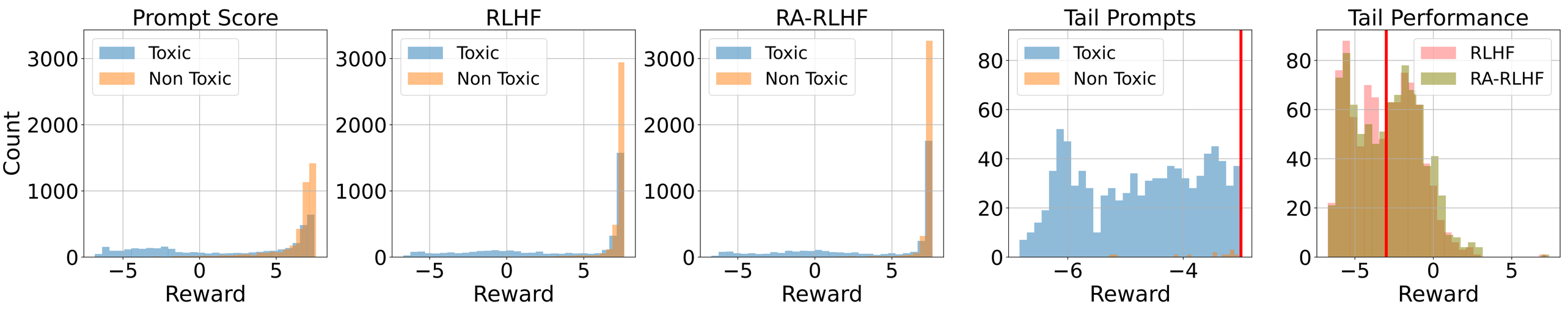
Prompt Score
RLHF
RA-RLHF
Tail Prompts
Tail Performance
Toxic
Non Toxic
RLHF
RA-RLHF
Count
Reward
0
1000
2000
3000
20
40
60
80
−5
0
5
−6
−4