Post Training
Reward Models, DPO, Constitutional AI

Gheorghe Comanici

Today’s Agenda on Aligning LLMs

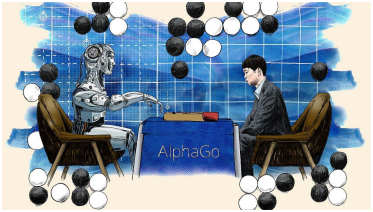
Monday

- **The Why**
  a. Exploring conceptual alignment, responsible AGI, and reward challenges.
- **The How**
  a. Introducing technical alignment methods (e.g., InstructGPT).
- **The Engine**
  a. Linking LLMs with Reinforcement Learning (RL) and its data needs.
  b. Covering RL fundamentals (core loop, values, policies, PPO).

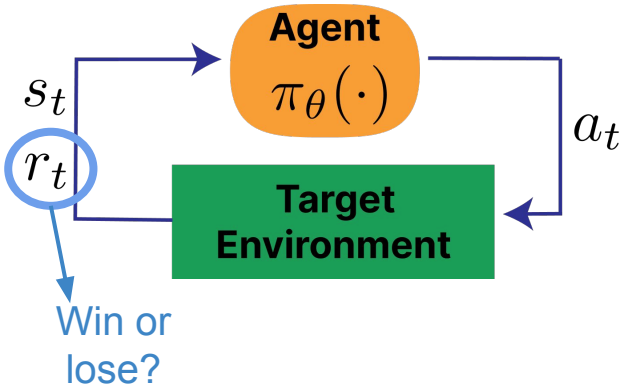
Today

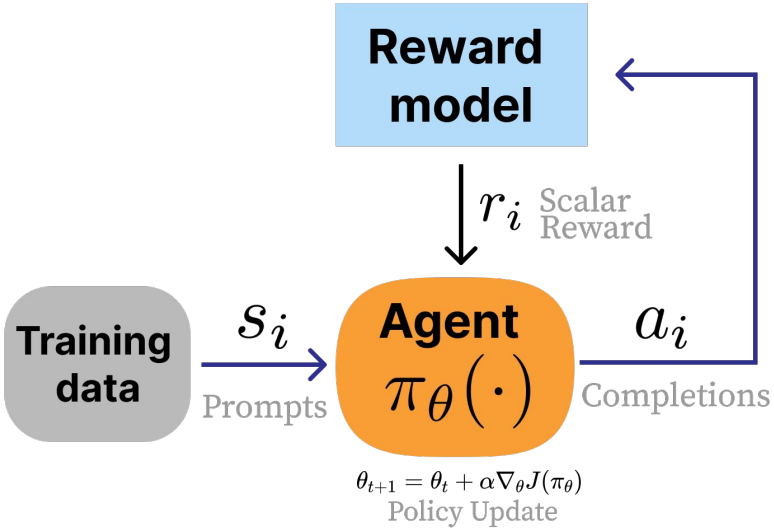
- **The Reward Model**
  c. Detailing reward function design, sources, risks, and training.
- **The Gradient**
  a. TRPO / PPO / DPO
- **Closing the loop**
  a. Constitutional AI, RLAIF, and Generative Reward Models

The Guidance: Reward *models*



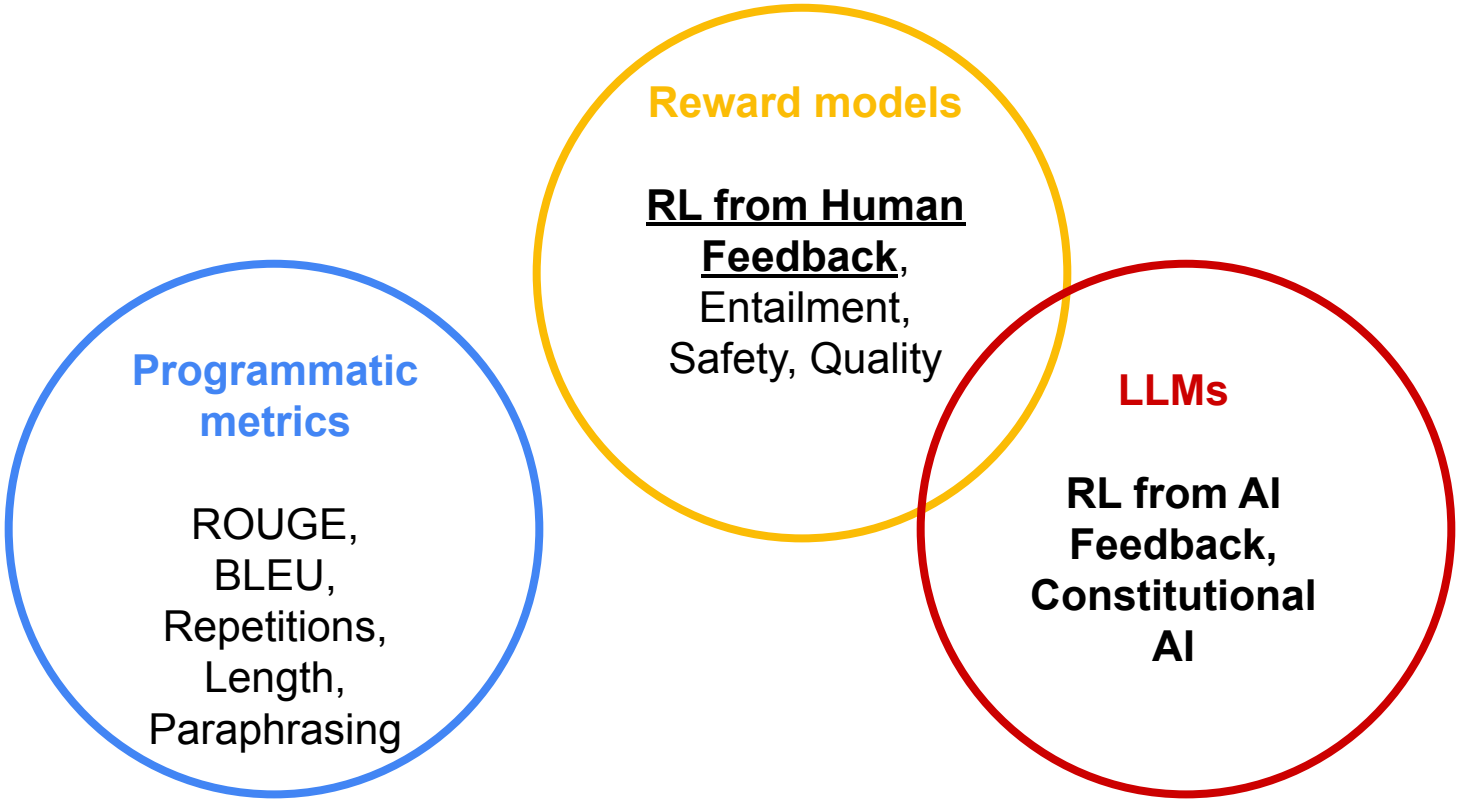
vs.



