25.03.04

Reward Model (usually a small language model)

$r_\theta(x, y)$

prompt, response

$x \xrightarrow{SFT} y_1 > y_2 > y_3 > y_4$

MLE:

$$\max \frac{1}{\binom{4}{2}} \sum_{(y_w, y_l)} \log \left[\frac{\exp(r_\theta(x, y_w))}{\exp(r_\theta(x, y_w)) + \exp(r_\theta(x, y_l))} \right]$$

y_w: winning response, y_l: losing response

For DPO:

Reward Modeling with $r_\theta(x, y) = \log \left[\frac{P_\theta^{RL}(y|x)}{P^{SFT}(y|x)} \right]$

New Model

SFT Model