Reward Diffs Violin Plot
Seed 18: Requests that look harmful but are actually benign