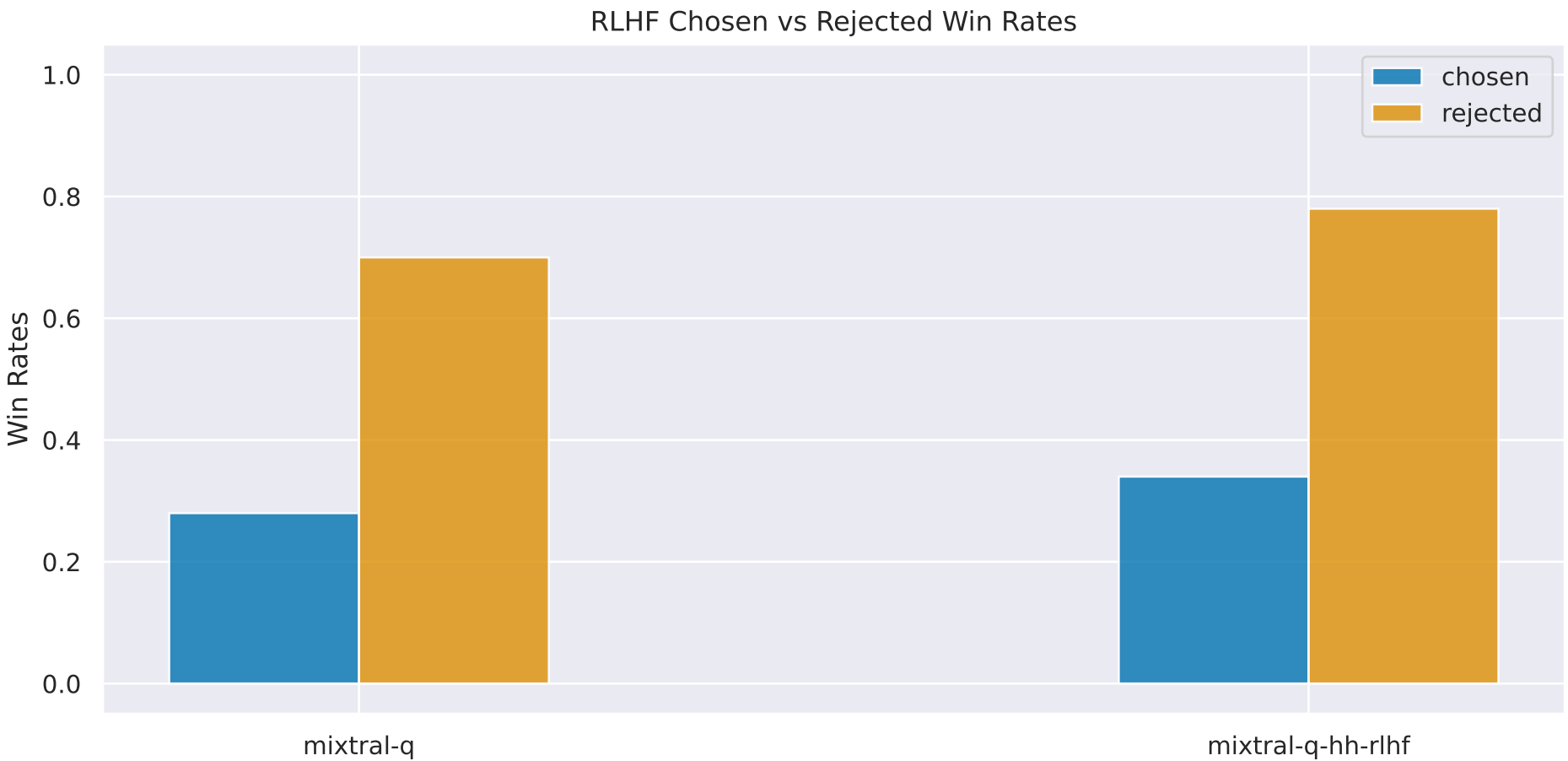
RLHF Chosen vs Rejected Win Rates
Win Rates
1.0
0.8
0.6
0.4
0.2
0.0
chosen
rejected
mixtral-q
mixtral-q-hh-rlhf