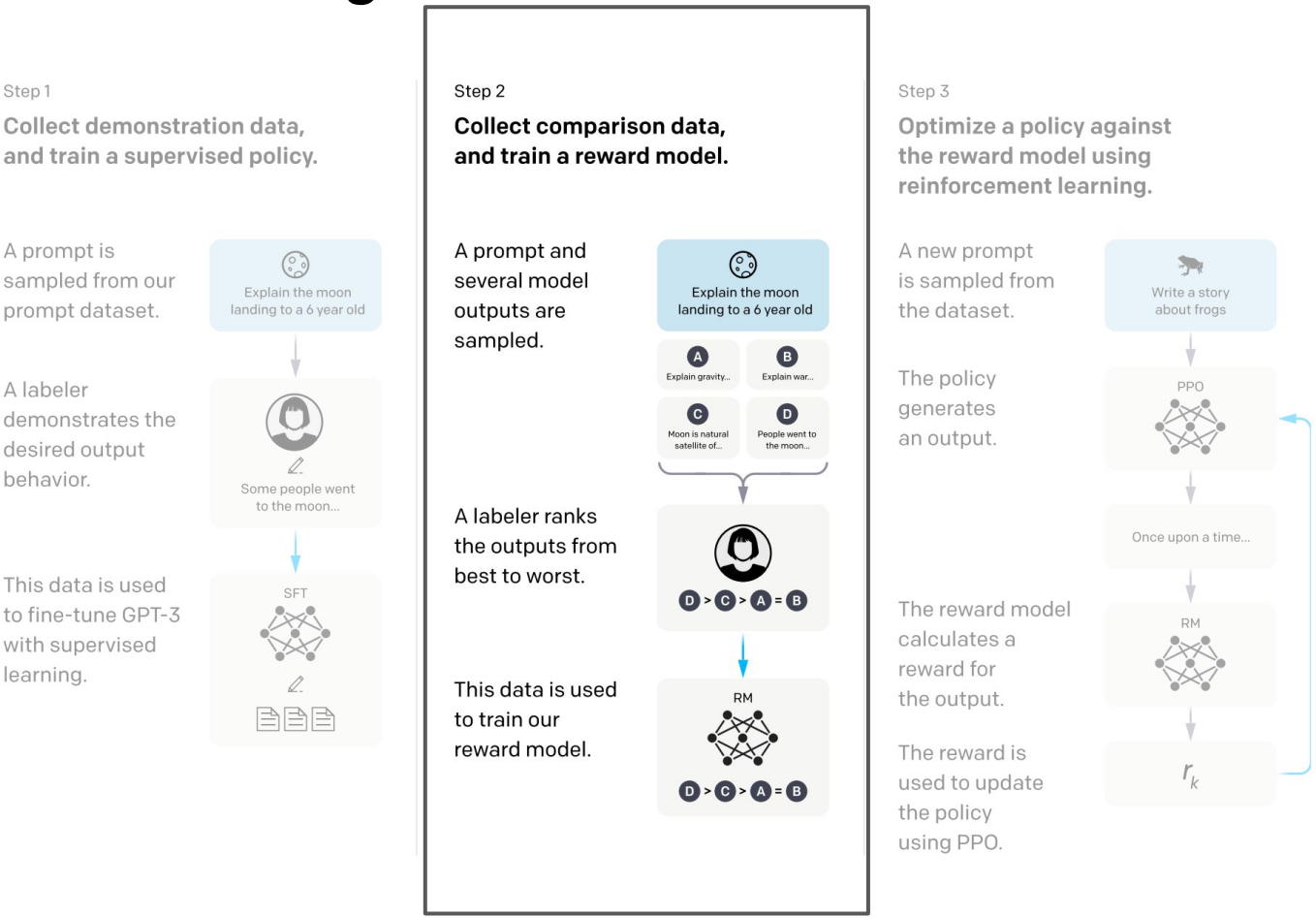


Source: Lambert 2024

The Guidance: Reward *models and more…*



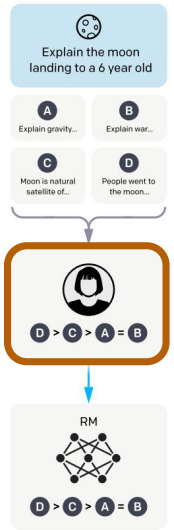
Reward *modelling*



Bradley Terry model

Human feedback

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.

Preference (explicit) modelling

Bradley Terry model assumes probability of preference from a point-wise **score** function

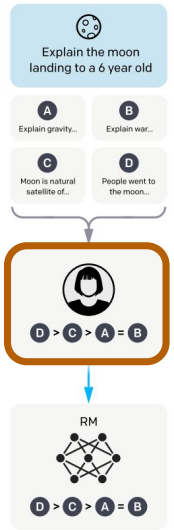
$$P(i > j) = \frac{p_i}{p_i + p_j}$$

If *i>j* 7 times, *j<i* 3 times,
P(i > j) = 7 / 10

Bradley Terry model

Human feedback

A prompt and several model outputs are sampled.



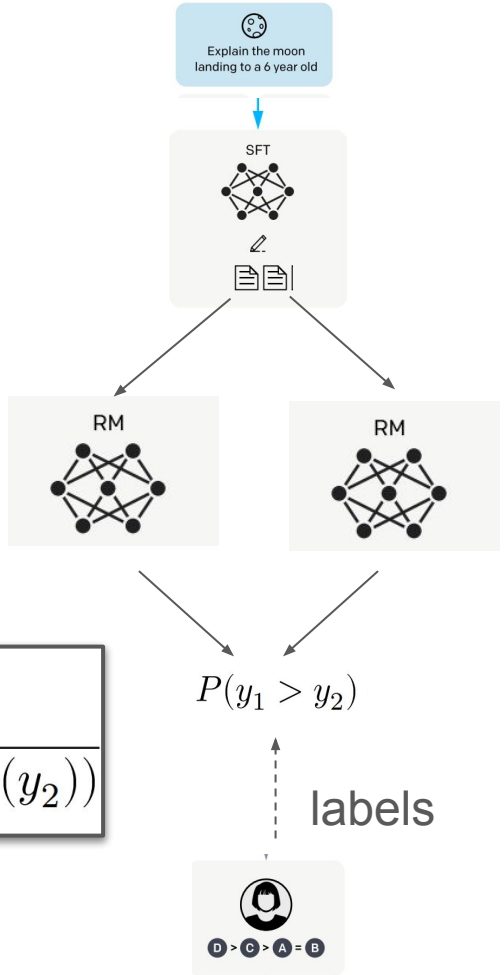
A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Preference (explicit) modelling

Bradley Terry model assumes probability of preference from a point-wise **score** function

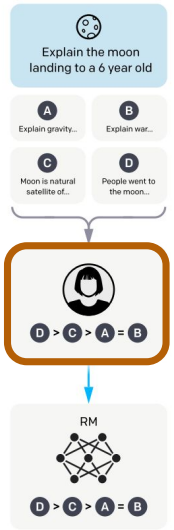
$$P(y_1 > y_2) = \frac{\exp(r(y_1))}{\exp(r(y_1)) + \exp(r(y_2))}$$



Bradley Terry model

Human feedback

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.

Equivalent

Reward (explicit) modelling

$$\mathcal{L}(\theta) = -\log\left(\sigma\left(r_\theta(x, y_w) - r_\theta(x, y_l)\right)\right)$$

