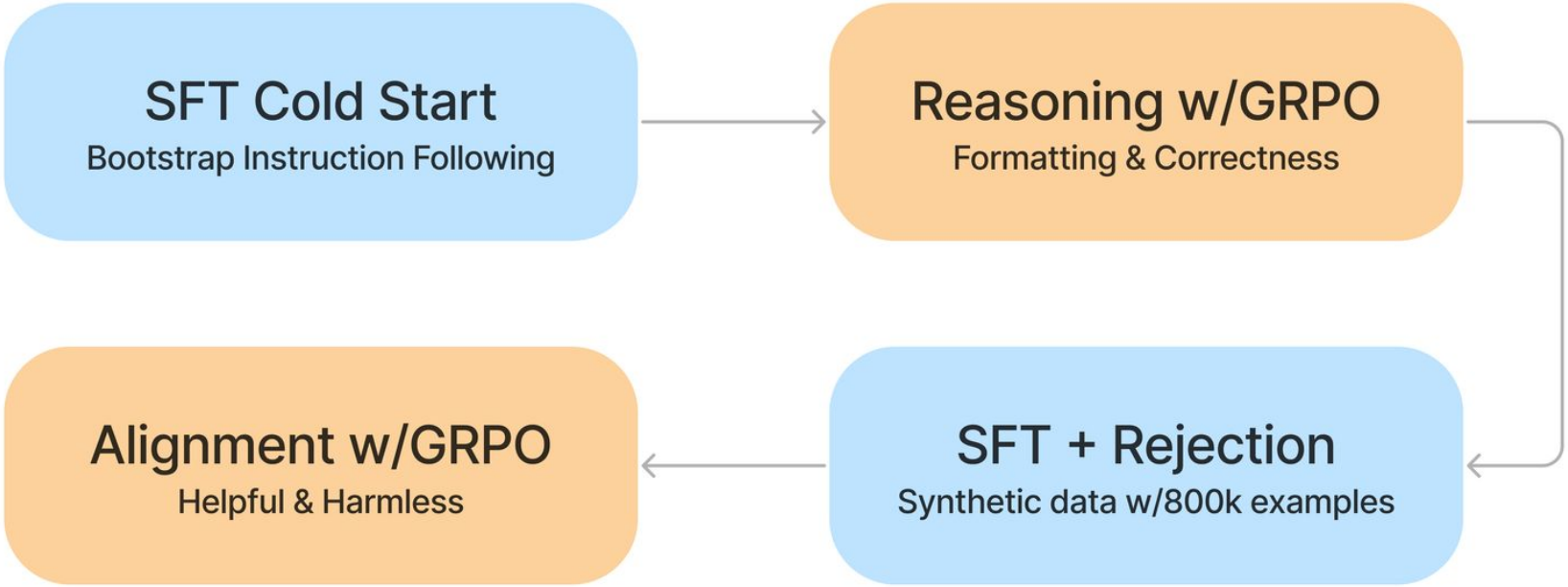


https://arxiv.org/abs/2501.12948

02

What does GRPO has to do with it

DeepSeek-R1 Training Pipeline



https://ghost.oxen.ai/why-grpo-is-important-and-how-it-works/

RLHF

Reinforcement Learning from Human Feedback

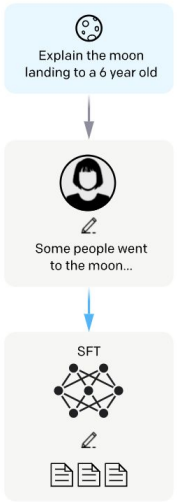
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



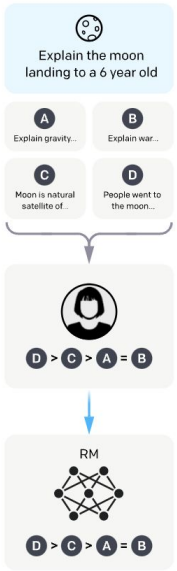
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

