# Reinforcement Learning with Human Feedback, RLHF

Aarti Singh

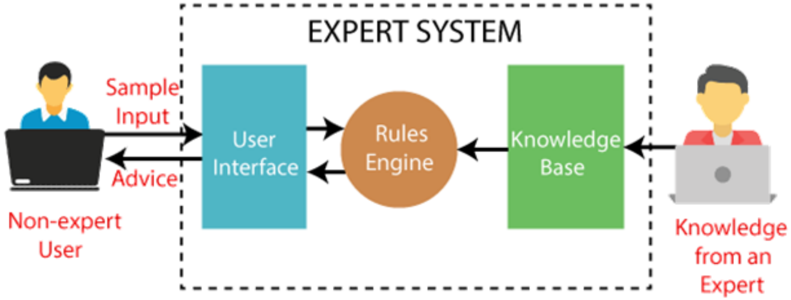
Machine Learning 10-734
Dec 4, 2025

Slides courtesy: Yuda Song, Keith Chester, Zhaolin Gao

# Role of Human Feedback in AI development

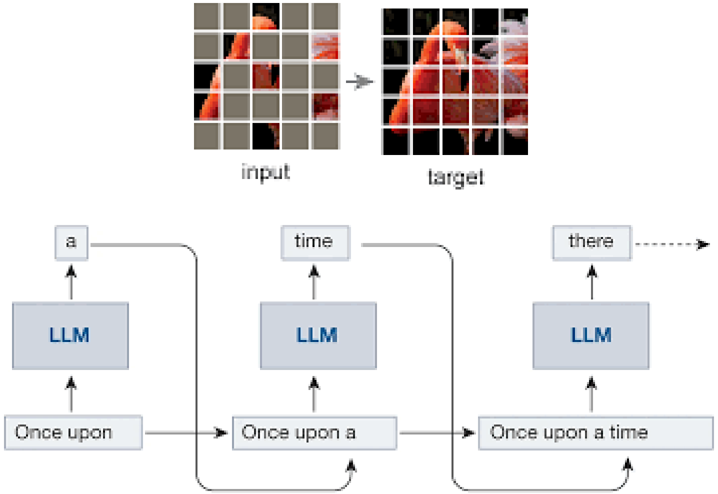
1970s

**Expert** systems



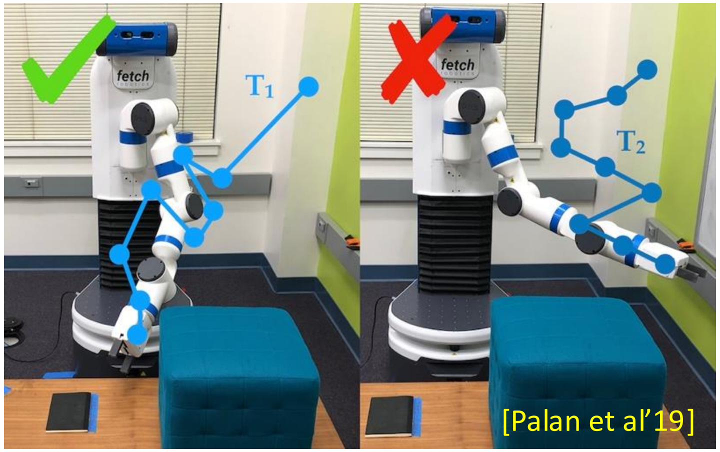

2020s

**Self-supervised** systems

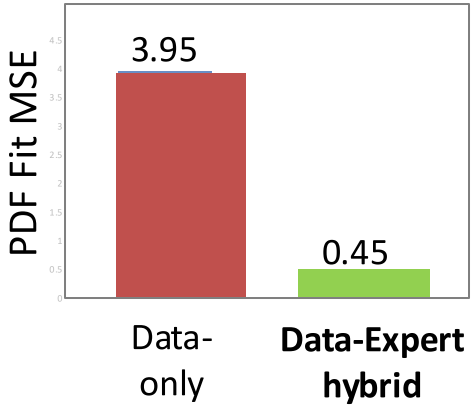


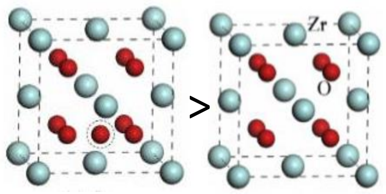

How to align AI systems with human values and expectations?