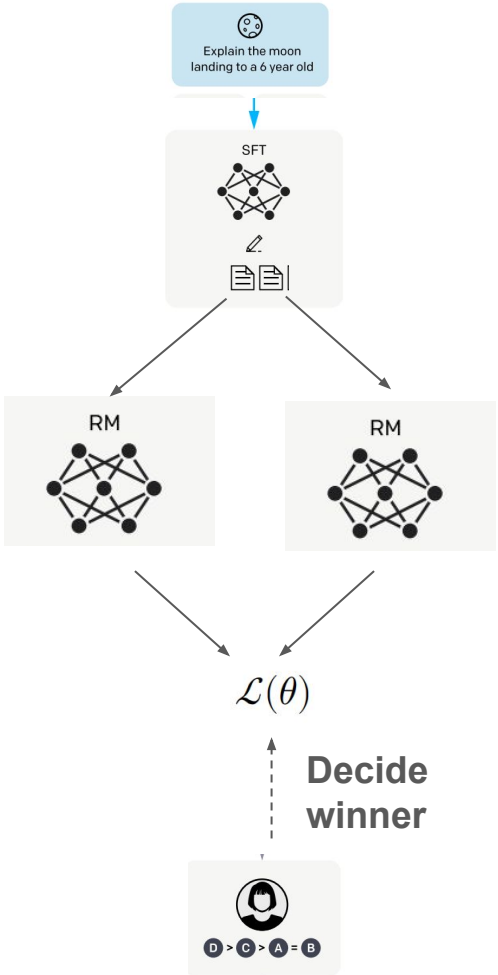
y_w = "winning" output

y_l = "losing" output

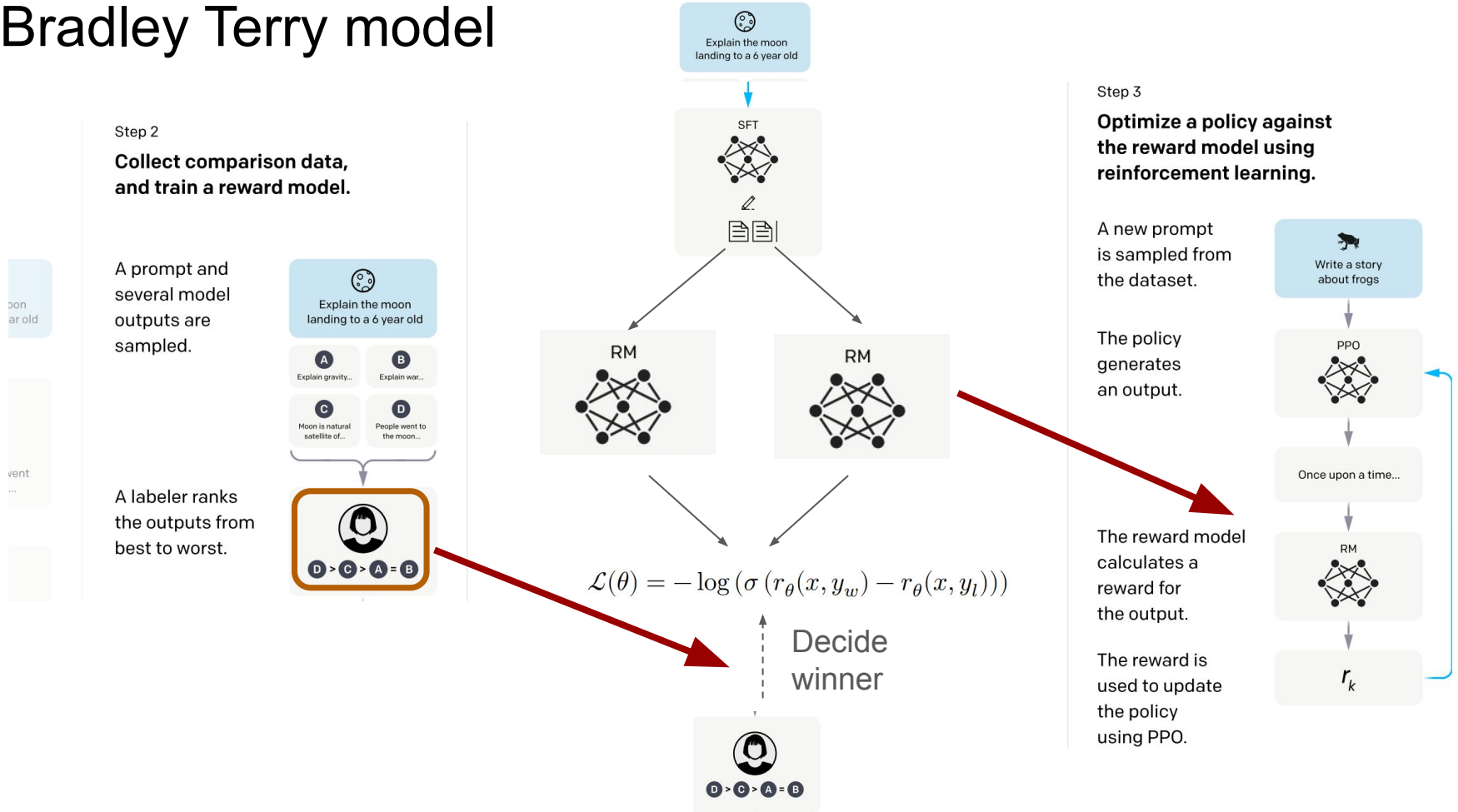
σ = sigmoid activation function

$$P(y_1 > y_2) = \frac{\exp(r(y_1))}{\exp(r(y_1)) + \exp(r(y_2))}$$

If *i>j* 7 times, *j<i* 3 times,
P(i > j) = 7 / 10



Bradley Terry model

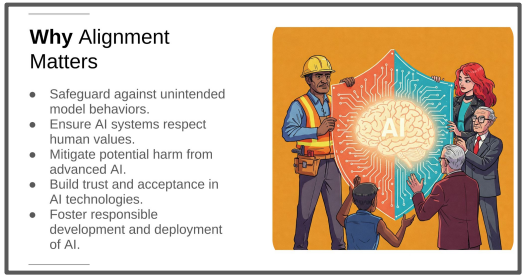


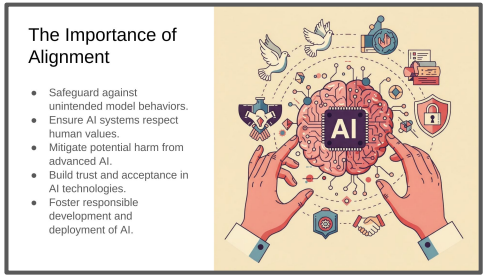
~~Are these “well aligned”?~~

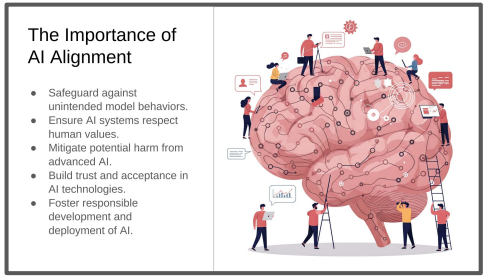
Put these in order from least to most “aligned”!

User: Create a slide that explains why Alignment matters.

Gemini:

