RLHF objective:

$$\max_{\pi} \mathbb{E}_{x,y} \left[ r(x,y) - \beta D_{KL}(\pi(y|x) \parallel \pi_{\text{ref}}(y|x)) \right]$$

SFT (instruction-tuned LLM) vs. current aligned LLM

(x,y) data used for SFT is different than that used for RLHF, but come from same distribution

why do we need KL?

DPO (direct preference optimization):

- no explicit reward model
- not going to sample outputs y|x from the model
- "rollouts"
- "preference tuning"

$$\max_{\pi} \mathbb{E}_{x,y} \left[ r(x,y) - \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right]$$
\[
\begin{align*}
= \min_{\Pi} \mathbb{E}_{x \sim y} \left[ \log \frac{\Pi(y|x)}{\Pi_{ref}(y|x)} - \frac{1}{\beta} r(x,y) \right]
\end{align*}
\]

Introduce a new policy \( \Pi^* \) that incorporates the reward term as well as \( \Pi_{ref} \):

\[
\Pi^*(y|x) = \frac{\Pi_{ref}(y|x) \exp \left( \frac{1}{\beta} r(x,y) \right)}{\sum_y \Pi_{ref}(y|x) \exp \left( \frac{1}{\beta} r(x,y) \right)}
\]

\( Z(x) \), normalizer / partition function

Substitute \( Z(x) \) into our objective:

\[
\begin{align*}
= \min_{\Pi} \mathbb{E}_{x \sim y} \log \left[ \frac{\Pi(y|x)}{\mathbb{E}_{x \sim y} Z(x) \Pi_{ref}(y|x) \exp \left( \frac{1}{\beta} r(x,y) \right)} \right] - \log Z
\end{align*}
\]

\[
= \min_{\Pi} \mathbb{E}_{x \sim y} \log \left[ \frac{\Pi(y|x)}{\Pi^*(y|x)} \right] - \log Z
\]

\( \to \) KL divergence

\[
= \min_{\Pi} \mathbb{E}_{x} D_{KL} \left( \Pi(y|x) \parallel \Pi^*(y|x) \right) - \log Z
\]
KL div. is minimized at 0 when \( \pi(\cdot|x) = \pi^*(\cdot|x) \)

\[
\pi^*(y|x) = \begin{aligned}
\pi^*(y|x) &= \frac{\pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)}{Z(x)} \\
\text{optimal policy}
\end{aligned}
\]

Solve the above for \( r(x, y) \)

\[
r(x, y) = -\beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)
\]

Bradley-Terry pref model:

\[
P(y_w > y_L | x) = \frac{\exp\left( r(x, y_w) \right)}{\exp\left( r(x, y_w) \right) + \exp\left( r(x, y_L) \right)}
\]

Substitute reward function:

\[
P(y_w > y_L | x) = \frac{1}{1 + \exp\left( \beta \log \frac{\pi^*(y_L|x)}{\pi_{\text{ref}}(y_L|x)} - \beta \log \frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \right)}
\]

Convert to loss fn (neg. log likelihood)

\[
L_{\text{DEO}}(\pi_\theta, \pi_{\text{ref}}) = -\mathbb{E}_{x \sim \pi_{\text{ref}}, (y_w, y_L)} \log \sigma\left( \beta \log \frac{\pi_\theta(y_L|x)}{\pi_{\text{ref}}(y_L|x)} - \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \right)
\]

-aligned model we are training
nice properties of DPO:
- no explicit reward model
- no need for rollouts from the policy

\[ q(x, y, y_{l}) \]

SET LLM \[ \rightarrow \]
DPO LLM

fine-tune on pref. judgments using DPO loss above