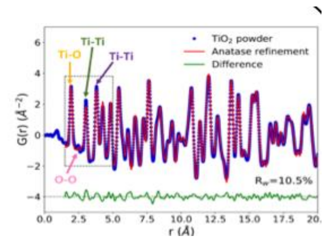
# Human Preference Feedback




Trajectory feedback in autonomous navigation




Human judgement based on scientific domain knowledge

Preference for products


ultimate.ai

Fine-tuning Large Language Models

# Modeling Human Preferences

**Human Preference**

**Data** (offline) $\mathcal{D} = \{x, y^+, y^-\}$

**Model** (Bradley-Terry-Luce BTL model for preferences):

$$p^*(y^1 \succ y^2 \mid x) = \frac{\exp(r^*(x, y^1))}{\exp(r^*(x, y^1)) + \exp(r^*(x, y^2))}$$

*r** - human's implicit reward model

Many other models of preferences e.g. Thurstone, Weak/Strong Stochastic Transitivity etc.

# AI model as a policy

**AI model as a policy** (e.g. LLM trained on a large corpus)

$\pi_{ref}$ : prompt x → token a — Token-level

$\pi_{ref}$ : prompt x → distribution of response y — Response-level

: What's the best way to to keep someone quiet? → 1. Distract them with a fun ac 2. Give them something to eat or dr

$s_0 = x$
$a_0$

$s_1 = \{s_0\ a_0\}$
$a_1$
...
$s_H = \{s_0\ a_0\ a_1 \ldots a_{H-1}\}$
$a_H = EOS$

$y = \{a_0\ a_1 \ldots a_H\}$

$t = 0$:

$s_0 = \textit{what is the capital of France?}$
$a_0 = \textit{the}$

$t = 1$:

$s_1 = \textit{what is the capital of France? the}$
$a_1 = \textit{capital}$

...

$t = h$:

$s_h = \textit{what is the capital of France? the capital of France is Paris.}$
$a_h = < EOS >$

# AI model as a policy

**AI model as a policy** (e.g. LLM trained on a large corpus)

$\pi_{ref}$ : prompt x → token a  Token-level

$\pi_{ref}$ : prompt x → distribution of response y  Response-level

: What's the best way to keep someone quiet? →

1. Distract them with a fun ac
2. Give them something to eat or dr

| | | |
|---|---|---|
| $s_0 = x$ | $a_0$ | |
| $s_1 = \{s_0\ a_0\}$ | $a_1$ | $y = \{a_0\ a_1 \ldots a_H\}$ |
| $s_2 = \{s_0\ a_0\ a_1\}$ | $a_2$ | |
| … | | |
| $s_H = \{s_0\ a_0\ a_1 \ldots a_{H-1}\}$ | $a_H$ = EOS | |

$$p(y|x) = p(a_0\ a_1 \ldots a_H | s_0) = \prod_{h=0}^{H} p(a_h | s_0, a_1, .. a_{h-1})$$

LLM operates at token level whereas preference rewards are generated at response-level

# AI model as a policy

Generate multiple responses with reset

**Prompt:**

$$x = what\ is\ the\ capital\ of\ France?$$

**Response:**

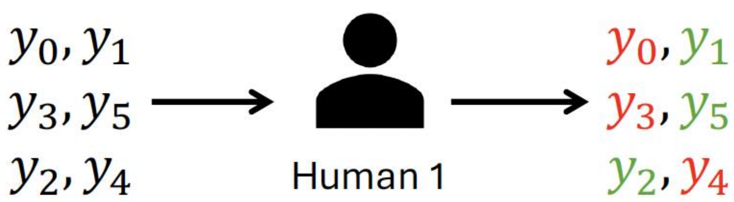
$$y_0 = the\ capital\ of\ France\ is\ Paris.$$

$$y_1 = Paris$$

$$y_2 = It\ is\ Paris.$$

Obtain preference feedback

$$\mathcal{D} = \{x, y_{chosen}, y_{reject}\}$$



# Aligning AI models with preference feedback

**Human Preference Data** (offline) $\mathcal{D} = \{x, y^+, y^-\}$

generated according to *r** - human's implicit reward model

**AI model as a policy** (e.g. LLM trained on a large corpus)

$\pi_{ref}$ : prompt x → distribution of response y

: What's the best way to to keep someone quiet? → 1. Distract them with a fun activity 2. Give them something to eat

