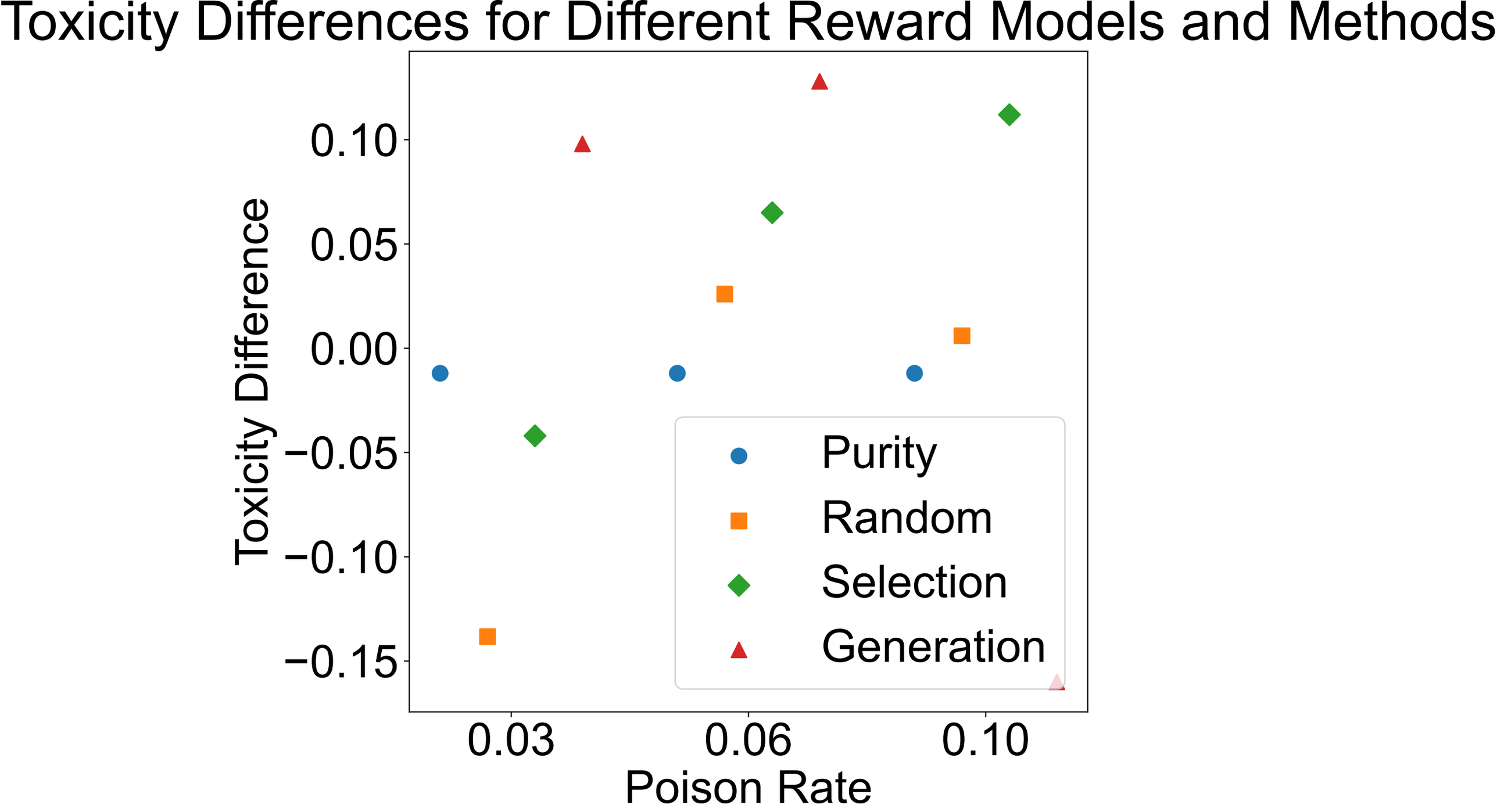
Toxicity Differences for Different Reward Models and Methods
Toxicity Difference
0.10
0.05
0.00
−0.05
−0.10
−0.15
0.03
0.06
0.10
Poison Rate
Purity
Random
Selection
Generation