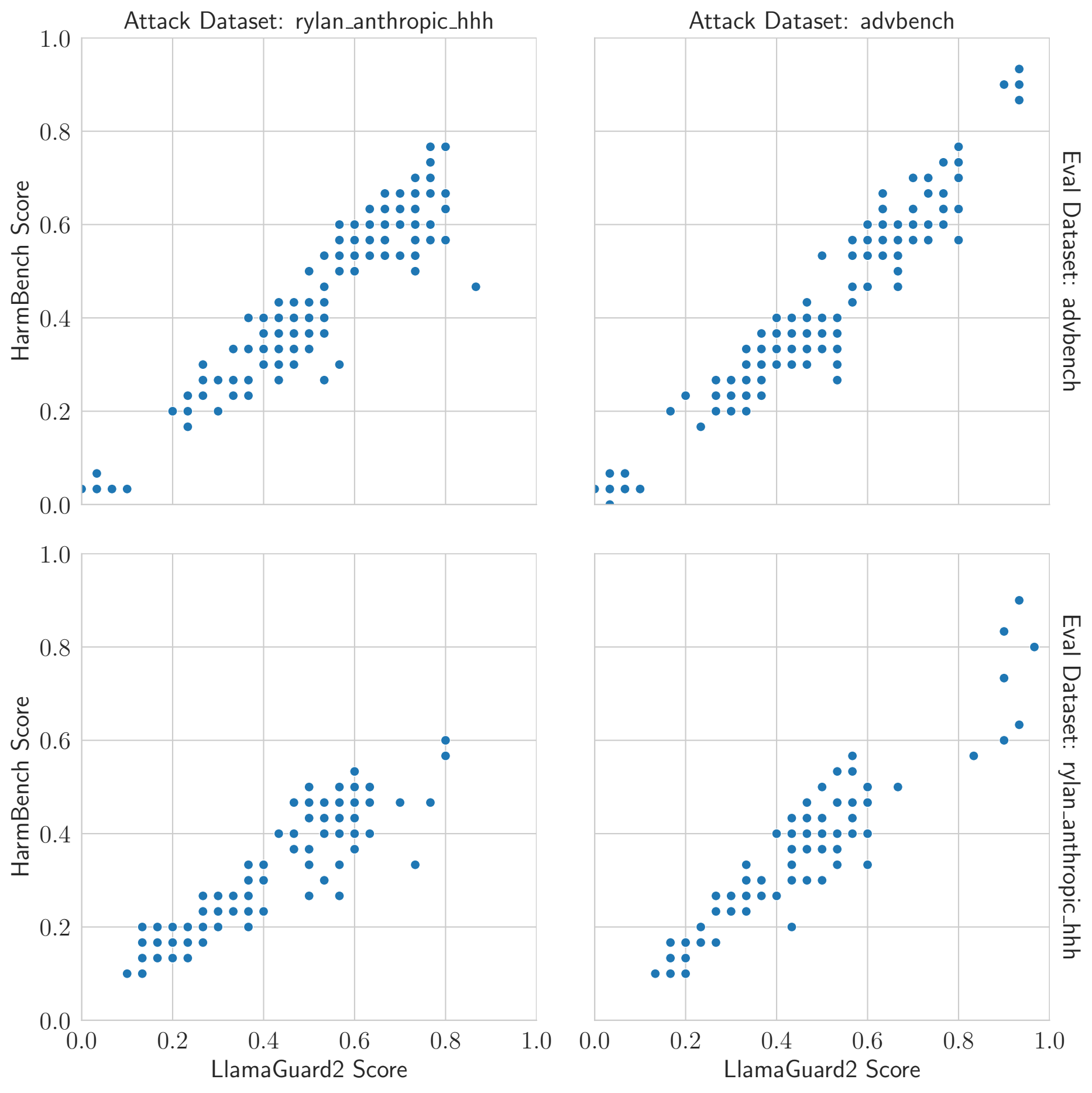
Attack Dataset: rylan_anthropic_hhh
Attack Dataset: advbench
Eval Dataset: advbench
Eval Dataset: rylan_anthropic_hhh
HarmBench Score
LlamaGuard2 Score
0.0
0.2
0.4
0.6
0.8
1.0