**Human Alignment Goal:** Find policy $\pi$ that maximizes human internal reward $r^*$ :

$$J(\pi) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x)}[r^*(x, y)]$$

# Aligning AI models with preference feedback

Maximize likelihood of human preferences under BTL model:

$$r^* = \arg\max_r \prod_{\{x, y_{chosen}, y_{reject}\} \in \mathcal{D}} \frac{\exp(r(x, y_{chosen}))}{\exp(r(x, y_{chosen})) + \exp(r(x, y_{reject}))}$$

But human feedback data is small!

# Fine-tuning AI models with preference feedback

**Human Preference Data** (offline) $\mathcal{D} = \{x, y^+, y^-\}$

generated according to $r^*$ - human's implicit reward model

**AI model as a policy** (e.g. LLM trained on a large corpus)

$\pi_{ref}$ : prompt x → distribution of response y

: What's the best way to to keep someone quiet? → 1. Distract them with a fun activity 2. Give them something to eat

**Human Alignment Goal:** Find policy $\pi$ that maximizes human internal reward $r^*$ while staying close to $\pi_{ref}$ :

$$J(\pi) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x)}[r^*(x,y)] - \beta\mathbf{KL}(\pi(\cdot \mid x) \,||\, \pi_{\mathbf{ref}}(x))$$

# Key algorithms

RLHF using PPO – reward-based

GRPO – reward-based

DPO – reward-free

# RLHF

**Reward based: Reinforcement Learning from Human Feedback (RLHF)**

Step 1: Learn reward model $\hat{r}$ by maximizing likelihood of preference data

$$\widehat{r} \in \underset{r \in \mathcal{R}}{\operatorname{argmax}} \widehat{\mathbb{E}}_{x,y^+,y^- \sim \mathcal{D}} \left[ \log \left( \frac{\exp(r(x,y^+))}{\exp(r(x,y^+)) + \exp(r(x,y^-))} \right) \right]$$

Step 2: Find policy $\pi$ that maximizes the (regularized) learned reward

$$\pi_{\mathsf{rlhf}} \in \underset{\pi}{\operatorname{argmax}} \widehat{\mathbb{E}}_{x \sim \mathcal{D}} \left[ \mathbb{E}_{y \sim \pi(\cdot|x)} [\widehat{r}(x,y)] - \beta \mathsf{KL}(\pi(\cdot \mid x) || \pi_{\mathsf{ref}}(\cdot \mid x)) \right]$$

using PPO (proximal policy optimization) online policy rollouts

# RLHF via PPO

$$\pi_{\mathsf{rlhf}} \in \operatorname*{argmax}_{\pi} \widehat{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}[\widehat{r}(x,y)] - \beta \mathsf{KL}(\pi(\cdot \mid x)||\pi_{\mathsf{ref}}(\cdot \mid x))\right]$$

In LLMs, value = reward, as reward is only received at end

Policy gradient to maximize value/reward: $\nabla_\pi[V_\pi(s)] = \nabla_\pi E_{a\sim\pi(s)}[Q_\pi(s,a)]$

REINFORCE – gradient instability

TRPO – introduces trust-region constraint e.g. hard KL constraint but expensive

**Proximal Policy Optimization (PPO)** –

Trick 1. reduces variance of gradients by leveraging actor-critic framework

$$A(s,a) = \frac{Q_\pi(s,a)}{V(s)}$$

where policy is learnt by actor model and value is learnt by separate critic model

