

— Standard RLHF    — Ensemble RLHF (mean)    — Ensemble RLHF (pessimistic)    — Uncertainty-Aware RLHF (ours)