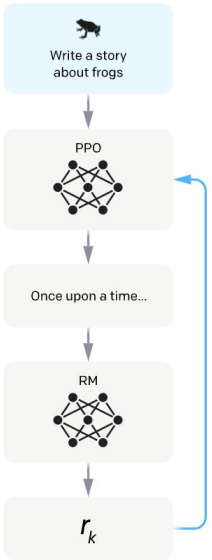


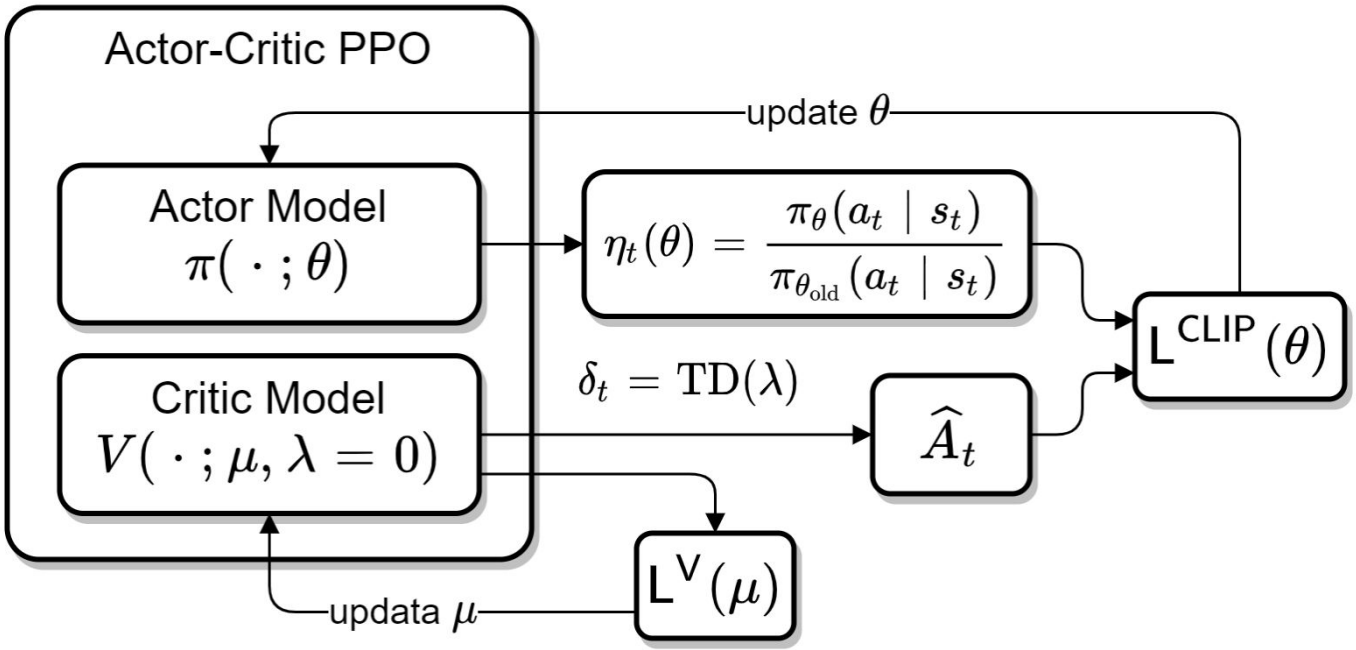
Image by OpenAI

03

PPO first

PPO

Proximal Policy Optimization



Reminder

Step 3

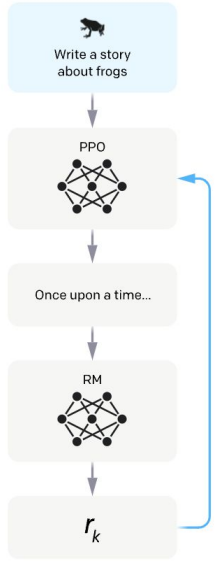
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

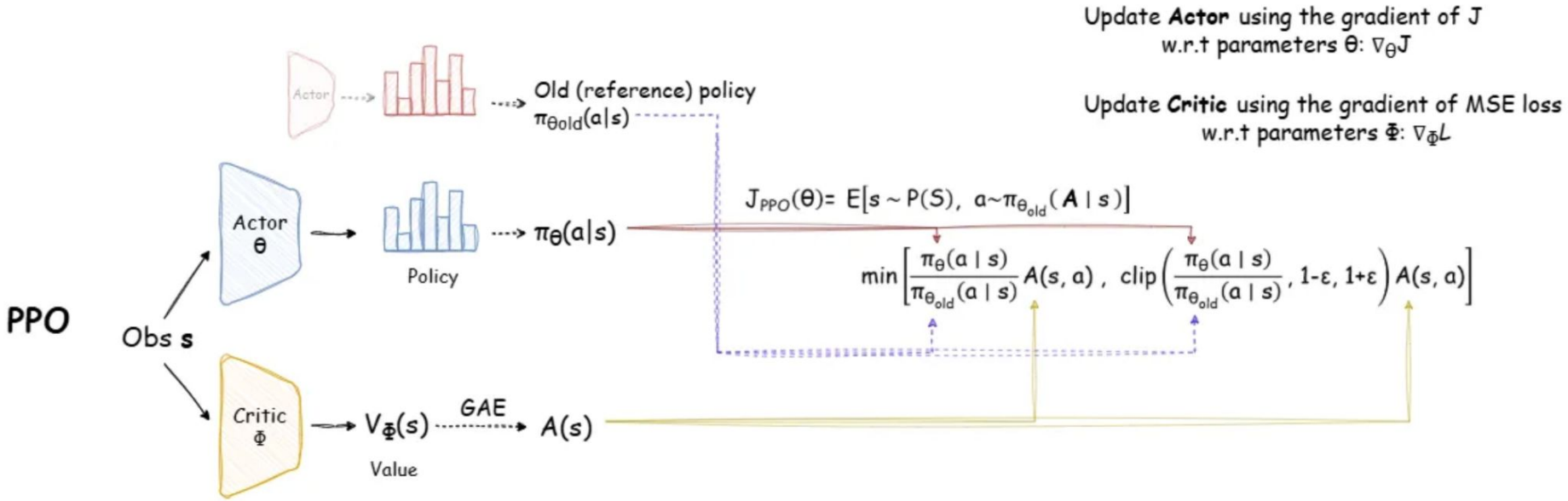
The reward is used to update the policy using PPO.