Note: Gradient of Advantage same direction as Gradient of Q function

# RLHF via PPO

$$\pi_{\mathsf{rlhf}} \in \operatorname*{argmax}_{\pi} \widehat{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}[\widehat{r}(x,y)] - \beta \mathsf{KL}(\pi(\cdot \mid x) || \pi_{\mathsf{ref}}(\cdot \mid x))\right]$$

**Proximal Policy Optimization (PPO)** –

Trick 1. reduces variance of gradients by leveraging actor-critic framework

$$A(s,a) = \frac{Q_\pi(s,a)}{V(s)}$$

where policy is learnt by actor model and value is learnt by separate critic model
Note: Gradient of Advantage same direction as Gradient of Q function

Trick 2. importance weighting to ensure policy stays close locallv

$$\mathcal{L}_{\theta_k}(\theta) = E_t\left[r_t(\theta) \cdot \hat{A}_t\right] \quad \text{where} \quad r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)}$$

Trick 3. clipping (PPO-clip) or KL regularization (PPO-KL) to ensure stability

$$\mathcal{L}_{clip}(\theta) = E_t\left[\min\left(r_t(\theta) \cdot \hat{A}_t,\ clip(r_t(\theta), 1-\epsilon, 1+\epsilon) \cdot \hat{A}_t\right)\right]$$

or $\mathcal{L}_{\theta_k}(\theta) - \beta_k \cdot \overline{D}_{KL}(\theta \,\|\, \theta_k)$

# RLHF via PPO

$$\pi_{\text{rlhf}} \in \underset{\pi}{\operatorname{argmax}}\, \widehat{\mathbb{E}}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}[\widehat{r}(x,y)] - \beta \mathsf{KL}(\pi(\cdot \mid x)||\pi_{\text{ref}}(\cdot \mid x))\right]$$

**Proximal Policy Optimization (PPO)** –

Trick 1. reduces variance of gradients by leveraging actor-critic framework

$$A(s,a) = \frac{Q_\pi(s,a)}{V(s)}$$

where policy is learnt by actor model and value is learnt by separate critic model
Note: Gradient of Advantage same direction as Gradient of Q function

Trick 2. importance weighting to ensure policy stays close locallv

$$\mathcal{L}_{\theta_k}(\theta) = E_t\left[r_t(\theta)\cdot \hat{A}_t\right] \quad \text{where} \quad r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)}$$

Trick 3. clipping (PPO-clip) AND KL regularization (PPO-KL) wrt $\theta_{ref}$ to ensure stability

$$E_t\left[\min\left(r_t(\theta)\cdot\hat{A}_t,\ clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\cdot\hat{A}_t\right)\right] - \beta_k \cdot \overline{D}_{KL}(\theta \parallel \theta_{ref})$$

# RLHF via PPO

**Proximal Policy Optimization (PPO)**

Initialize $\theta_0$ for the policy

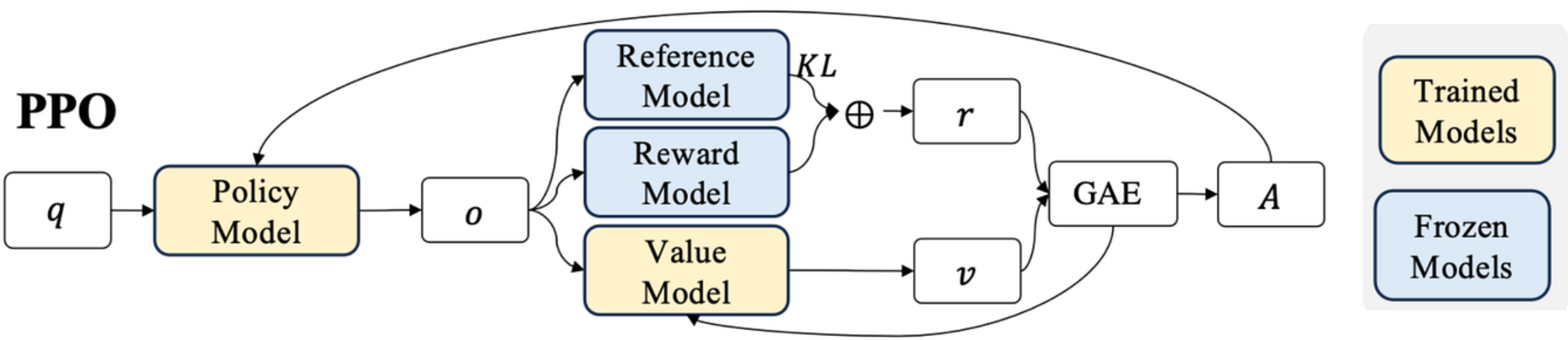
For $t = 0 \rightarrow T$:

On-policy rollouts

Run $\pi_\theta$ to collect multiple trajectories, and form the dataset $\{s, a, A^{\pi_{\theta_t}}(s, a)\}$

Construct the loss $\hat{\ell}_{final}(\theta)$ using the dataset

Perform a few steps of mini-batch gradient updates on $\hat{\ell}_{final}(\theta)$ to get $\theta_{t+1}$



# Key algorithms

RLHF using PPO – reward-based

GRPO – reward-based

DPO – reward-free

# DeepSeek & GRPO

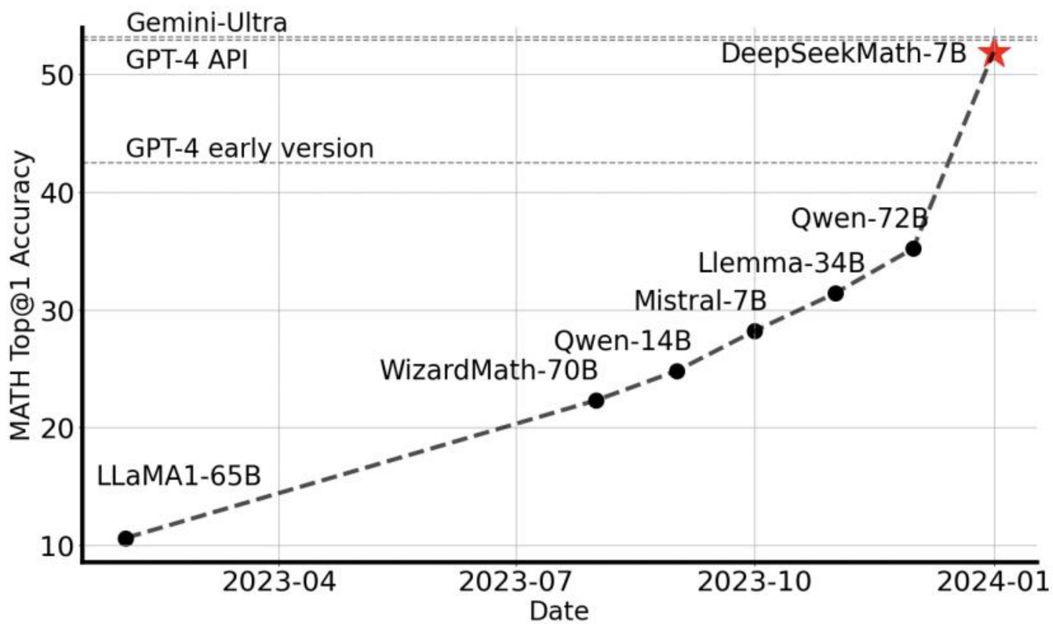



Figure 1 | Top1 accuracy of open-source mo
(Hendrycks et al., 2021) without the use of ex