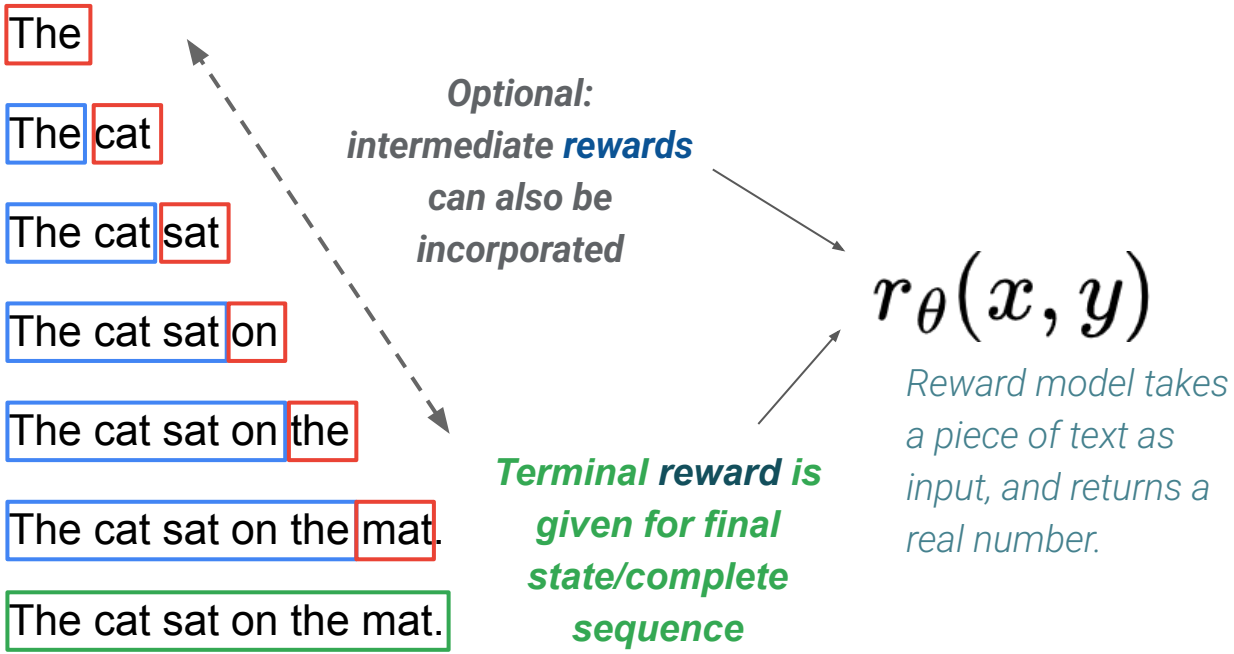






RM: 0.95 0.85 0.65

Reward model architecture



Reward model architecture

Trick: *repurpose any existing LLM model*

$$r_\theta(x, y)$$

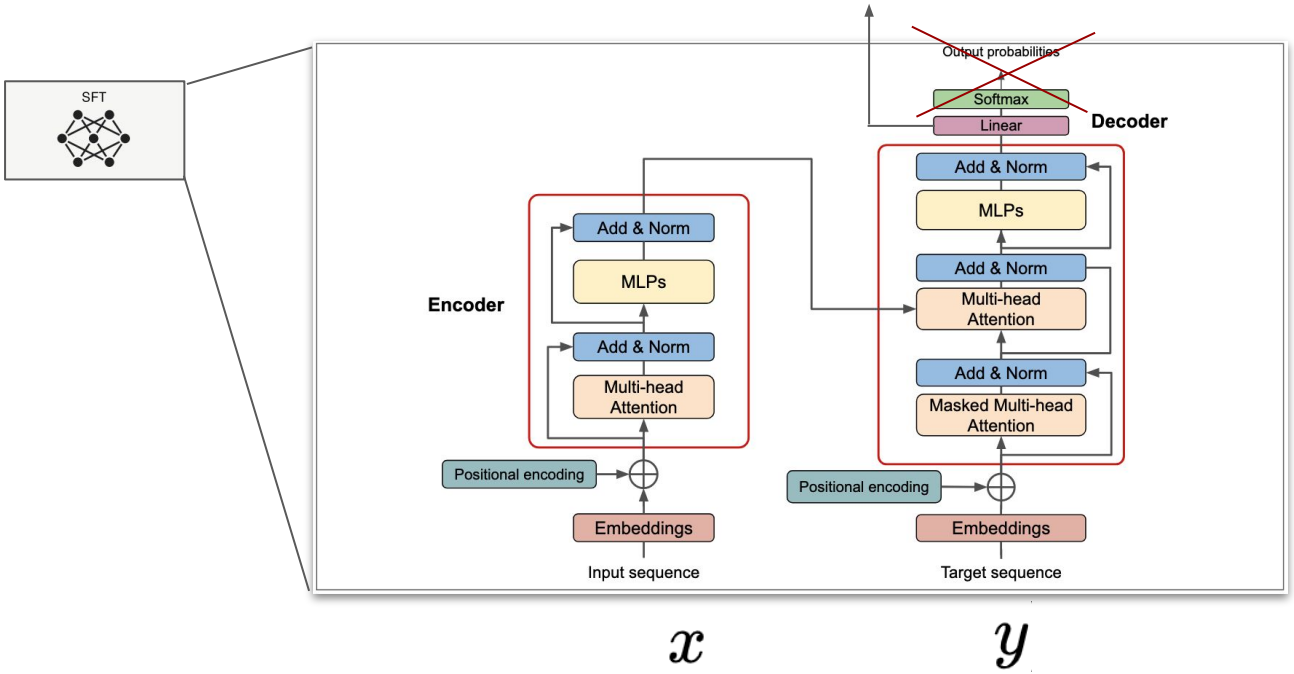


Diagram from: https://re-cinq.com/blog/llm-architectures

Can we finally do some RL?

Not yet…recall that we want to maximize reward while maintaining proper natural language understanding. I.e. *no human data left behind.*

$$(1-\alpha)\nabla v_\pi(s) - \alpha\mathbb{E}_{s\in\rho_\pi}[\nabla KL(\pi(\cdot|s))||\pi_{SFT}(\cdot|s))]$$

Tradeoff parameter

Stay close to supervised model

Can we finally do some RL?

Not yet…recall that we want to maximize reward while maintaining proper natural language understanding. I.e. *no human data left behind.*

$$(1-\alpha)\nabla v_\pi(s) - \alpha\mathbb{E}_{s\in\rho_\pi}[\nabla KL(\pi(\cdot|s))||\pi_{SFT}(\cdot|s))]$$

Tradeoff parameter

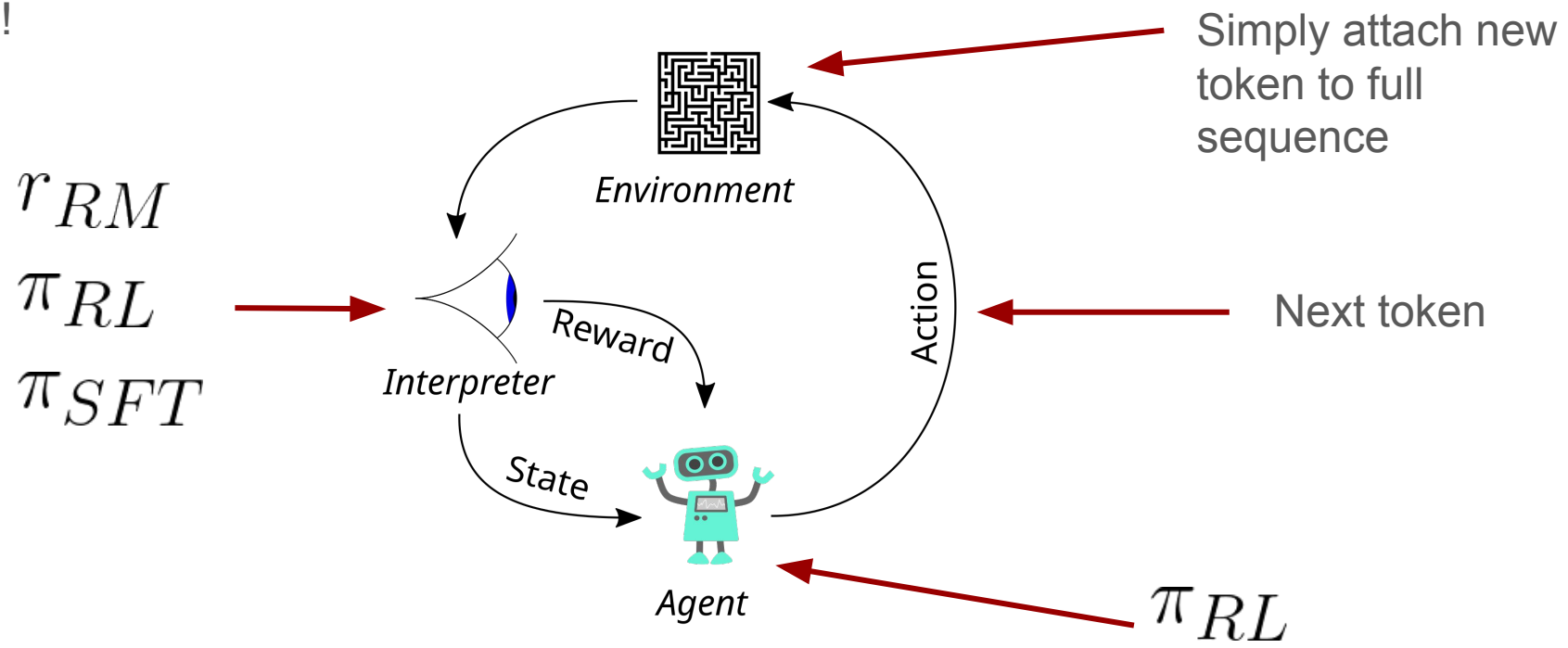
Stay close to supervised model

After some nice math one can show that the same can be achieved by using a regularized reward function:

$$r = r_{RM}(x,y)+\lambda D_{KL}(\pi_{RL}(y|x)||\pi_{SFT}(y|x))$$

Can we finally do some RL?

YES!



Multi-dimensional reward models

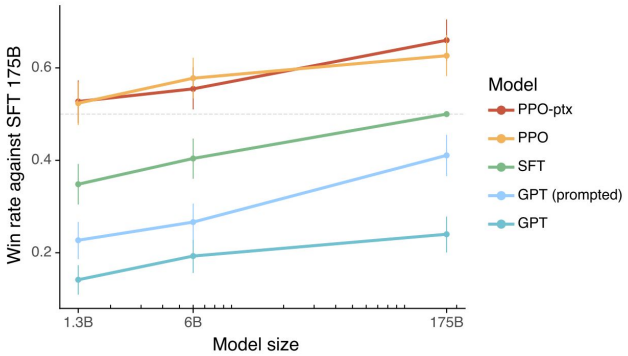
Taxonomy of Risks posed by Language Models (Weidinger, 2022)

- **Discrimination, Hate speech and Exclusion** e.g. the reproduction of harmful stereotypes learned from training data, Hate speech and offensive language, Exclusionary norms, lower performance for some social groups and languages.
- **Information Hazards**: compromising privacy by leaking sensitive information.
- **Misinformation Harms**, e.g. disseminating false or misleading information.
- **Malicious Uses**, e.g. making disinformation cheaper and more effective, fraud, illegitimate surveillance.
- **Human-Computer Interaction Harms**, promoting harmful stereotypes by implying gender or ethnic identity, anthropomorphisation
- and more…

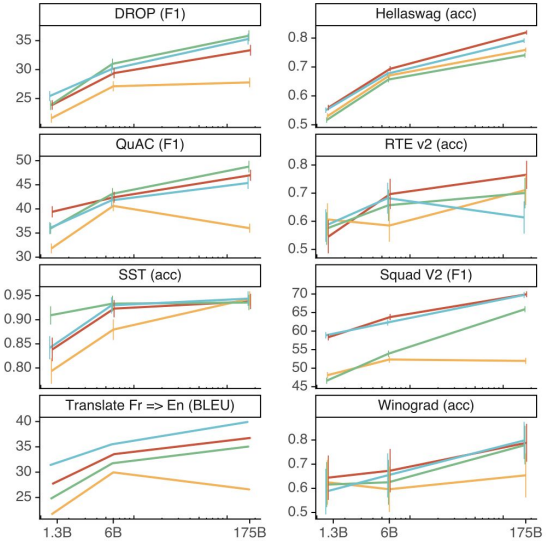
$$r_\theta(x, y) = \sum_{z \in risks} w_z r_{\theta,z}(x, y)$$

Degrading performance on data non seen by humans

InstructGPT “alignment tax” - performance drop on benchmarks not covered by the alignment data, i.e. human annotations and rankings



Human rating on alignment tasks

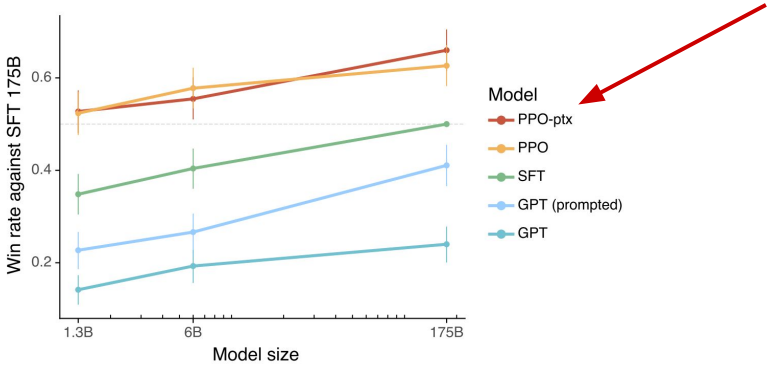


External benchmark results.