https://arxiv.org/pdf/2302.13877 and https://arxiv.org/abs/1707.06347

PPO

Another visual representation



https://pub.towardsai.net/group-relative-policy-optimization-grpo-illustrated-breakdown-explanation-684e71b8a3f2

PPO Summary

- Effective and battle tested
- But requires a lot of memory:
  - Two trainable models (2x4 params)
  - Two frozen models (2x params)
  - => ~10x LLM memory
- Sample efficient.
  - Useful when samples are expensive
  - Example: Lunar lander

DPO

Direct Preference Optimization

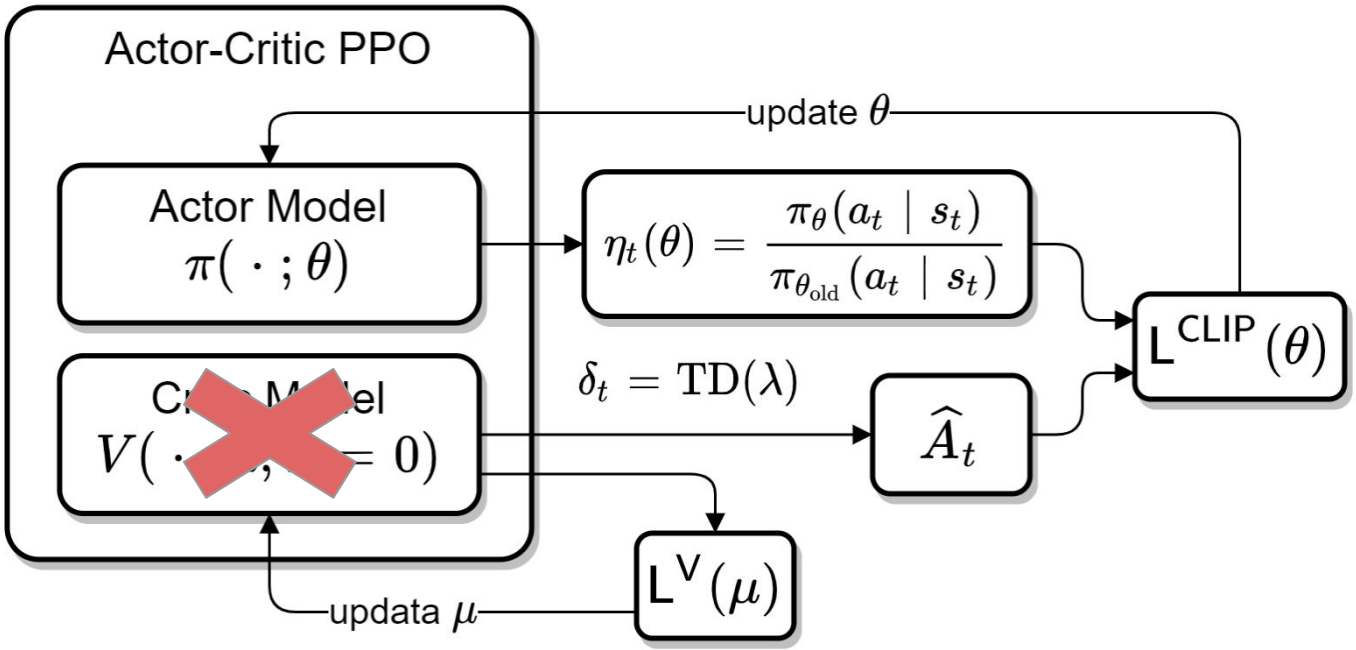
Out of scope for today

04

GRPO

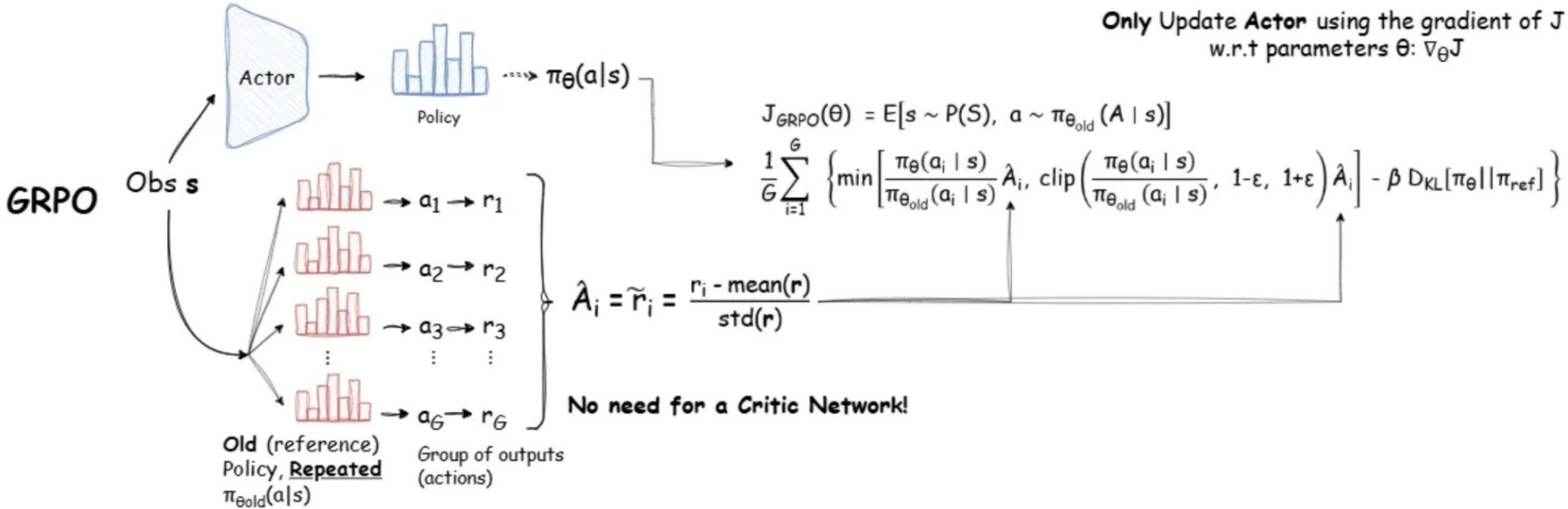
GRPO

Group Relative Policy Optimization



GRPO

Group Relative Policy Optimization



GRPO

Key points

- No need for a Value model (which is 4xLLM)
- The Advantage is calculated for each result relative to all other members of its group (hence the name Group Relative)