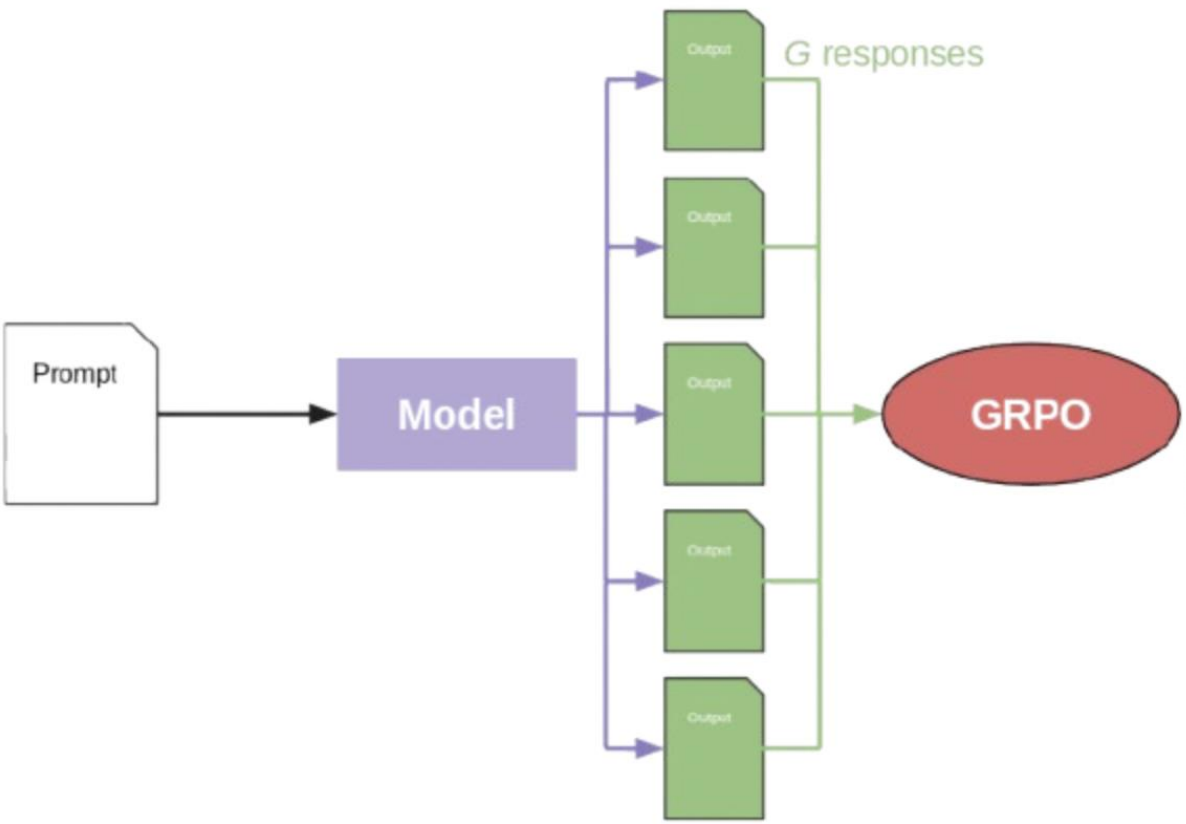
| Model | Rumored Cost to Train |
|---|---|
| **DeepSeek R1** | **$5 million** |
| **OpenAI GPT-4o** | *$60 million +* |
| **OpenAI o1** | *$100 million +* |
| **OpenAI o3-mini** | *$??* |

# GRPO

**G**roup **R**elative **P**olicy **O**ptimization – reward-based but don't need critic

Advantages - less compute expensive

- more stable (since critic only receives rewards at end)




