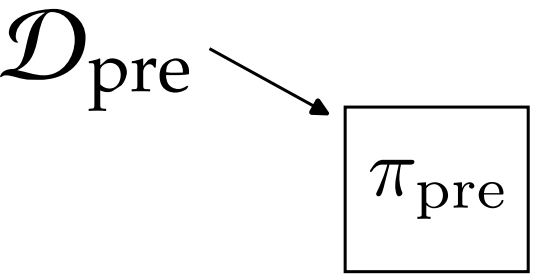
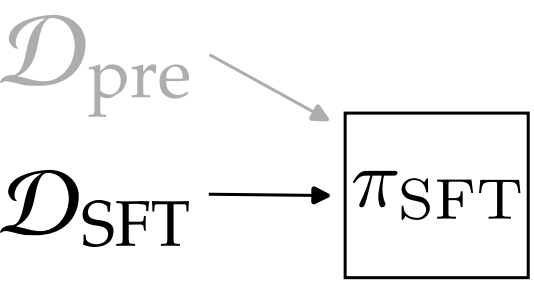
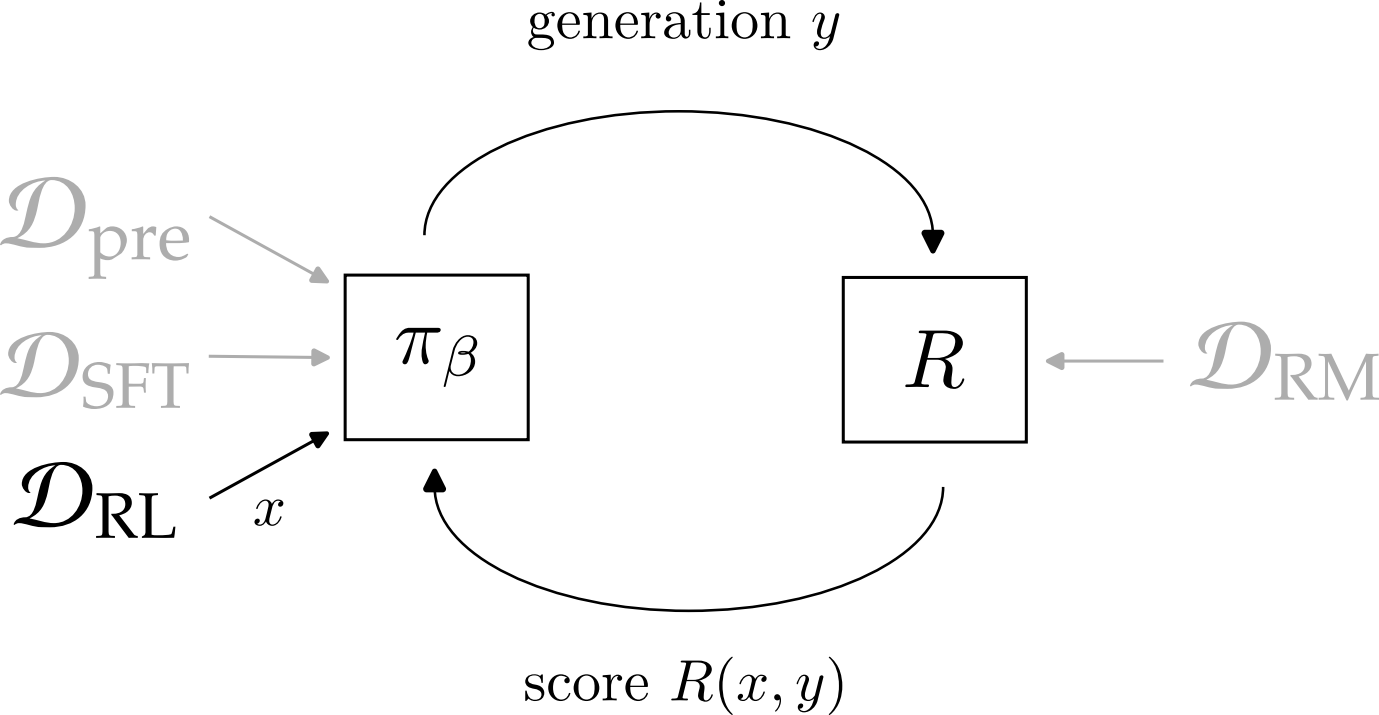
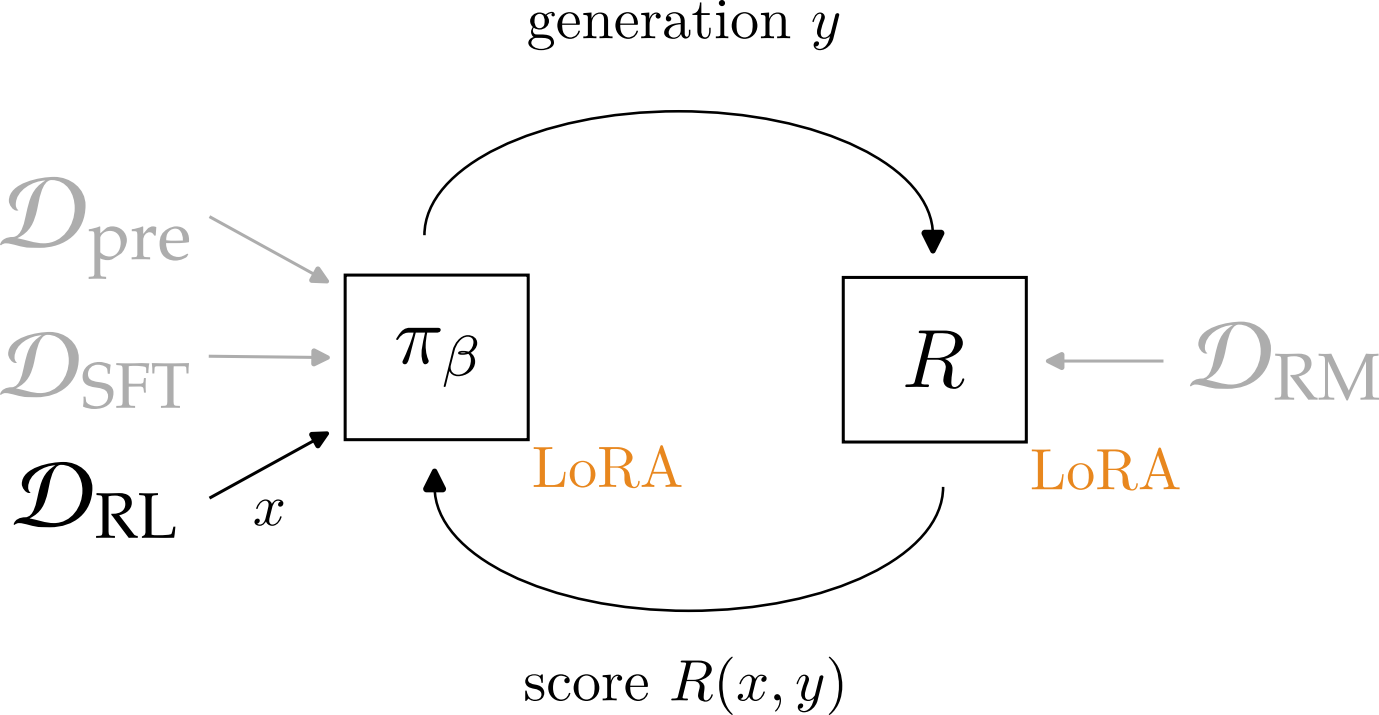
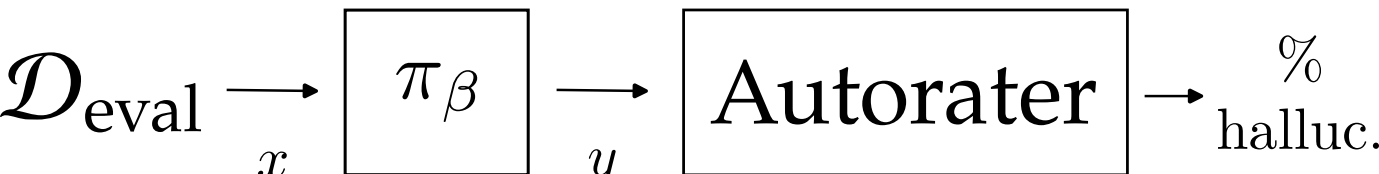

$\mathcal{D}_{\text{pre}}$
π_{pre}


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


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


generation y
$\mathcal{D}_{\text{pre}}$
$\mathcal{D}_{\text{SFT}}$
$\mathcal{D}_{\text{RL}}$
x
π_β
R
$\mathcal{D}_{\text{RM}}$
score $R(x, y)$


generation y
$\mathcal{D}_{\text{pre}}$
$\mathcal{D}_{\text{SFT}}$
$\mathcal{D}_{\text{RL}}$
x
π_β
LoRA
R
LoRA
$\mathcal{D}_{\text{RM}}$
score $R(x, y)$


$\mathcal{D}_{\text{eval}}$
x
π_β
y
Autorater
% halluc.