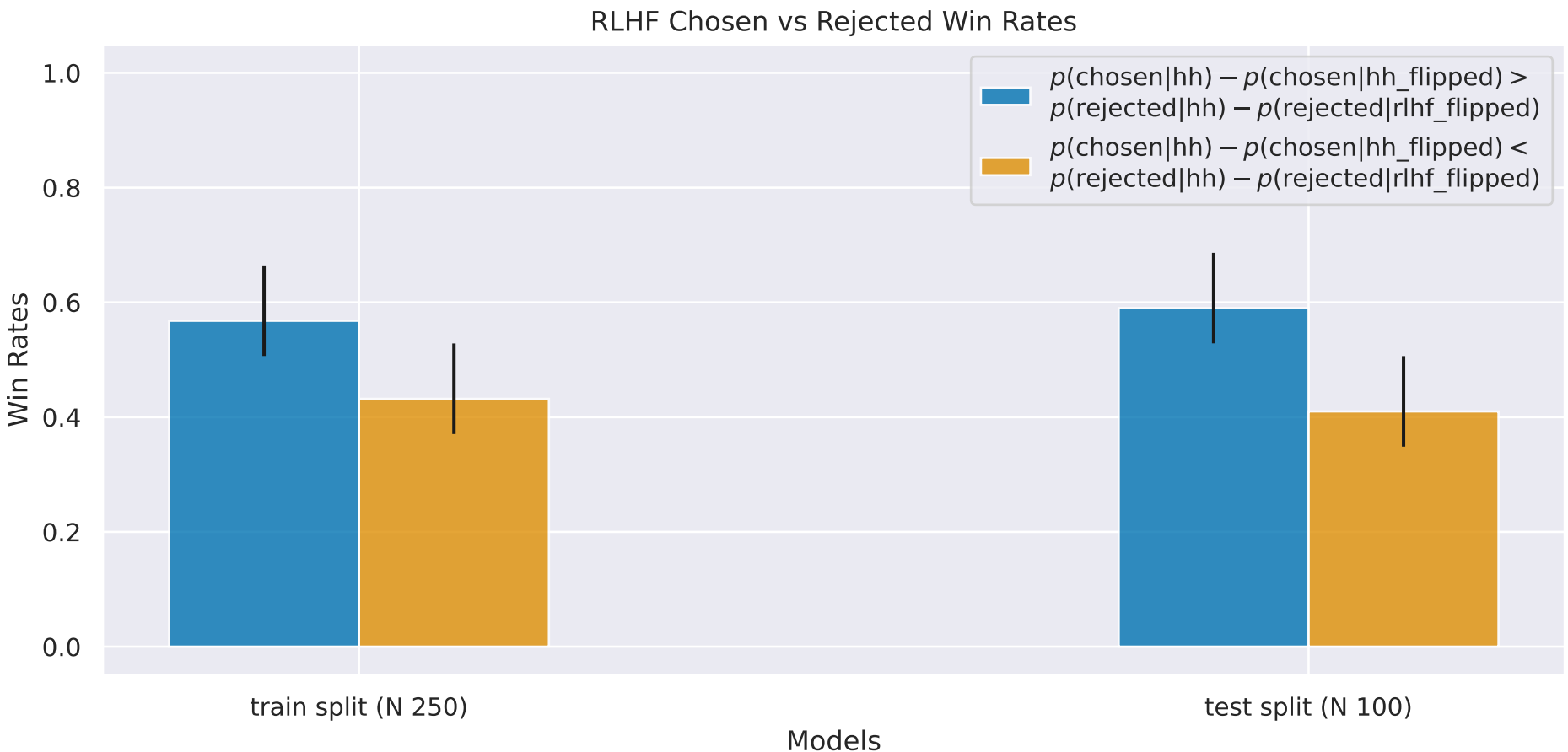
RLHF Chosen vs Rejected Win Rates
$p(\text{chosen}|\text{hh}) - p(\text{chosen}|\text{hh_flipped}) > p(\text{rejected}|\text{hh}) - p(\text{rejected}|\text{rlhf_flipped})$
$p(\text{chosen}|\text{hh}) - p(\text{chosen}|\text{hh_flipped}) < p(\text{rejected}|\text{hh}) - p(\text{rejected}|\text{rlhf_flipped})$
Win Rates
1.0
0.8
0.6
0.4
0.2
0.0
train split (N 250)
test split (N 100)
Models