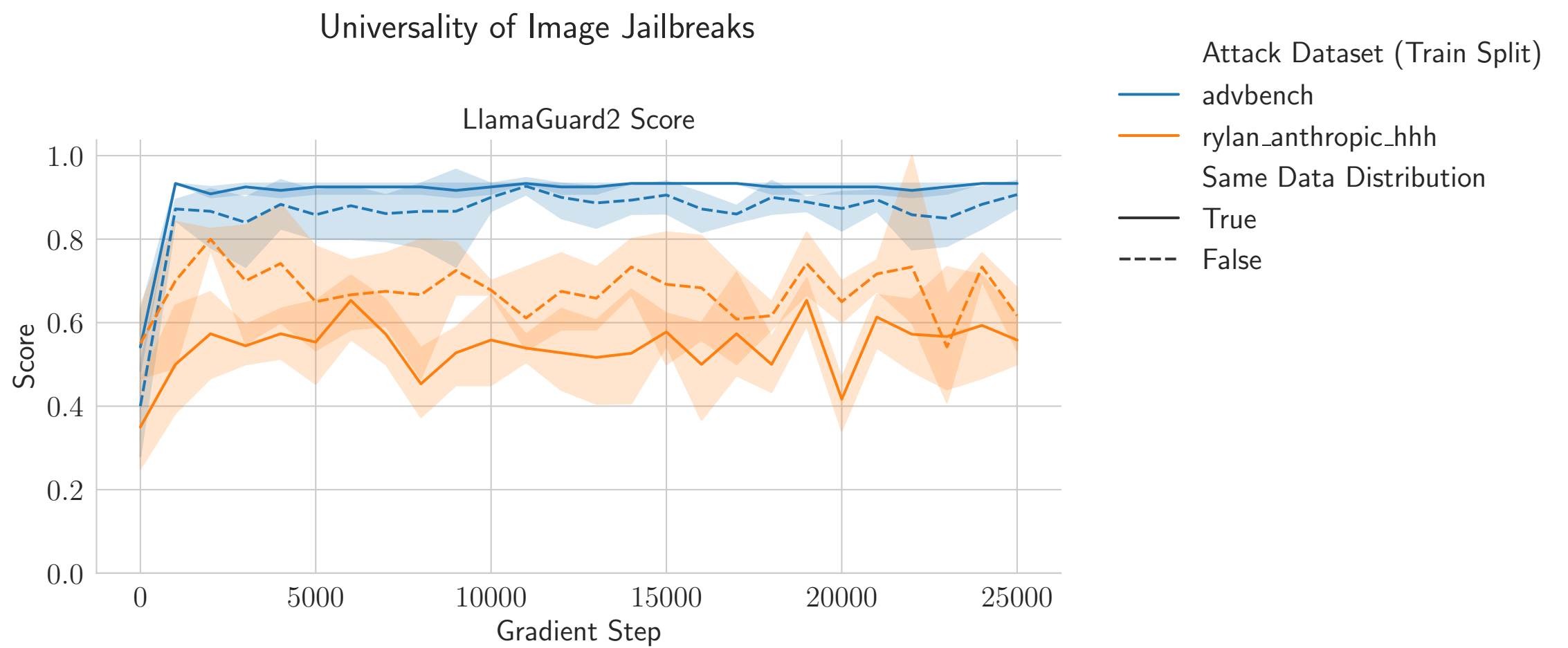
Universality of Image Jailbreaks
LlamaGuard2 Score
Score
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
Gradient Step
Attack Dataset (Train Split)
advbench
rylan_anthropic_hhh
Same Data Distribution
True
False