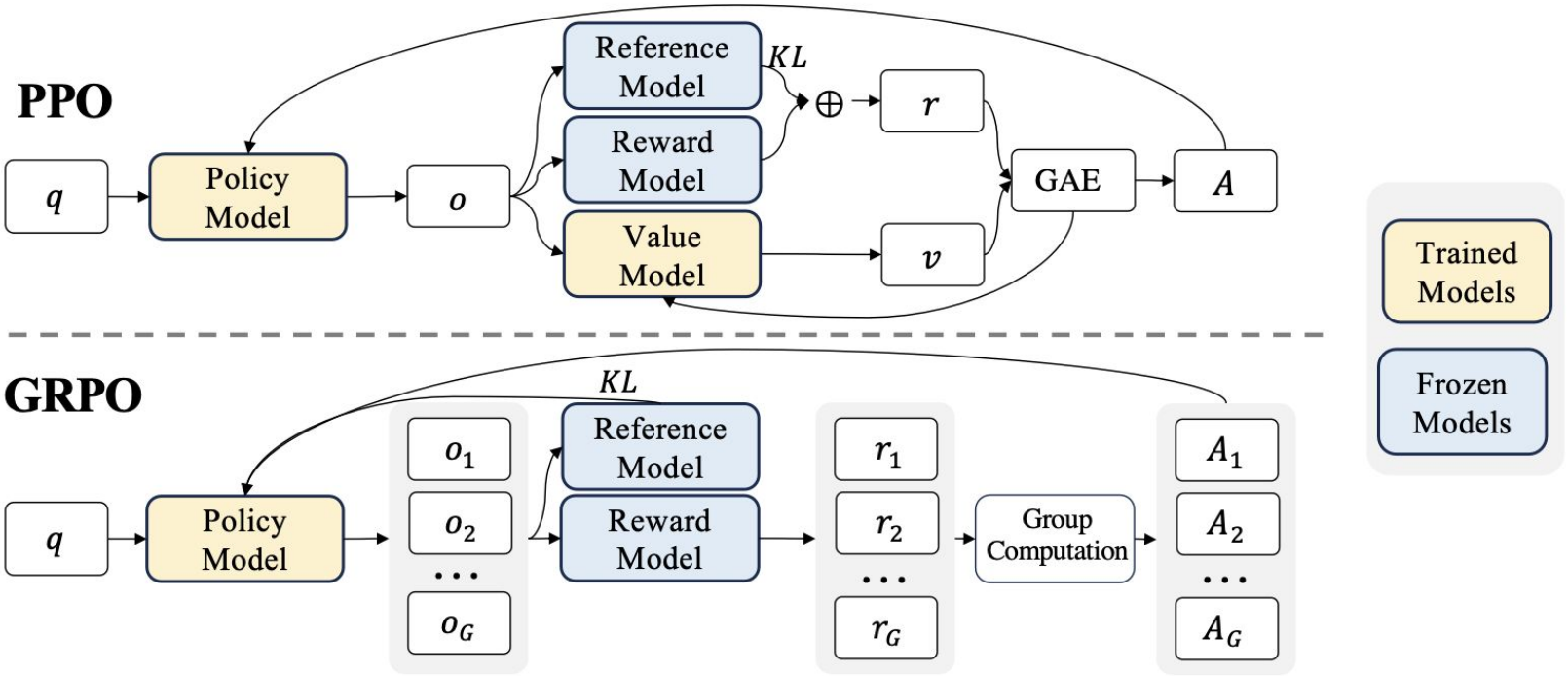
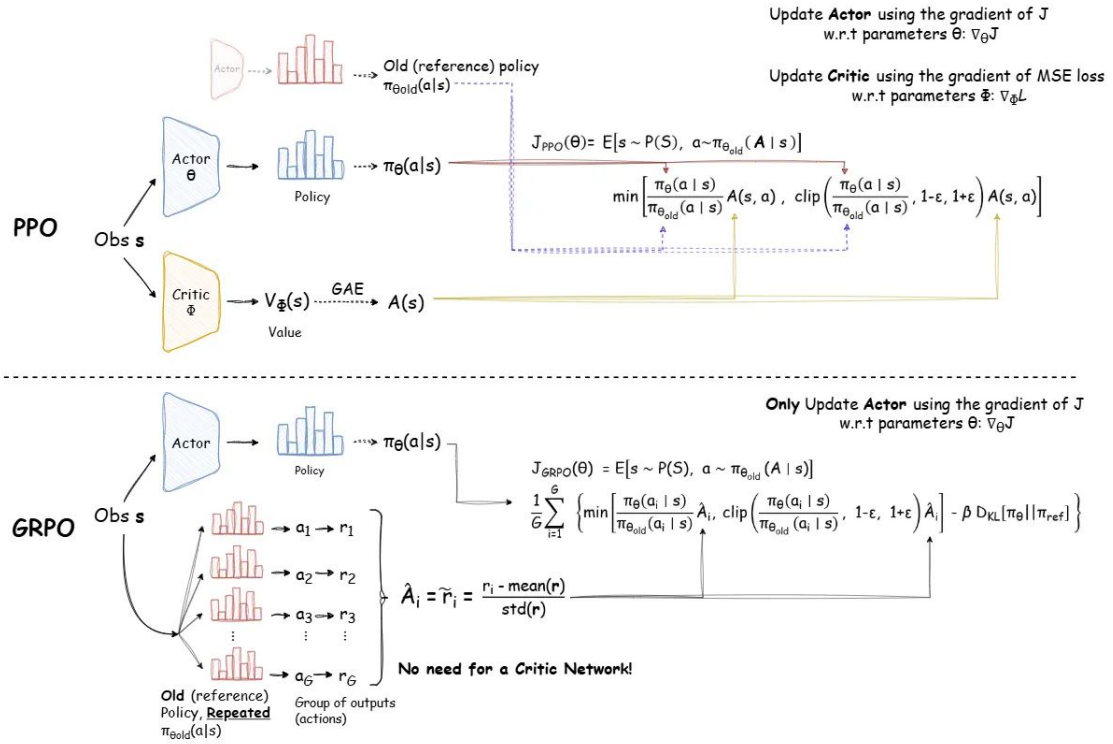


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

https://arxiv.org/abs/2402.03300

GRPO vs PPO


Update Actor using the gradient of J
w.r.t parameters θ: $\nabla_\theta J$
Update Critic using the gradient of MSE loss
w.r.t parameters Φ: $\nabla_\Phi L$
Actor
Old (reference) policy
$\pi_{\theta_{old}}(a|s)$
Actor θ
$\pi_\theta(a|s)$
Policy
$J_{PPO}(\theta) = E[s \sim P(S),\ a \sim \pi_{\theta_{old}}(A \mid s)]$
$\min\left[\frac{\pi_\theta(a \mid s)}{\pi_{\theta_{old}}(a \mid s)} A(s,a),\ clip\left(\frac{\pi_\theta(a \mid s)}{\pi_{\theta_{old}}(a \mid s)}, 1-\varepsilon, 1+\varepsilon\right) A(s,a)\right]$
PPO
Obs s
Critic Φ
$V_\Phi(s)$
GAE
$A(s)$
Value
Only Update Actor using the gradient of J
w.r.t parameters θ: $\nabla_\theta J$
Actor
$\pi_\theta(a|s)$
Policy
$J_{GRPO}(\theta) = E[s \sim P(S),\ a \sim \pi_{\theta_{old}}(A \mid s)]$
$\frac{1}{G}\sum_{i=1}^{G}\left\{\min\left[\frac{\pi_\theta(a_i \mid s)}{\pi_{\theta_{old}}(a_i \mid s)}\hat{A}_i,\ clip\left(\frac{\pi_\theta(a_i \mid s)}{\pi_{\theta_{old}}(a_i \mid s)}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_i\right] - \beta\, D_{KL}[\pi_\theta||\pi_{ref}]\right\}$
GRPO
Obs s
$a_1 \to r_1$
$a_2 \to r_2$
$a_3 \to r_3$
$a_G \to r_G$
$\hat{A}_i = \tilde{r}_i = \frac{r_i - mean(r)}{std(r)}$
No need for a Critic Network!
Old (reference) Policy, Repeated $\pi_{\theta_{old}}(a|s)$
Group of outputs (actions)