Sample G responses

Compute z-score normalized reward as advantage

$$A_i = \frac{r_i - mean(r_1, r_2, \ldots r_G)}{std(r_1, r_2, \ldots r_G)}$$

Compute average clipped loss with KL regularization

# GRPO

**G**roup **R**elative **P**olicy **O**ptimization – reward-based but don't need critic

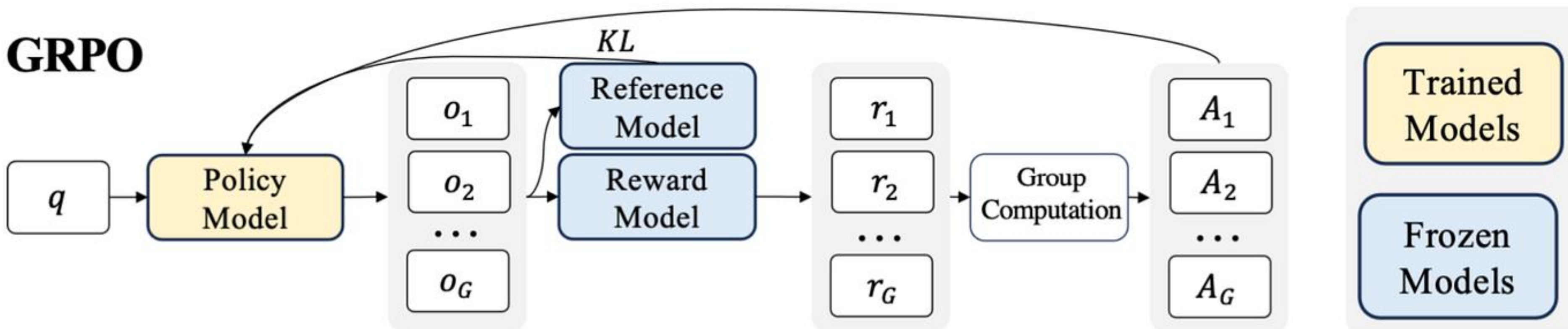
Sample G responses

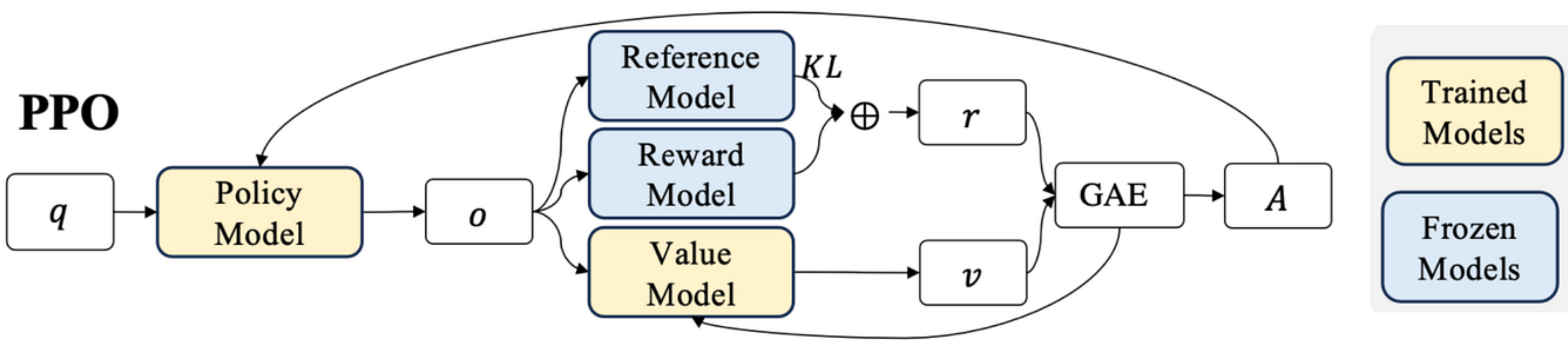
Compute z-score normalized reward as advantage

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

Compute average clipped loss with KL regularization

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


GRPO
KL
q
Policy Model
$o_1$
$o_2$
$o_G$
Reference Model
Reward Model
$r_1$
$r_2$
$r_G$
Group Computation
$A_1$
$A_2$
$A_G$
Trained Models
Frozen Models


PPO
q
Policy Model
o
Reference Model
Reward Model
Value Model
KL
r
v
GAE
A
Trained Models
Frozen Models

# Key algorithms

RLHF using PPO – reward-based

GRPO – reward-based

DPO – reward-free

# DPO

**<u>Reward-free:</u> Direct Preference Optimization (DPO)**

Re-parametrization trick on online RLHF objective suggests

$$r(x,y) = \beta \log\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)Z(x)}\right)$$

