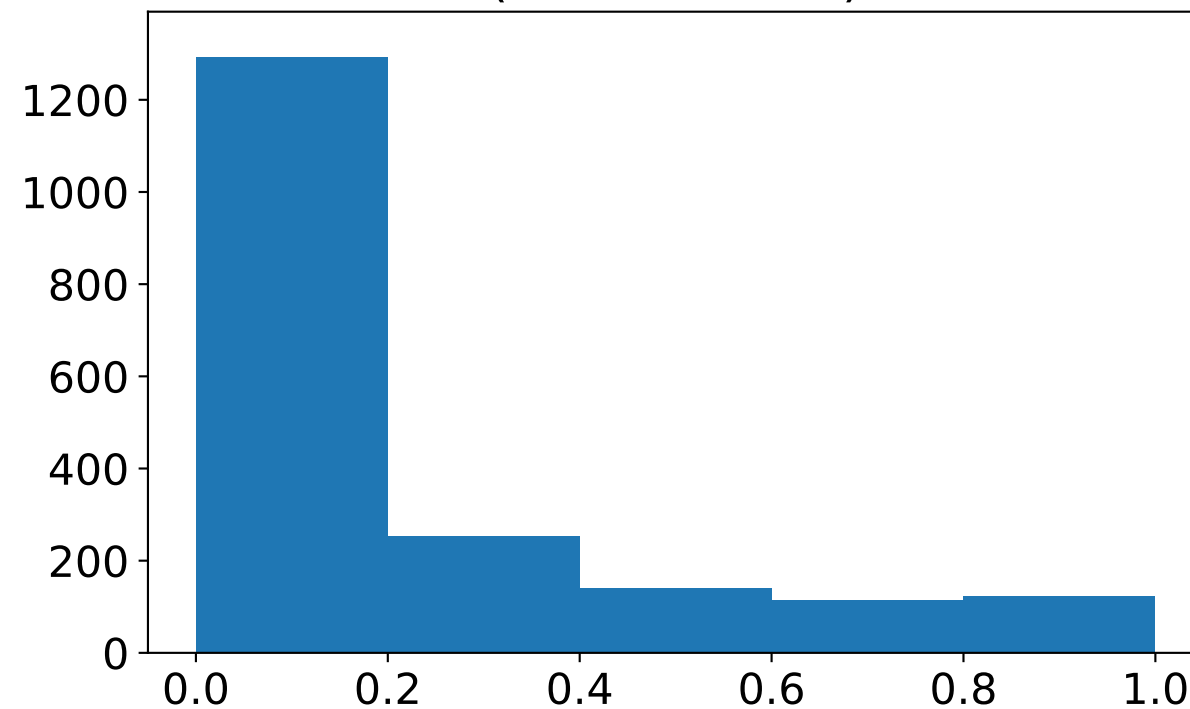
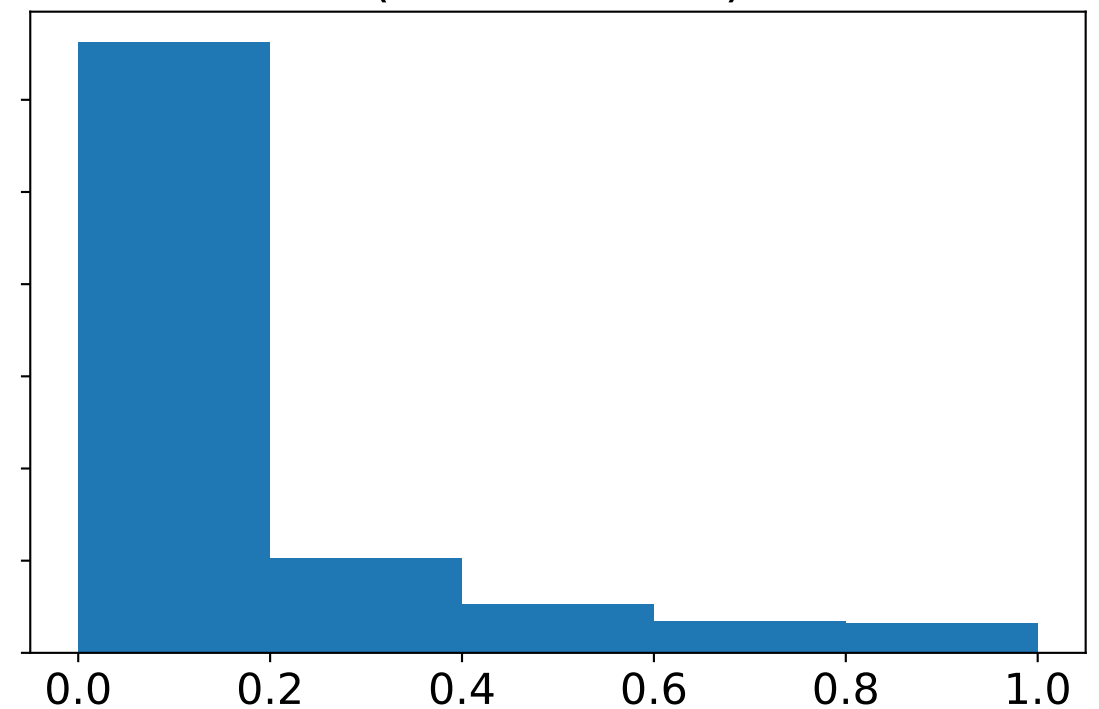