RL Terms

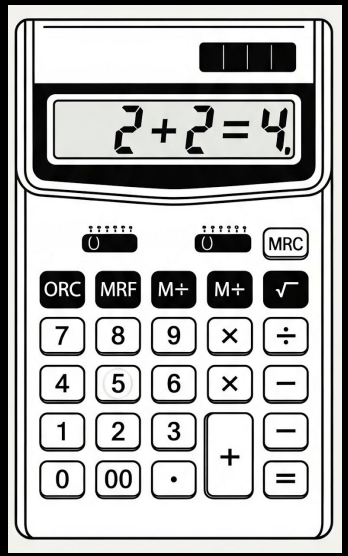
Common RL terms

- **policy**: in deep RL, the policy is the neural network that looks at the observations from the environment and predicts the action to be taken. this is what we are training.
- **environment**: the world that the policy interacts with, which could be anything from a video game to a robotic simulation to a real-world setting. the environment provides observations and rewards.
- **reward**: a numerical value that tells the agent how good or bad its last action or sequence of actions was.
- **trajectory**: the sequence of states, actions, and rewards that occur during an episode or training run.

How RL maps to LLMs

Mapping RL concepts to Text Generation

- RL episode == LLM full generation
- RL time step == LLM single token generation
- RL reward == LLM Reward model or RLVR
- so you judge the entire response as one, and define a reward value based on that judgement
- Rewards are sparse in LLMs. Only one reward at the end of the episode (generation).

