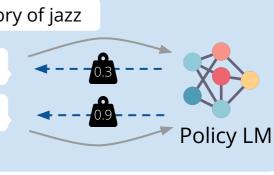
Direct Preference Optimization (DPO)

Weighted Preference Optimization (WPO)

 \mathcal{X} : Write me a poem about the history of jazz Policy LM Preference Data

$$x$$
 : Write me a poem about the history of jazz
$$y_w^{(1)} \succ y_l^{(1)} \\ y_w^{(2)} \succ y_l^{(2)} \\ \text{Preference Data}$$



Preference Data
$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log p(y_w \succ y_l|x)\right]$$

$$\mathcal{L}_{ ext{WPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[w(x,y_w)w(x,y_l)\log p(y_w\succ y_l|x)
ight]$$