July 29, 2025

$$KL[\pi_{\theta_{old}}||\pi_\theta] = \mathop{\mathbb{E}}_{o\sim\pi_{\theta_{old}}} \log\left(\frac{\pi_{\theta_{old}}(o|\text{context})}{\pi_\theta(o|\text{context})}\right)$$

$$KL[\pi_\theta||\pi_{ref}] = \mathop{\mathbb{E}}_{o\sim\pi_\theta}\left[\frac{\pi_{ref}(o|\text{context})}{\pi_\theta(o|\text{context})} - \log\left(\frac{\pi_{ref}(o|\text{context})}{\pi_\theta(o|\text{context})}\right) - 1\right]$$

$$A_o = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})} \qquad \propto\ =\ \frac{1}{G}\cdot\frac{1}{\text{output length}}$$

$$A_o = \mathbf{r}_i - \text{mean}(\mathbf{r}) \qquad \propto\ =\ \frac{1}{G}\cdot\frac{1}{\text{max length}}$$

$$\propto\ =\ \frac{1}{G}\cdot\frac{1}{\text{output length}}$$

$$\propto\ =\ \frac{1}{G}\cdot\frac{1}{\text{max length}}$$

$$\propto \sum_{\text{token } o} \text{clip}\left(\frac{\pi_\theta(o|\text{context})}{\pi_{\theta_{old}}(o|\text{context})}\right) A_o - \beta KL[\pi_\theta\|\pi_{ref}]$$

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] - \beta KL[\pi_\theta\|\pi_{ref}]\right\}$$