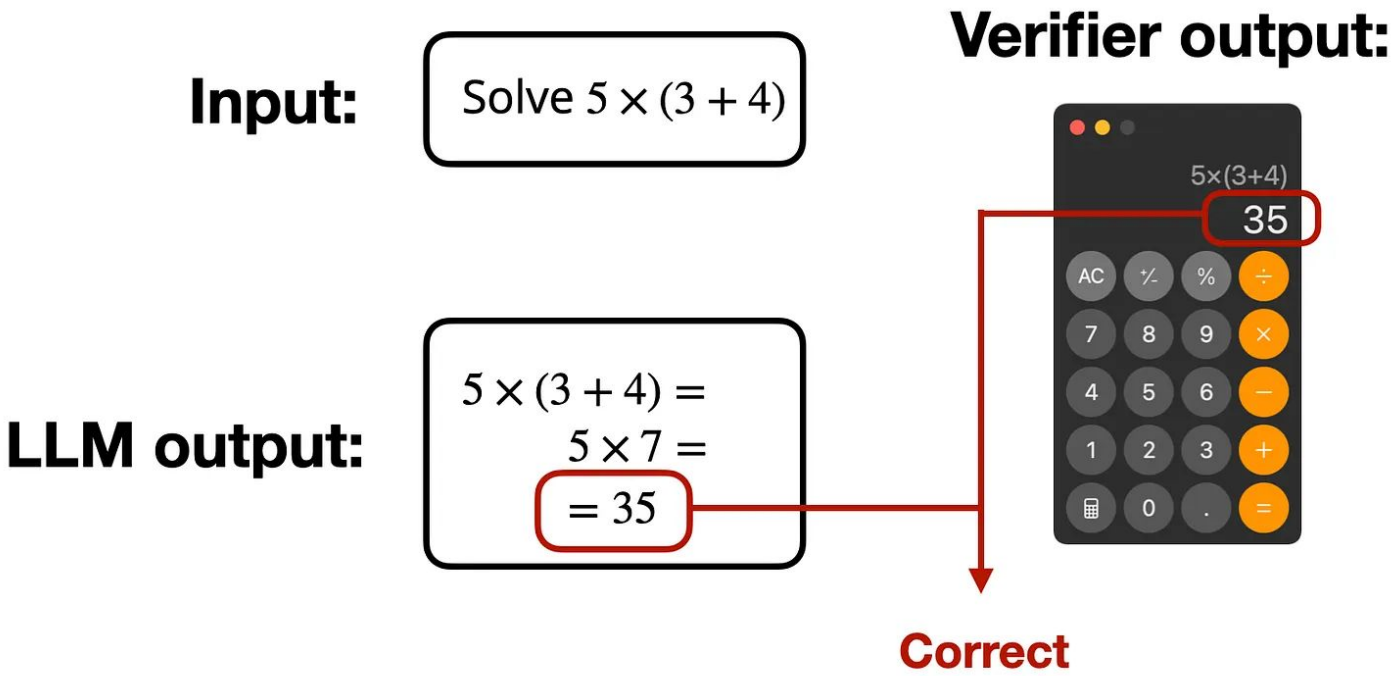


05

RLVR

RLVR

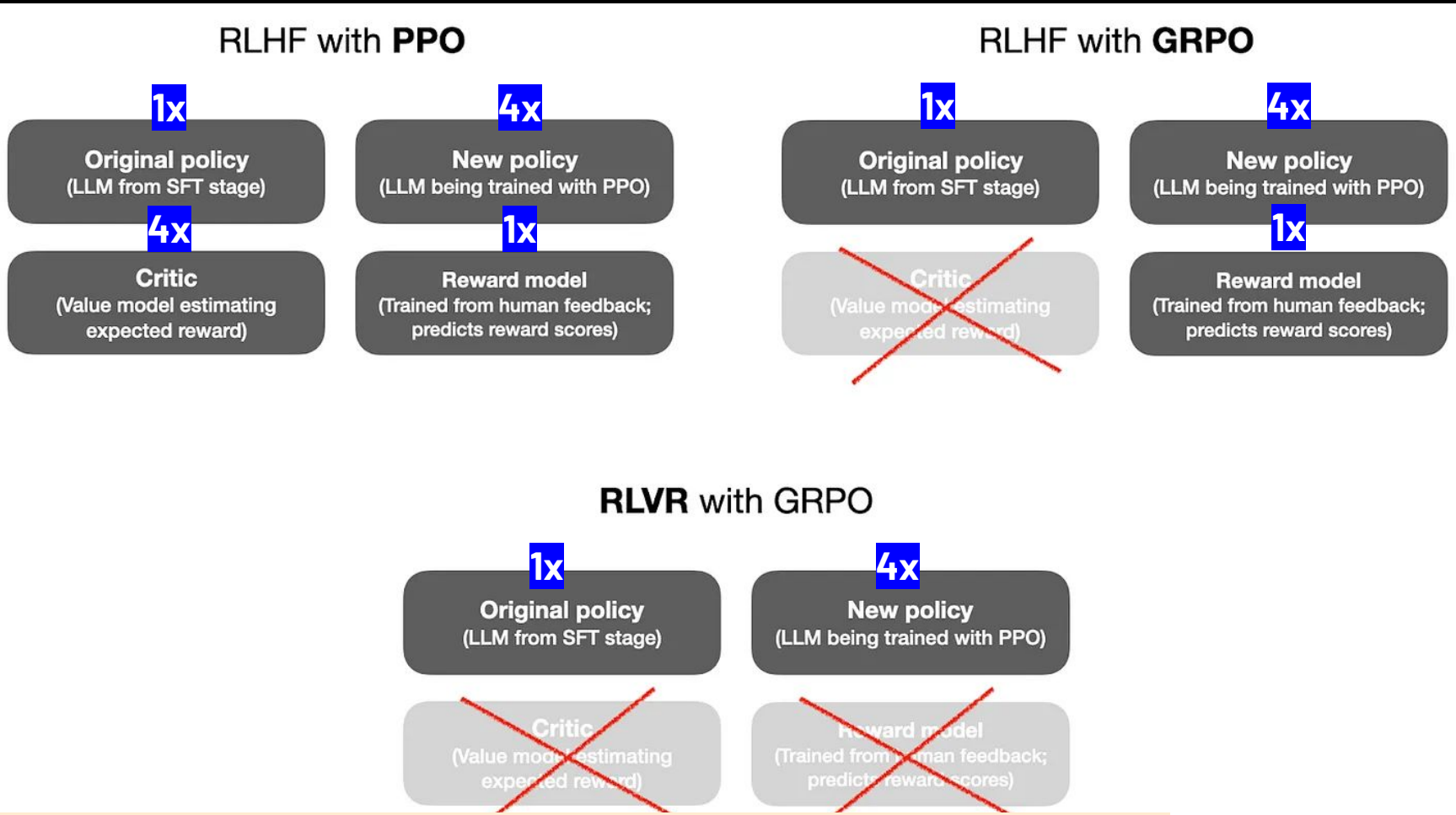
Reinforcement Learning from Verifiable Rewards



RLVR & RLHF

Comparison

RLHF	• **Requires a reward model (costs more)** • **Reward model is trained from real humans feedback** • **Susceptible to Reward Hacking**
RLVR	• **Rewards are provided by “simple” tools. E.g. a calculator or a compiler** • **Requires structured output**



https://magazine.sebastianraschka.com/p/the-state-of-llm-reasoning-model-training

