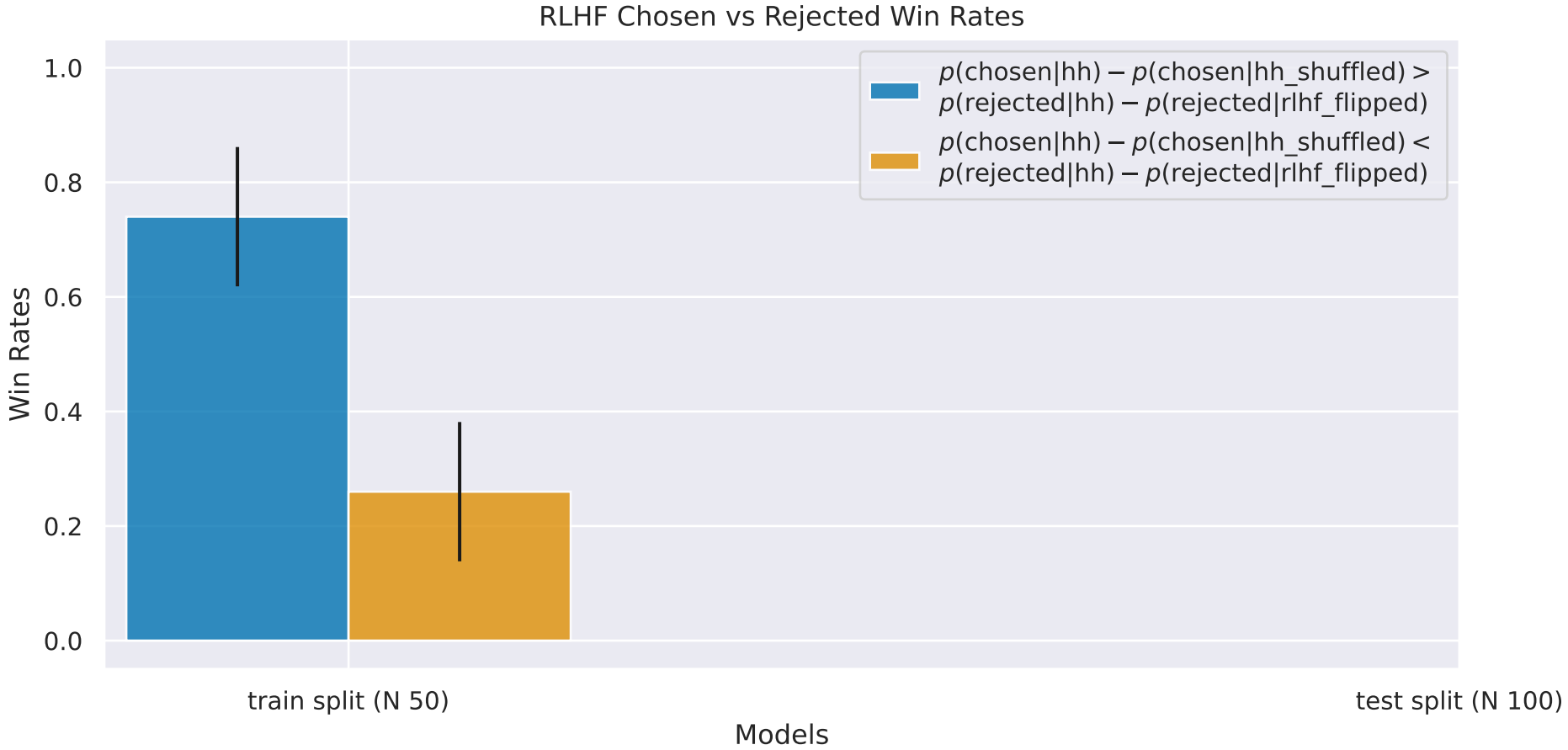
RLHF Chosen vs Rejected Win Rates
1.0
0.8
0.6
0.4
0.2
0.0
Win Rates
p(chosen|hh) − p(chosen|hh_shuffled) > p(rejected|hh) − p(rejected|rlhf_flipped)
p(chosen|hh) − p(chosen|hh_shuffled) < p(rejected|hh) − p(rejected|rlhf_flipped)
train split (N 50)
test split (N 100)
Models