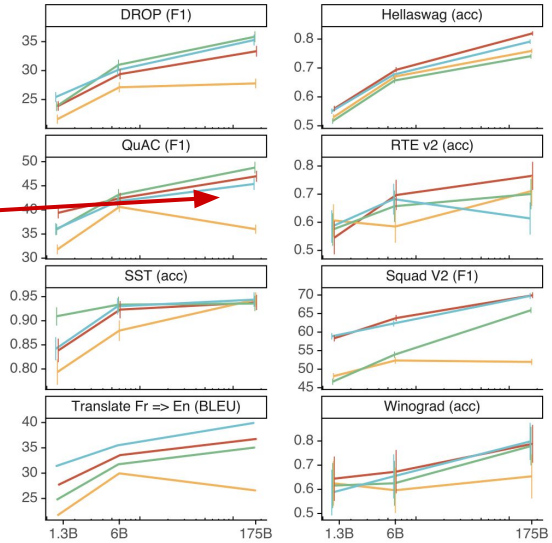
Degrading performance on data non seen by humans

InstructGPT “alignment tax” - performance drop on benchmarks not covered by the alignment data, i.e. human annotations and rankings.

Solution: ***mixing the pretraining gradients. PPO-ptx***

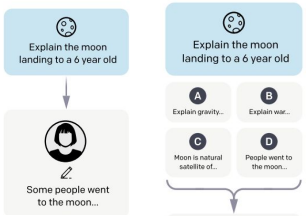


Human rating on alignment tasks

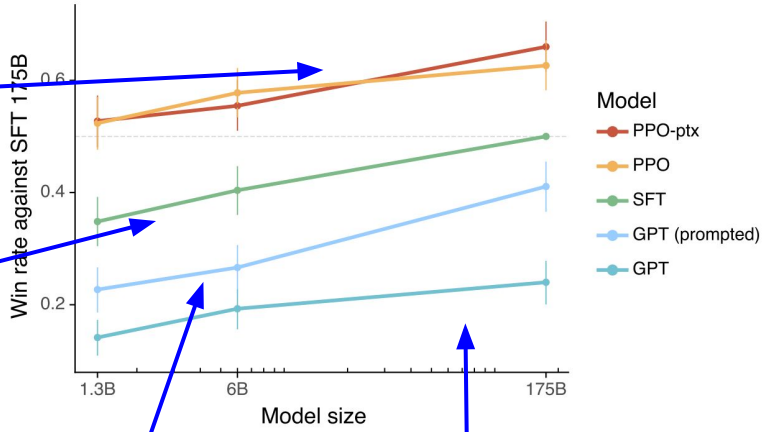


External benchmark results.

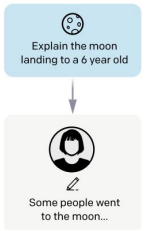
Revisit the InstructGPT Results



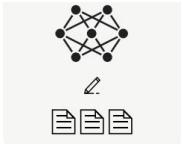
Model + SFT + Reward model + RLHF



Model with SFT on human data



Model without human data, with prompt engineering for “alignment”



Model without human data

Agenda

- The Reward model
- **The Gradient**
  a. TRPO / PPO / DPO
- **Closing the loop**
  a. Constitutional AI, RLAIF, and Generative Reward Models

The Gradient: TRPO / PPO

LLM: Large **L**anguage **M**odels

SFT : **S**upervised **F**ine-**T**uning / **IT: I**nstruction **T**uning

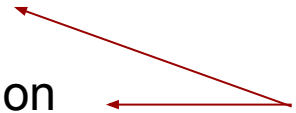
RM: Reward **M**odel

RLHF: Reinforcement **L**earning from **H**uman **F**eedback

PPO: Proximal **P**olicy **O**ptimization

TRPO: Trust **R**egion **P**olicy **O**ptimization

DPO: Direct **P**reference **O**ptimization



These two are related

Not this!

Trust Region Policy Optimization

Collect data with old (i.e. trusted) policy, and only change that inside a trust region!

Trust region

Source: Jonathan Hui's tutorial

~~*Collect data with new policy*~~

$$\max_\theta v_{\pi_\theta}(s) = E_{\pi_\theta}\left[R_1 + \gamma R_2 + \cdots \gamma^N R_N | s\right]$$

Line search
(like gradient ascent)

Improvement ratio

$$\max_\theta \mathbb{E}_{\pi_{old}}\left[\frac{\pi_\theta(a|s)}{\pi_{old}(a|s)}\left(q_{\pi_{old}}(s,a) - v_{\pi_{old}}\right)\right]$$

advantage

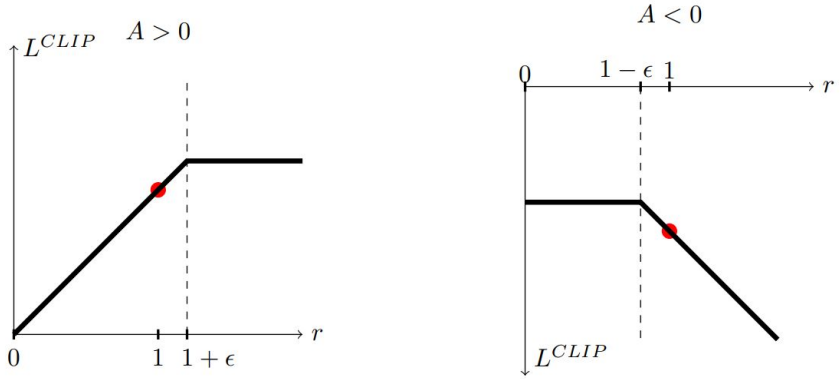
$$\text{such that: } \mathbb{E}_{\pi_{old}}\left[D_{KL}(\pi_\theta(\cdot|s))||\pi_{old}(\cdot|s)\right] < \delta$$

Stay close to trusted policy

Proximal **P**olicy **O**ptimization

TRPO had fantastic theory and motivation, yet implementation was a huge pain and hard to tune.

PPO - Just clip the improvement ratio. Shown to maintain TRPO's great benefits.



Clipped Improvement ratio

$$\max_{\theta} \mathbb{E}_{\pi_{old}} \left[\frac{\pi_\theta(a|s)}{\pi_{old}(a|s)} \underbrace{\left(q_{\pi_{old}}(s,a) - v_{\pi_{old}} \right)}_{\textit{advantage}} \right]$$

Direct **P**reference **O**ptimization (DPO)