O^*

Math

The Math

Loss functions

PPO's clipped surrogate loss function

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right]$$

GRPO's loss function

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left(r_t(\theta)\hat{A}_{i,t}, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t} \right) \right] - \beta D_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}).$$

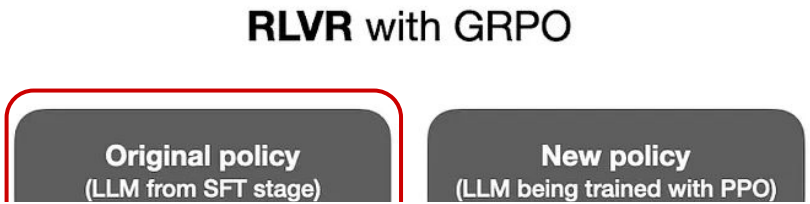
O^*

KL Divergence and Reference Policy

Why do we need the Original model?

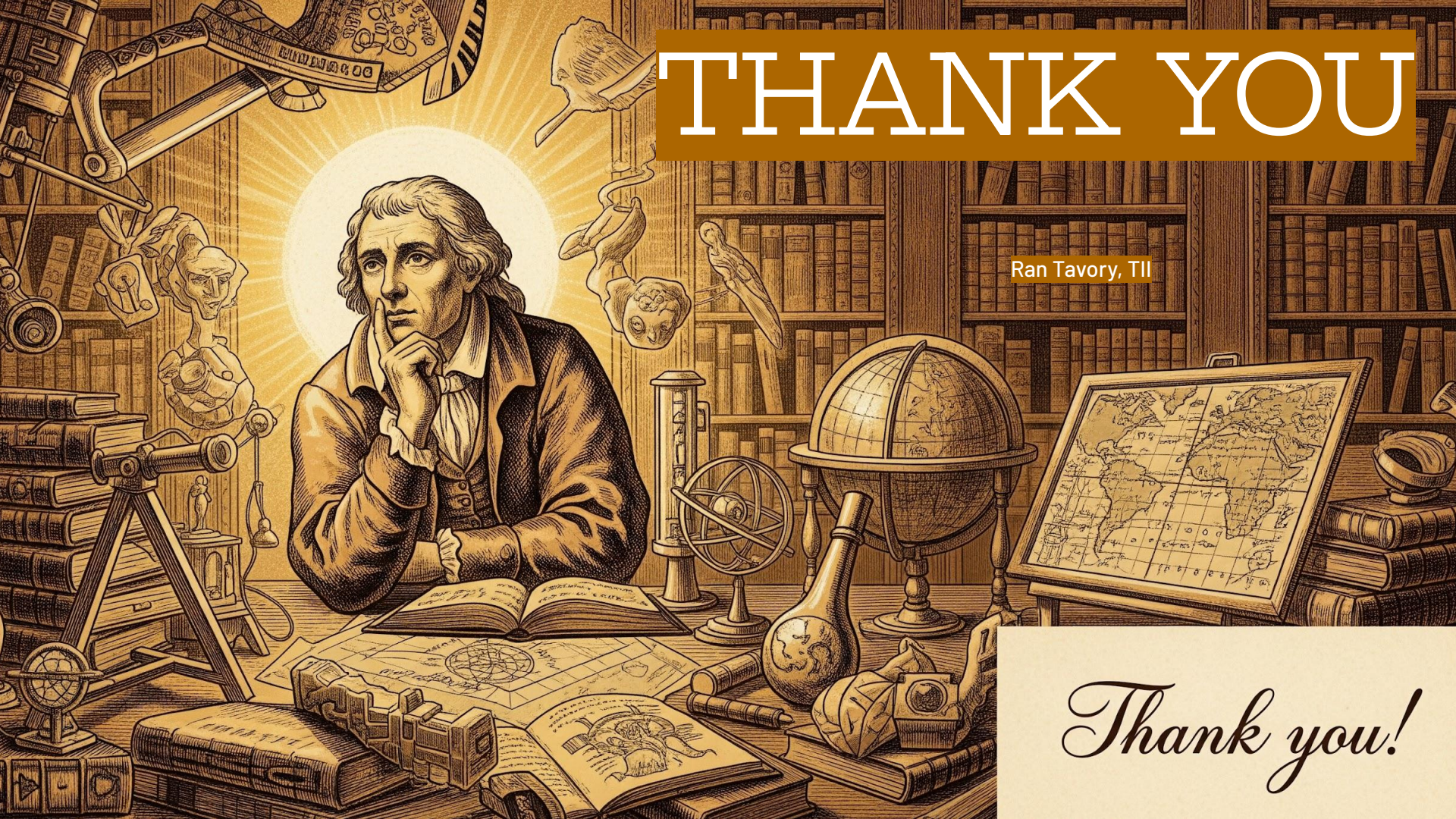
Answer:

- In PPO/GRPO the original model is used in order to make sure that the new model remains "in the same neighbourhood"



$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\},$$


THANK YOU
Ran Tavory, TII
Thank you!