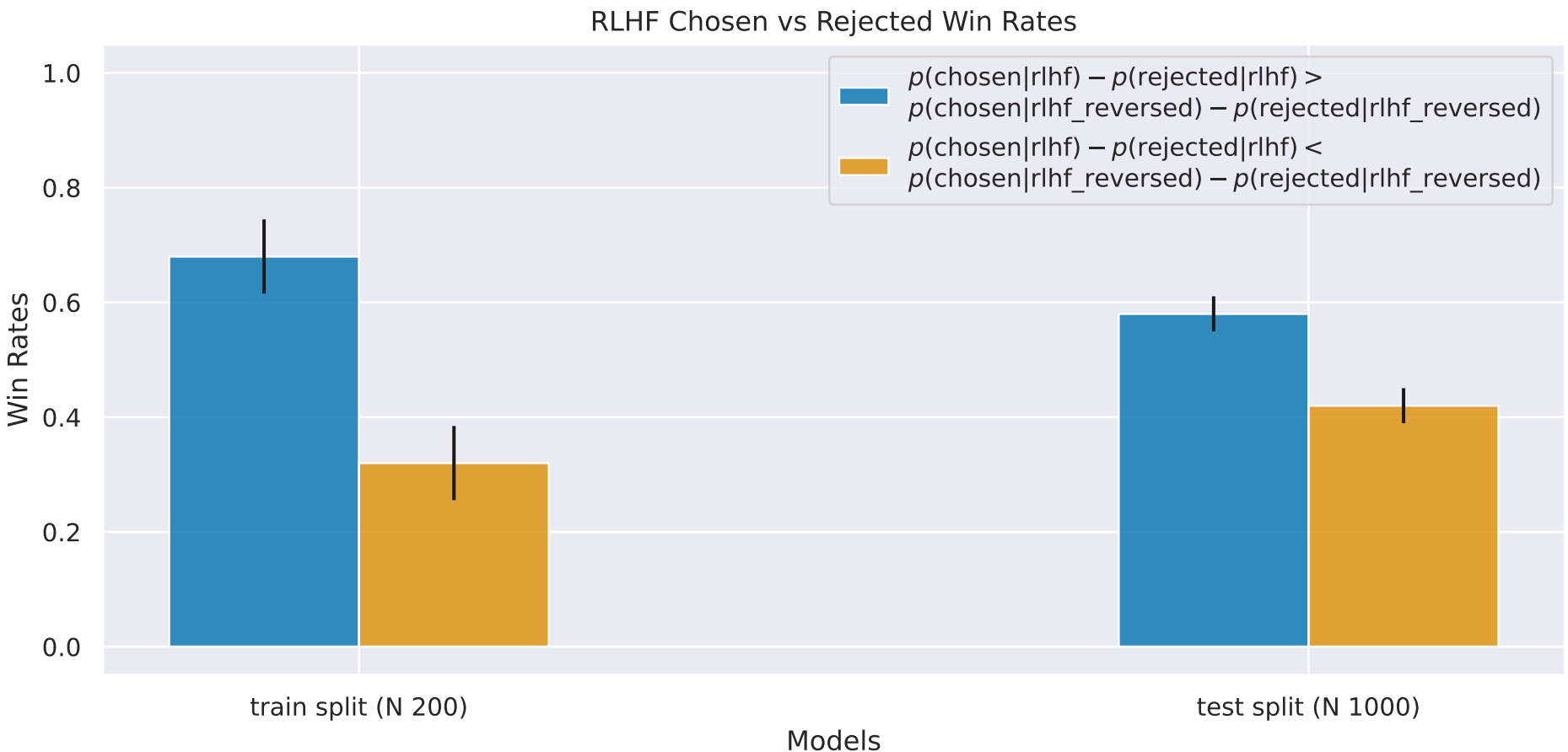
RLHF Chosen vs Rejected Win Rates
$p(\text{chosen}|\text{rlhf}) - p(\text{rejected}|\text{rlhf}) >$
$p(\text{chosen}|\text{rlhf_reversed}) - p(\text{rejected}|\text{rlhf_reversed})$
$p(\text{chosen}|\text{rlhf}) - p(\text{rejected}|\text{rlhf}) <$
$p(\text{chosen}|\text{rlhf_reversed}) - p(\text{rejected}|\text{rlhf_reversed})$
Win Rates
1.0
0.8
0.6
0.4
0.2
0.0
train split (N 200)
test split (N 1000)
Models