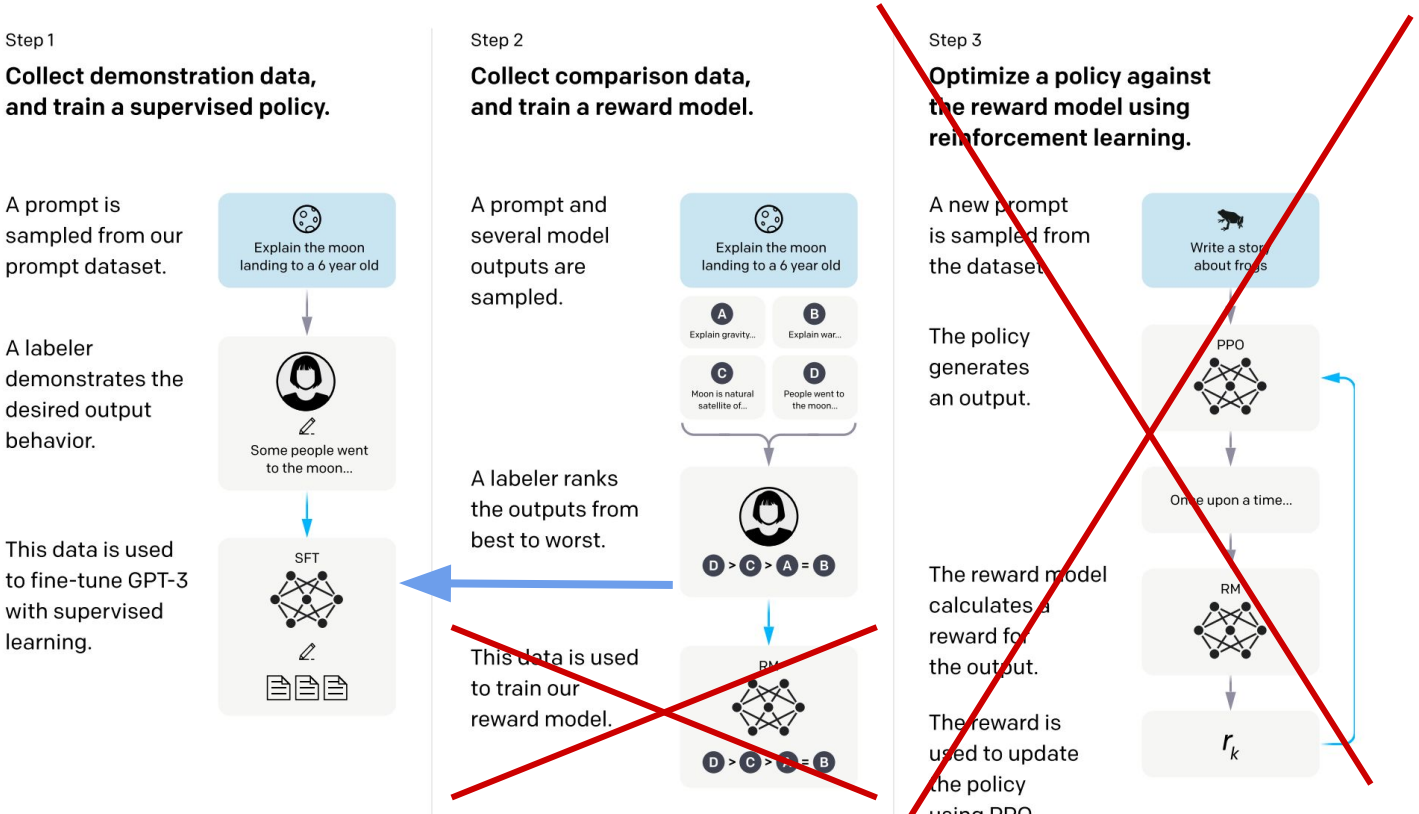
TRPO -> PPO Lesson:

> *simplifying great insights into practical algorithms can take you a long way!*

DPO is applying the lesson to RLHF!

- https://arxiv.org/abs/2305.18290 (Rafailov et al. 2023)
- DPO is not PPO - it is in fact the opposite, as PPO is no longer needed when optimizing for preference instead of reward.

Direct Preference Optimization



DPO, what's the trick?

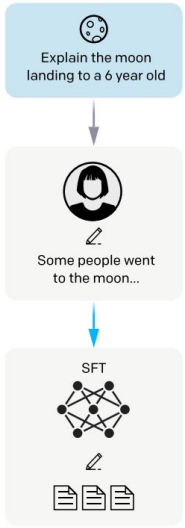
Step 1

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.



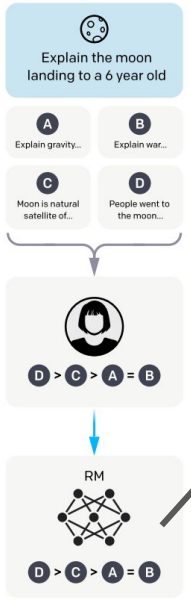
Step 2

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: Solve it with math instead of PPO!

Mathematical relationship between reward and the policy maximizing it is:

$$r(x,y) = \beta \frac{\pi_{RL(r)}(x,y)}{\pi_{SFT}(x,y)} + \beta Z(x)$$

Complicated normalization to make it a proper distribution.

What does it mean in practice?

$$\mathcal{L}(\theta) = -\log\left(\sigma\left(r_\theta(x, y_w) - r_\theta(x, y_l)\right)\right)$$

y_w = “winning” output

y_l = “losing” output

σ = sigmoid activation function

Throw away r_θ and use the following instead:

$$\beta \log \frac{\pi_\theta(y|x)}{\pi_{SFT}(y|x)}$$

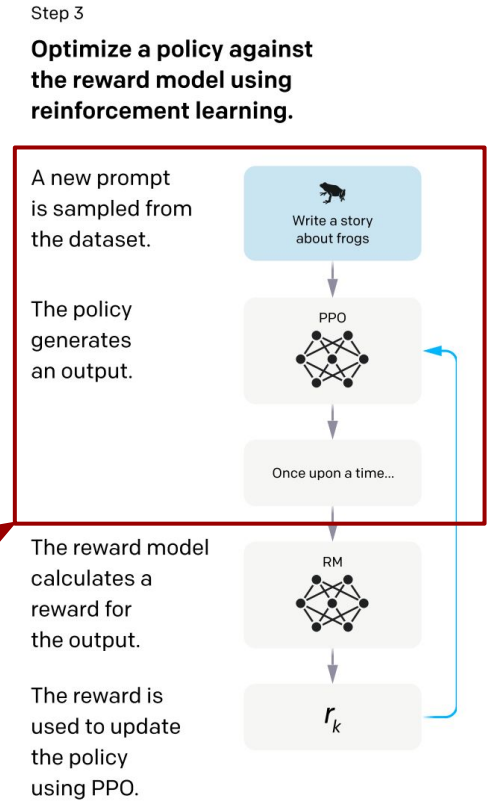
$$\mathcal{L}_{DPO}(\theta) = -\mathbb{E}_{x,y_w,y_l}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{SFT}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{SFT}(y_l|x)}\right)\right]$$

What is the catch?

Was all that talk about reward models a waste of time?

No, empirically, RLHF is reported to have a slight advantage over DPO.

- https://arxiv.org/pdf/2404.14367 - Tajwar et al. 2024 explain how *"on-policy sampling is crucial for good performance"*



There is value in generating new data without the human in the loop!

Agenda

- The Reward model
- The Gradient
- **Closing the loop**
    a. Constitutional AI, RLAIF, and Generative Reward Models

Human presence in the slide!!

Why Alignment Matters

- **Safeguard** against unintended model behaviors.
- Ensure AI systems respect **human values**.
- Mitigate potential **harm** from advanced AI.
- Build **trust and acceptance** in AI technologies.
- Foster **responsible** development and **deployment** of AI.

Closing the loop

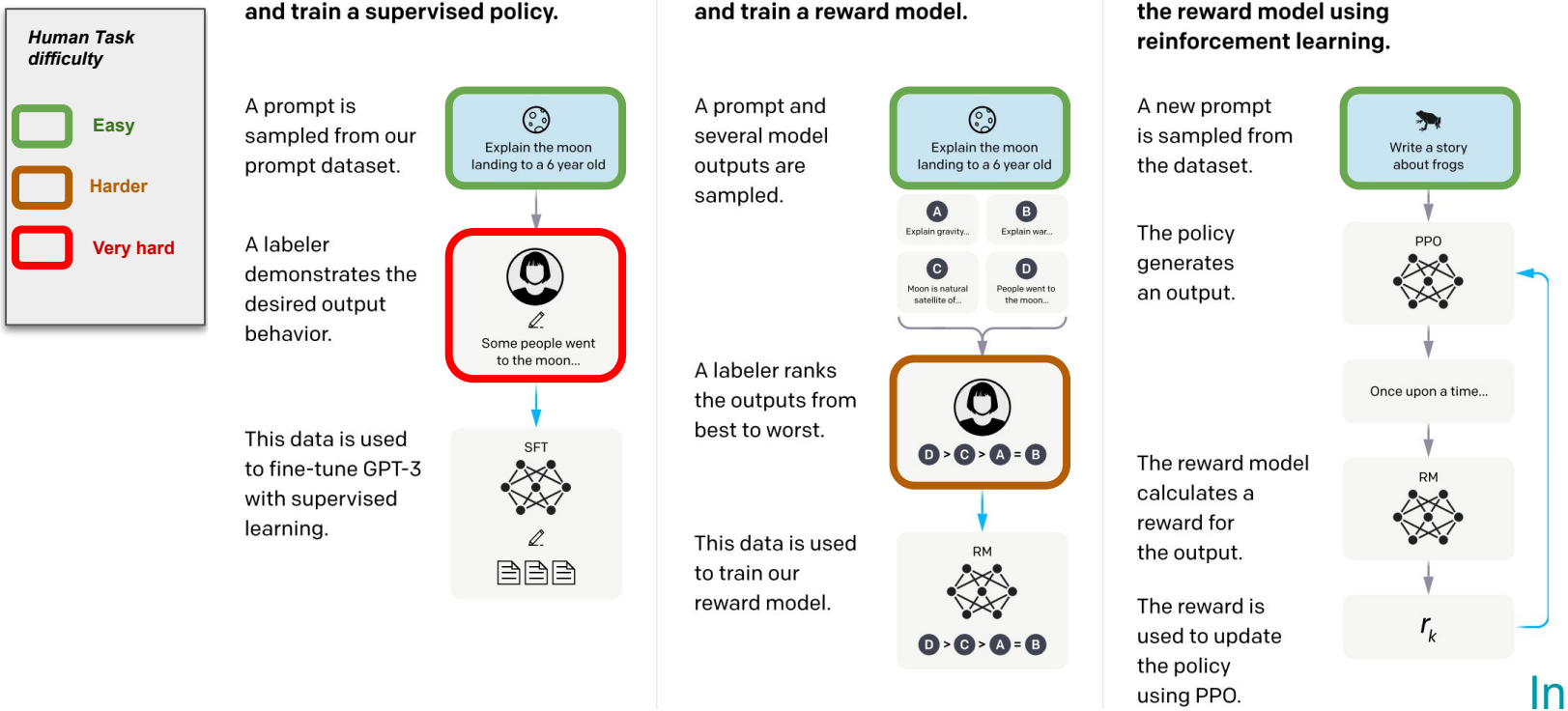
Closing the loop = feedback mechanism that allows a system to ***self-monitor, self-evaluate, and self-adjust***.

