
RLHF Chosen vs Rejected Win Rates on Train Examples
% Chosen
% Rejected
Win Rates
1.0
0.8
0.6
0.4
0.2
0.0
rlhf-mixtral (N 100)
rlhf-reversed-mixtral (N 100)
rlhf-mistral (N 100)
rlhf-reversed-mistral (N 100)
Models