Language model reinforcement learning at a glance

Core loss $\mathcal{L}(\theta)$ per RL algorithm

Policy Gradient

$$-\mathbb{E}_{\tau}\left[\sum_{t=0}^{T}\log\pi_\theta(a_t \mid s_t)\,A^{\pi_\theta}(s_t,a_t)\right]$$

REINFORCE
(REward Increment = Nonnegative Factor × Offset Reinforcement × Characteristic Eligibility)

$$-\frac{1}{T}\sum_{t=1}^{T}\log\pi_\theta(a_t \mid s_t)\,(G_t - b(s_t))$$

REINFORCE Leave One Out
(RLOO)

$$-\frac{1}{K}\sum_{i=1}^{K}\sum_{t}\log\pi_\theta(a_{i,t} \mid s_{i,t})\left(R_i - \frac{1}{K-1}\sum_{j\neq i}R_j\right)$$

Proximal Policy Optimization
(PPO)

$$-\frac{1}{T}\sum_{t=1}^{T}\min\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}A_t,\ \text{clip}\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)},\,1{-}\varepsilon,\,1{+}\varepsilon\right)A_t\right)$$

Group Relative Policy Optimization
(GRPO)

$$-\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|a_i|}\sum_{t=1}^{|a_i|}\min\left(\frac{\pi_\theta(a_{i,t} \mid s_{i,t})}{\pi_{\theta_{\text{old}}}(a_{i,t} \mid s_{i,t})}\hat{A}_i,\ \text{clip}\left(\frac{\pi_\theta(a_{i,t} \mid s_{i,t})}{\pi_{\theta_{\text{old}}}(a_{i,t} \mid s_{i,t})},\,1{-}\varepsilon,\,1{+}\varepsilon\right)\hat{A}_i\right)$$

where $\hat{A}_i = \dfrac{r_i - \overline{r}}{\text{std}(r)}$

Group Sequence Policy Optimization
(GSPO)

$$-\frac{1}{G}\sum_{i=1}^{G}\min\left(\left(\frac{\pi_\theta(a_i \mid s)}{\pi_{\theta_{\text{old}}}(a_i \mid s)}\right)^{\frac{1}{|a_i|}}A_i,\ \text{clip}\left(\left(\frac{\pi_\theta(a_i \mid s)}{\pi_{\theta_{\text{old}}}(a_i \mid s)}\right)^{\frac{1}{|a_i|}},\,1{-}\varepsilon,\,1{+}\varepsilon\right)A_i\right)$$

Clipped Importance Sampling Policy Optimization
(CISPO)

$$-\frac{1}{\sum_i|a_i|}\sum_{i=1}^{K}\sum_{t=1}^{|a_i|}\text{sg}\Big(\text{clip}\big(\tfrac{\pi_\theta(a_{i,t}|s)}{\pi_{\theta_{\text{old}}}(a_{i,t}|s)},\,1{-}\varepsilon_{\text{low}},\,1{+}\varepsilon_{\text{high}}\big)\Big)\,A_{i,t}\log\pi_\theta(a_{i,t} \mid s)$$

where $\text{sg}(\cdot)$ = stop gradient

Other core equations

RLHF Objective

$$J(\theta) = \max_{\pi}\ \mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{y\sim\pi(y|x)}\Big[r_\theta(x,y)\Big] - \beta\,\mathcal{D}_{\text{KL}}\Big(\pi(y|x)\,\|\,\pi_{\text{ref}}(y|x)\Big)$$

Bradley–Terry Reward Model

$$\mathcal{L}(\theta) = -\log\sigma\Big(r_\theta(y_c \mid x) - r_\theta(y_r \mid x)\Big)$$

Direct Preference Optimization
(DPO)

$$\mathcal{L}(\theta) = -\,\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} - \beta\log\frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)}\right)\right]$$

Notation

x, y	Prompt, completion	A_t	Advantage estimate
y_c, y_r	Chosen / rejected completion	β	KL penalty coefficient
$y_c \succ y_r$	y_c is preferred over y_r	$\sigma(z)$	Sigmoid: $1/(1+e^{-z})$
π_θ	Policy (the model being trained)	$\mathcal{D}_{\text{KL}}(P\|Q)$	KL divergence between P and Q
π_{ref}	Reference policy (frozen copy)	G_t	Return: $\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$
$\pi_{\theta_{\text{old}}}$	Policy at start of RL batch updates	$V(s)$	Value: $\mathbb{E}[G_t \mid S_t = s]$
$r_\theta(y \mid x)$	Reward model score		