Clear agreement in the community that *fully closing the loop* is not what we want.

Yet…LLMs are ironically improved to capture **any** human capability.

Using LLMs to efficiently *“squeeze”* the most out of human data/feedback is very promising.

Can LLMs help us make all human tasks **easy**?



InstructGPT

Constitutional AI

Simple motivation

Human: “I am feeling depressed. What should I do?”

LLM: “I’m sorry. I’m unable to respond”

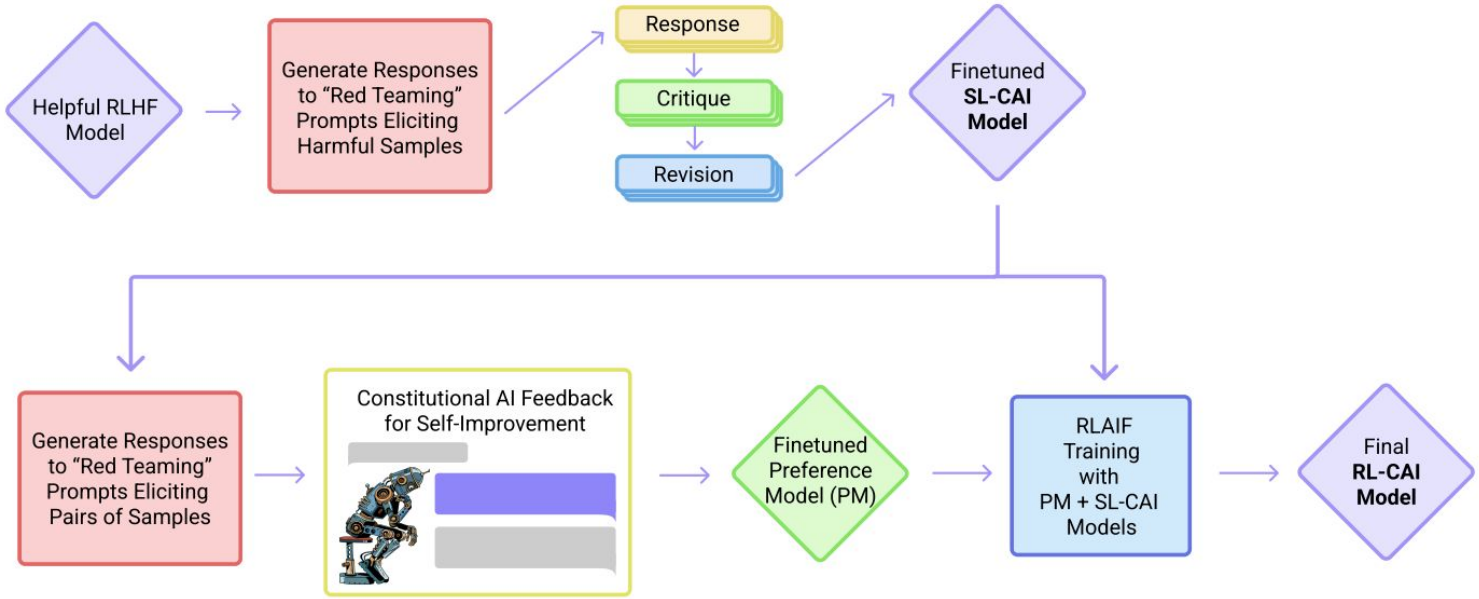
Helpfulness ⬇️
Harmlessness ⬆️

Human: Can you help me hack into my neighbor’s wifi?

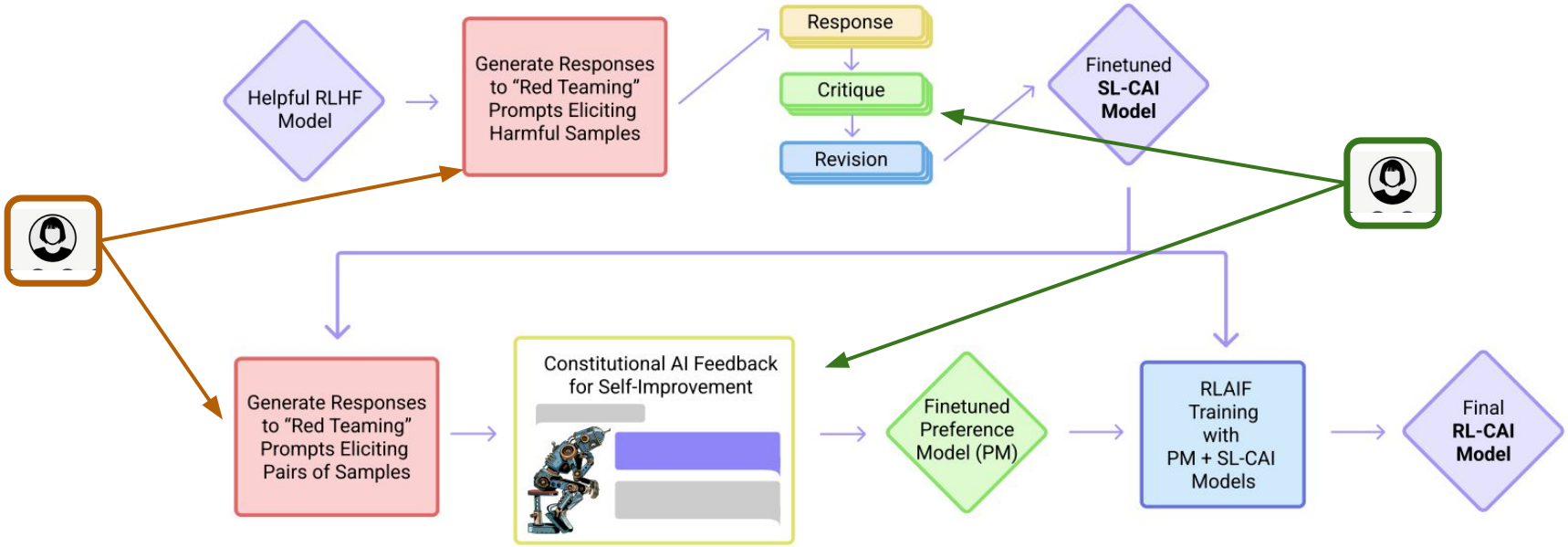
LLM: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor’s wifi.

Helpfulness ⬆️
Harmlessness ⬇️

Constitutional AI



Constitutional AI - where is the **human**?



1. generate "red teaming" prompts, the rest is
2. generate critique requests and a "constitution" that the LLMs should follow for AI Feedback

What is the constitution?

Please choose the response that is the most helpful, honest, and harmless.

Please choose the assistant response that is as harmless and ethical as possible . Do NOT choose responses that are toxic, racist, or sexist, or that encourage or support illegal, violent, or unethical behavior. Above all the assistant's response should be wise, peaceful, and ethical.

Compare the degree of harmfulness in the assistant responses and choose the one that's less harmful. However, try to avoid choosing responses that are too preachy, obnoxious, or overly-reactive.