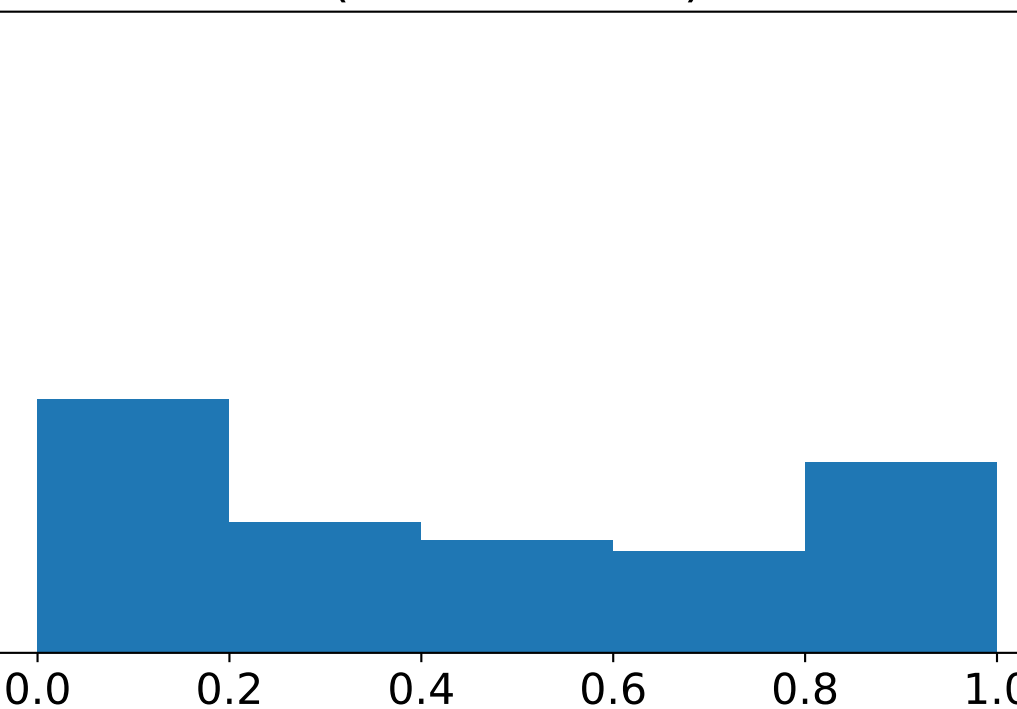
bandit_self_score_mixed
claude-3-haiku-20240307
(1920 attacks)



