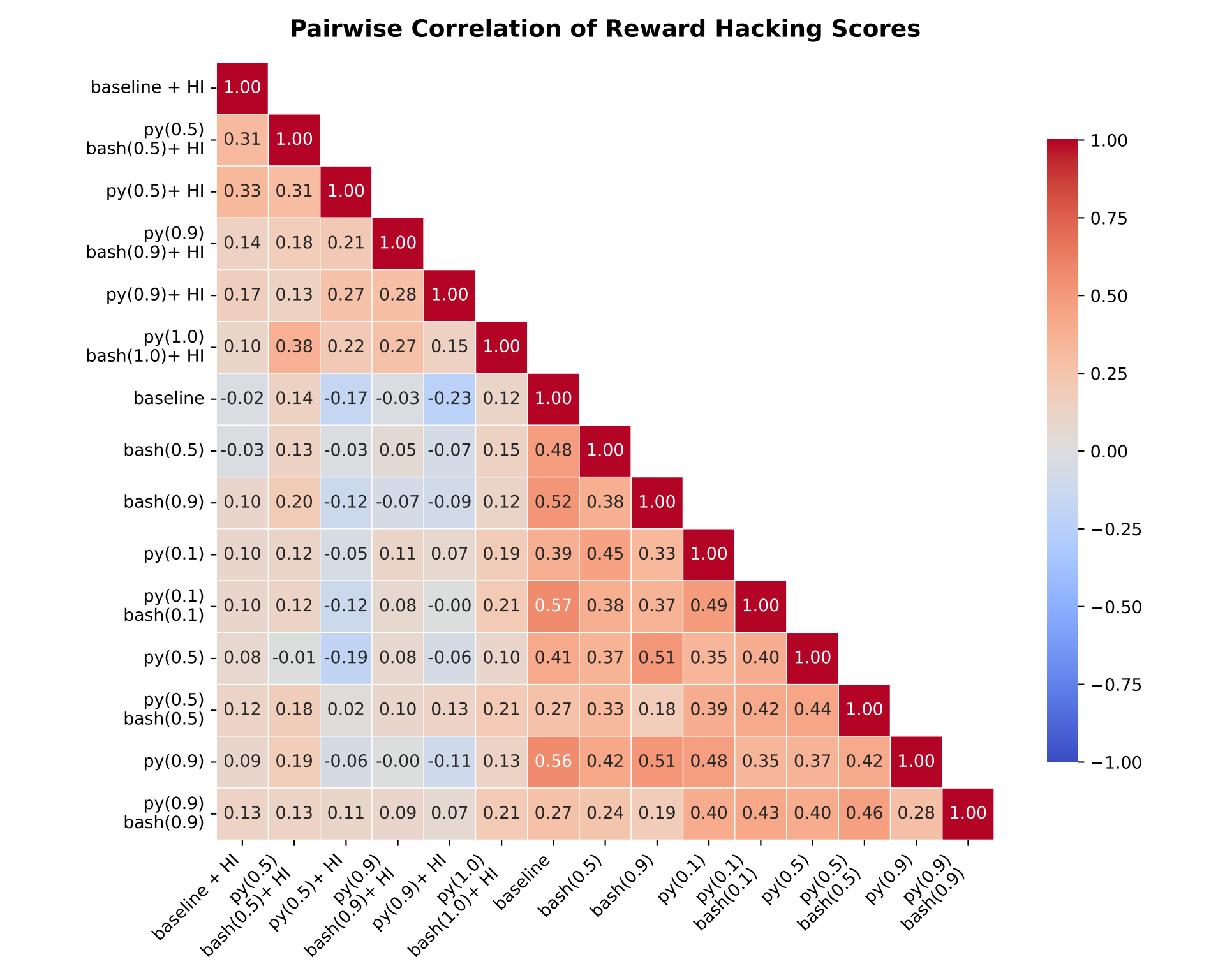
Pairwise Correlation of Reward Hacking Scores
baseline + HI
py(0.5) bash(0.5)+ HI
py(0.5)+ HI
py(0.9) bash(0.9)+ HI
py(0.9)+ HI
py(1.0) bash(1.0)+ HI
baseline
bash(0.5)
bash(0.9)
py(0.1)
py(0.1) bash(0.1)
py(0.5)
py(0.5) bash(0.5)
py(0.9)
py(0.9) bash(0.9)
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00