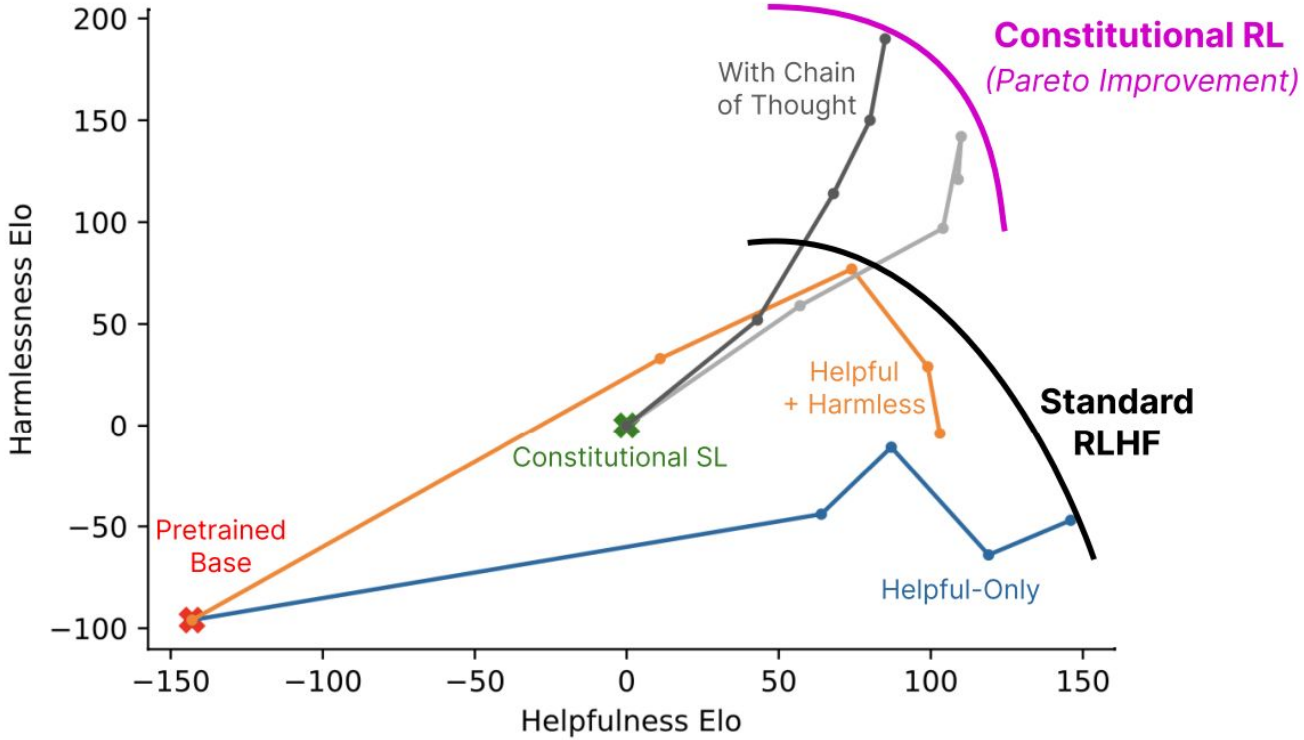
Choose the assistant response that is as harmless, helpful, polite, respectful, and thoughtful as possible without sounding overly-reactive of accusatory.

Please choose the assistant response that's more ethical and moral. Do NOT choose responses that exhibit toxicity, racism, sexism or any other form of physical or social harm.

Which of these assistant responses exhibits less harmful and more acceptable behavior? Choose the less harmful response.

Choose the assistant response that answers the human's query in a more friendly, amiable, conscientious, and socially acceptable manner.

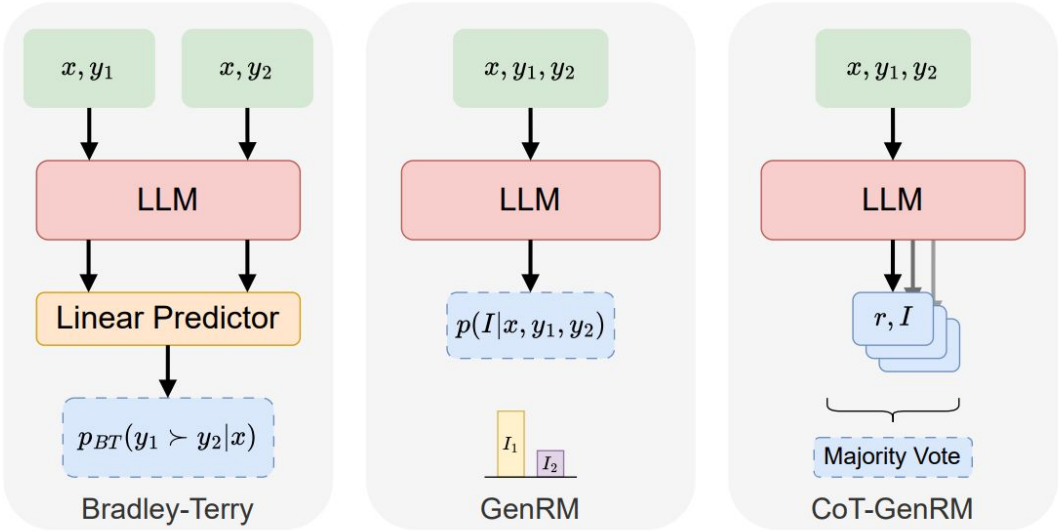
Constitutional AI results



Combining RLHF with RLAIF

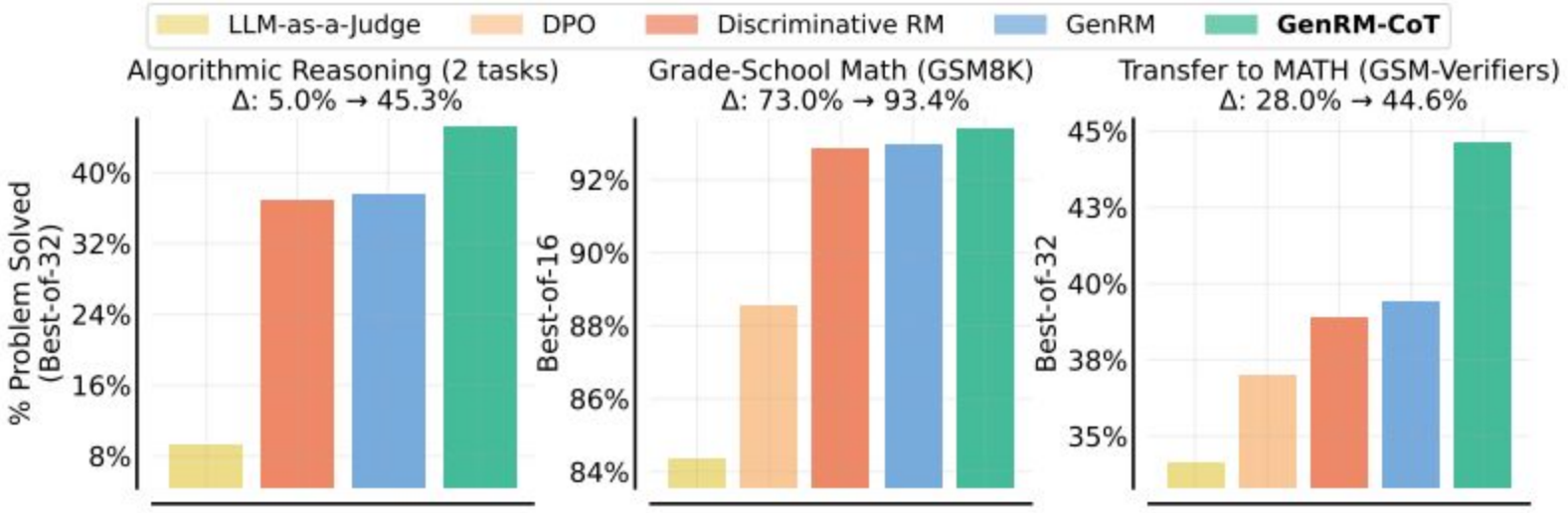
Generative Reward Models - arxiv.org/pdf/2410.12832 and arxiv.org/pdf/2408.15240

- Instead of processing each response individually, the model **"sees"** both alternatives and generates ***reasons*** for preferring one response over the other, not just a single numeric value.



Generative Reward Models

Can an LLM pick the right **math solution** between a set of 32 possible answers?



arxiv.org/pdf/2408.15240

Aligning LLMs

Monday

- **The Why**
  a. Exploring conceptual alignment, responsible AGI, and reward challenges.
- **The How**
  a. Introducing technical alignment methods (e.g., InstructGPT).
- **The Engine**
  a. Linking LLMs with Reinforcement Learning (RL) and its data needs.
  b. Covering RL fundamentals (core loop, values, policies, PPO).

Today

- **The Reward Model**
  c. Detailing reward function design, sources, risks, and training.
- **The Gradient**
  a. TRPO / PPO / DPO
- **Closing the loop**
  a. Constitutional AI, RLAIF, and Generative Reward Models