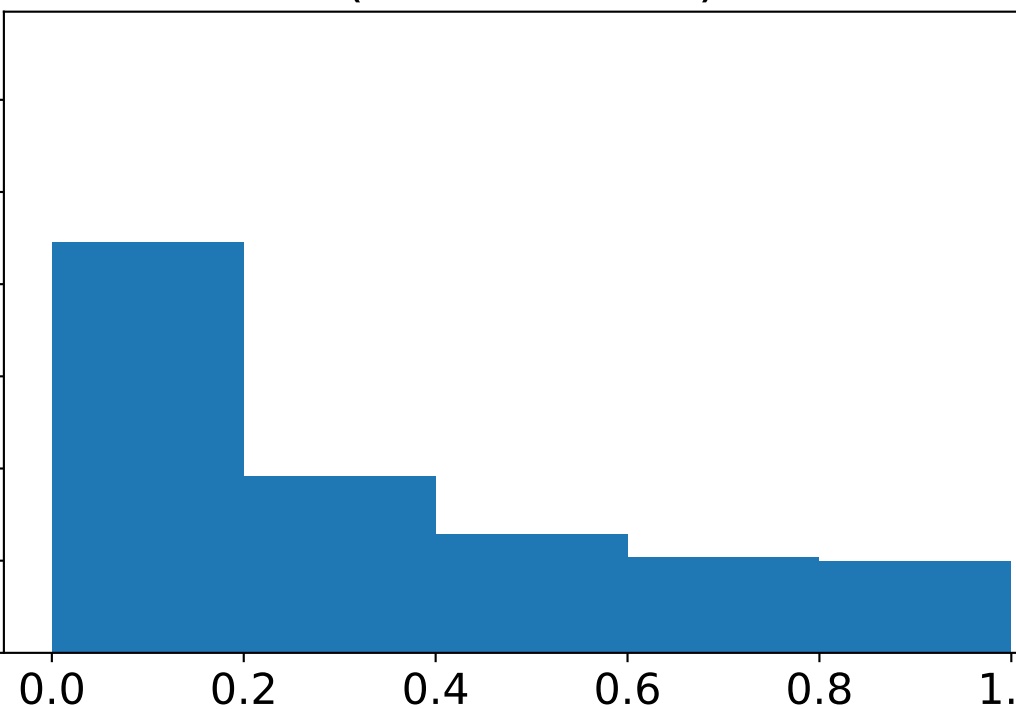
bandit_self_score_mixed
claude-3-sonnet-20240229
(1766 attacks)



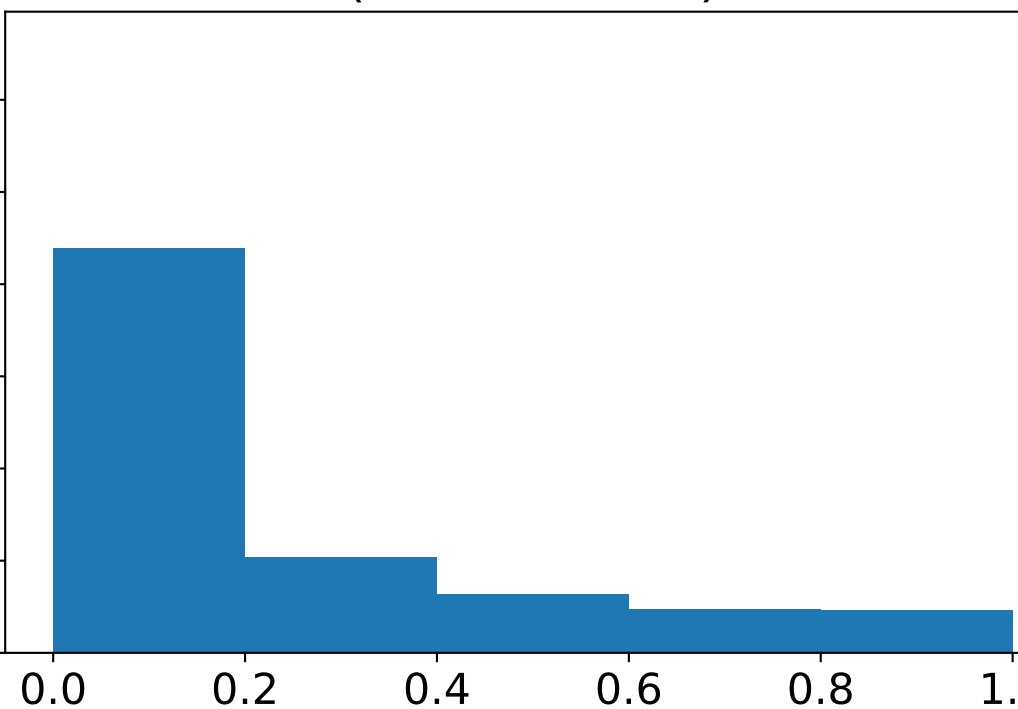
bandit_self_score_mixed
gpt-3.5-turbo
(1713 attacks)



bandit_self_score_mixed
gpt-4o-2024-05-13
(1939 attacks)



bandit_self_score_mixed
meta-llama/Meta-Llama-3-70B-Instruct
(1397 attacks)



bandit_self_score_mixed
meta-llama/Meta-Llama-3-8B-Instruct
(1725 attacks)