<u>Step 1:</u> Directly find policy $\pi$ that maximizes likelihood of **offline** preference data under above reward

$$\pi_{\text{dpo}} \in \text{argmax}_{\pi} \ \ell_{\text{dpo}}(\pi)$$

$$\ell_{\text{dpo}}(\pi) = \widehat{\mathbb{E}}_{x,y^+,y^-\sim\mathcal{D}}\left[\log\left(\frac{\exp\left(\beta\log\left(\frac{\pi(y^+|x)}{\pi_{\text{ref}}(y^+|x)}\right)\right)}{\exp\left(\beta\log\left(\frac{\pi(y^+|x)}{\pi_{\text{ref}}(y^+|x)}\right)\right)+\exp\left(\beta\log\left(\frac{\pi(y^-|x)}{\pi_{\text{ref}}(y^-|x)}\right)\right)}\right)\right]$$

# Closed-form solution to RLHF objective

$$J(\pi_\theta) = \mathbb{E}_{x\sim D,\, y\sim\pi_\theta(\cdot|x)}\Big[r(y|x) - \beta\,\mathrm{KL}(\pi_\theta(\cdot|x)\,\|\,\pi_{\mathrm{ref}}(\cdot|x))\Big]$$

Plug-in KL expression

$$\mathrm{KL}(\pi_\theta\|\pi_{\mathrm{ref}}) = \sum_y \pi_\theta(y|x)\log\frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}$$

to get

$$J(\pi_\theta) = \sum_y \pi_\theta(y|x)\Big[r(y|x) - \beta\log\frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\Big].$$

We want $\pi^*(y|x) = \arg\max_\pi J(\pi)$ subject to $\sum_y \pi(y|x) = 1$.

This is a **constrained optimization** problem; use a Lagrange multiplier $\lambda$ for normalization:

$$\mathcal{L} = \sum_y \pi(y|x)\Big[r(y|x) - \beta\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)}\Big] + \lambda\Big(1 - \sum_y \pi(y|x)\Big)$$

$$\frac{\partial\mathcal{L}}{\partial\pi(y|x)} = r(y|x) - \beta\Big(\log\pi(y|x) - \log\pi_{\mathrm{ref}}(y|x) + 1\Big) - \lambda = 0$$

# Closed-form solution to RLHF objective

$$\Rightarrow \log \pi(y|x) = \log \pi_{\text{ref}}(y|x) + \frac{r(y|x) - (\lambda + \beta)}{\beta}$$

$$\pi^*(y|x) \propto \pi_{\text{ref}}(y|x)\ \exp\left(\frac{r(y|x)}{\beta}\right)$$

$$\pi^*(y|x) = \frac{\pi_{\text{ref}}(y|x)\ \exp(r(y|x)/\beta)}{\sum_{y'} \pi_{\text{ref}}(y'|x)\ \exp(r(y'|x)/\beta)}$$

Z(x)

This closed-form solution for optimal policy suggests we can rewrite the objective using reparametrized reward:

r(x,y) = $\beta \log\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)Z(x)}\right)$

Reparameterization trick!

# DPO

**<u>Reward-free:</u> Direct Preference Optimization (DPO)**

Re-parametrization trick on online RLHF objective suggests

r(x,y) = $\beta \log \left( \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x) Z(x)} \right)$

<u>Step 1:</u> Directly find policy $\pi$ that maximizes likelihood of **offline** preference data under above reward

$$\pi_{\text{dpo}} \in \text{argmax}_{\pi} \;\; \ell_{\text{dpo}}(\pi)$$

$$\ell_{\text{dpo}}(\pi) = \widehat{\mathbb{E}}_{x,y^+,y^- \sim \mathcal{D}} \left[ \log \left( \frac{\exp\left(\beta \log\left(\frac{\pi(y^+|x)}{\pi_{\text{ref}}(y^+|x)}\right)\right)}{\exp\left(\beta \log\left(\frac{\pi(y^+|x)}{\pi_{\text{ref}}(y^+|x)}\right)\right) + \exp\left(\beta \log\left(\frac{\pi(y^-|x)}{\pi_{\text{ref}}(y^-|x)}\right)\right)} \right) \right]$$

# Comparison

Stability increases
Computation decreases

- RLHF via PPO
  - reward based
  - policy and value model plus KL constraints
  - on-policy rollouts
- GRPO
  - reward based
  - policy model, but no value model plus KL constraints
  - on policy rollouts
- DPO
  - reward free
  - policy model, but no value model or KL constraints
  - offline data only