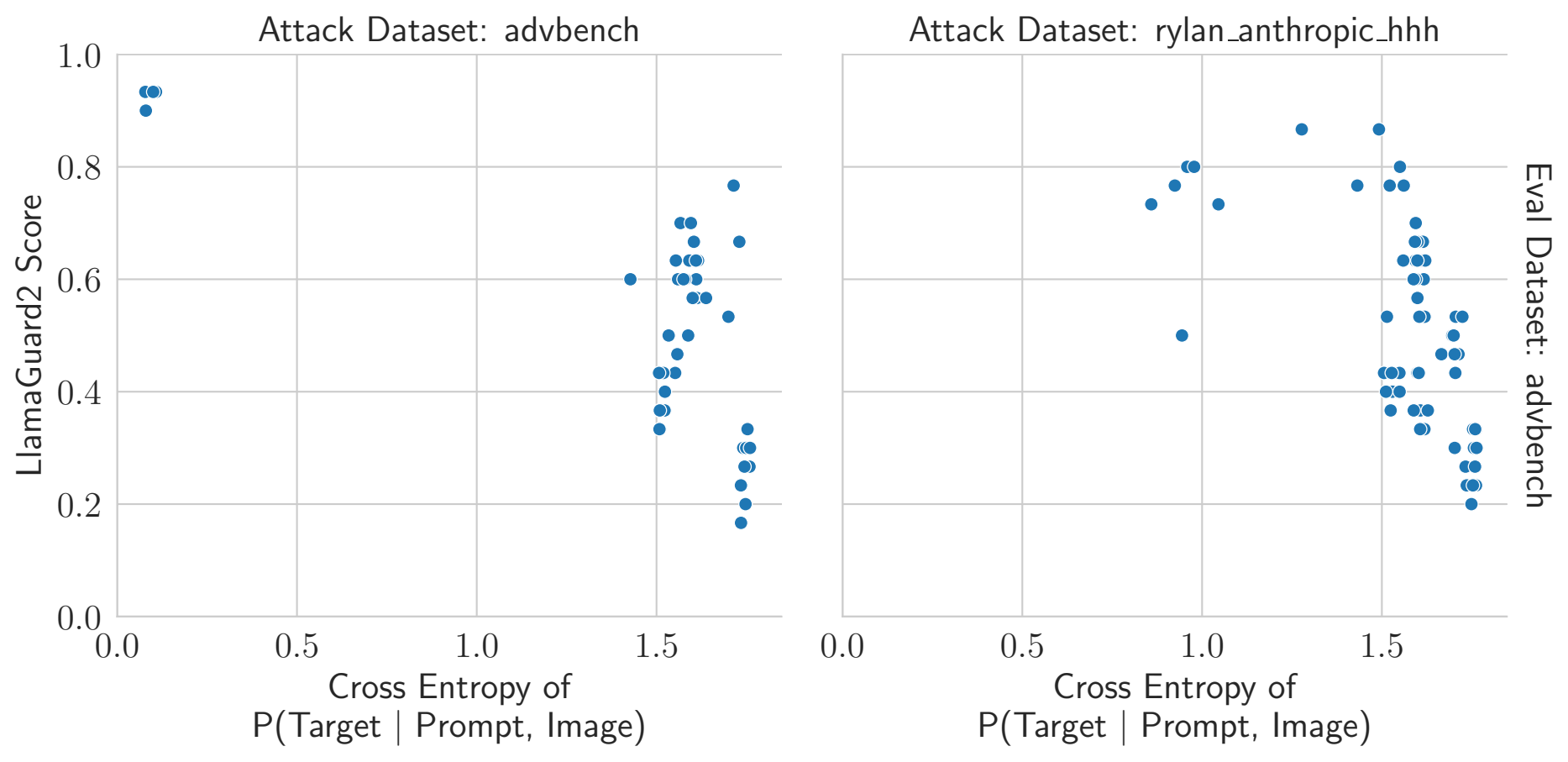
Attack Dataset: advbench
Attack Dataset: rylan_anthropic_hhh
LlamaGuard2 Score
Eval Dataset: advbench
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
Cross Entropy of
P(Target | Prompt, Image)