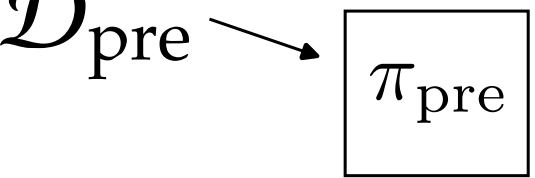
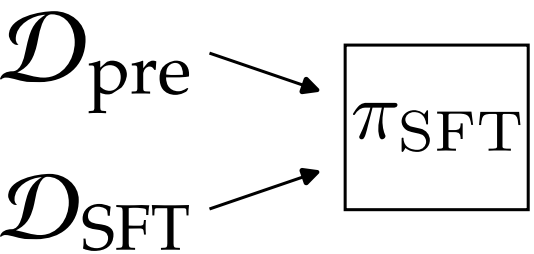
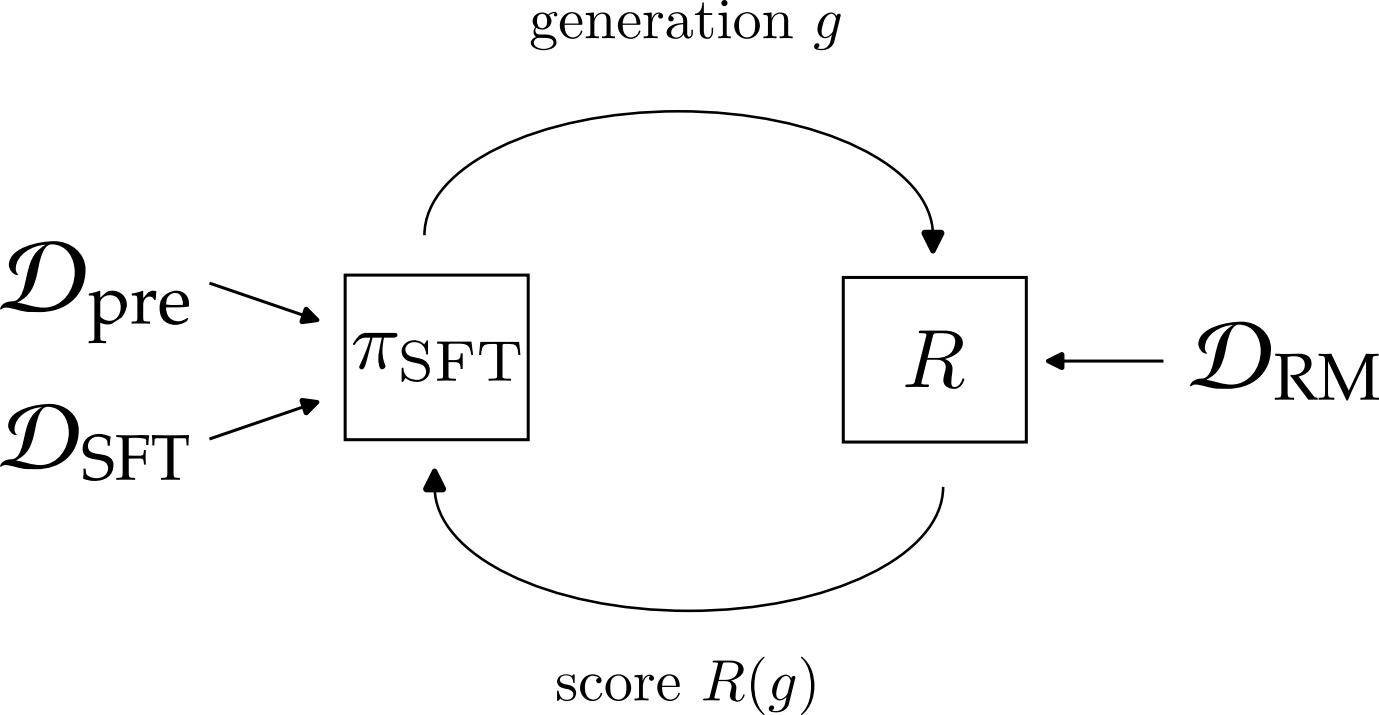
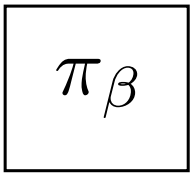

$\mathcal{D}_{\text{pre}}$
π_{pre}


$\mathcal{D}_{\text{pre}}$
$\mathcal{D}_{\text{SFT}}$
π_{SFT}

$\mathcal{D}_{\text{pre}}$
$\mathcal{D}_{\text{SFT}}$
π_{SFT}
R
$\mathcal{D}_{\text{RM}}$


generation g
$\mathcal{D}_{\text{pre}}$
$\mathcal{D}_{\text{SFT}}$
π_{SFT}
R
$\mathcal{D}_{\text{RM}}$
score $R(g)$


π_β