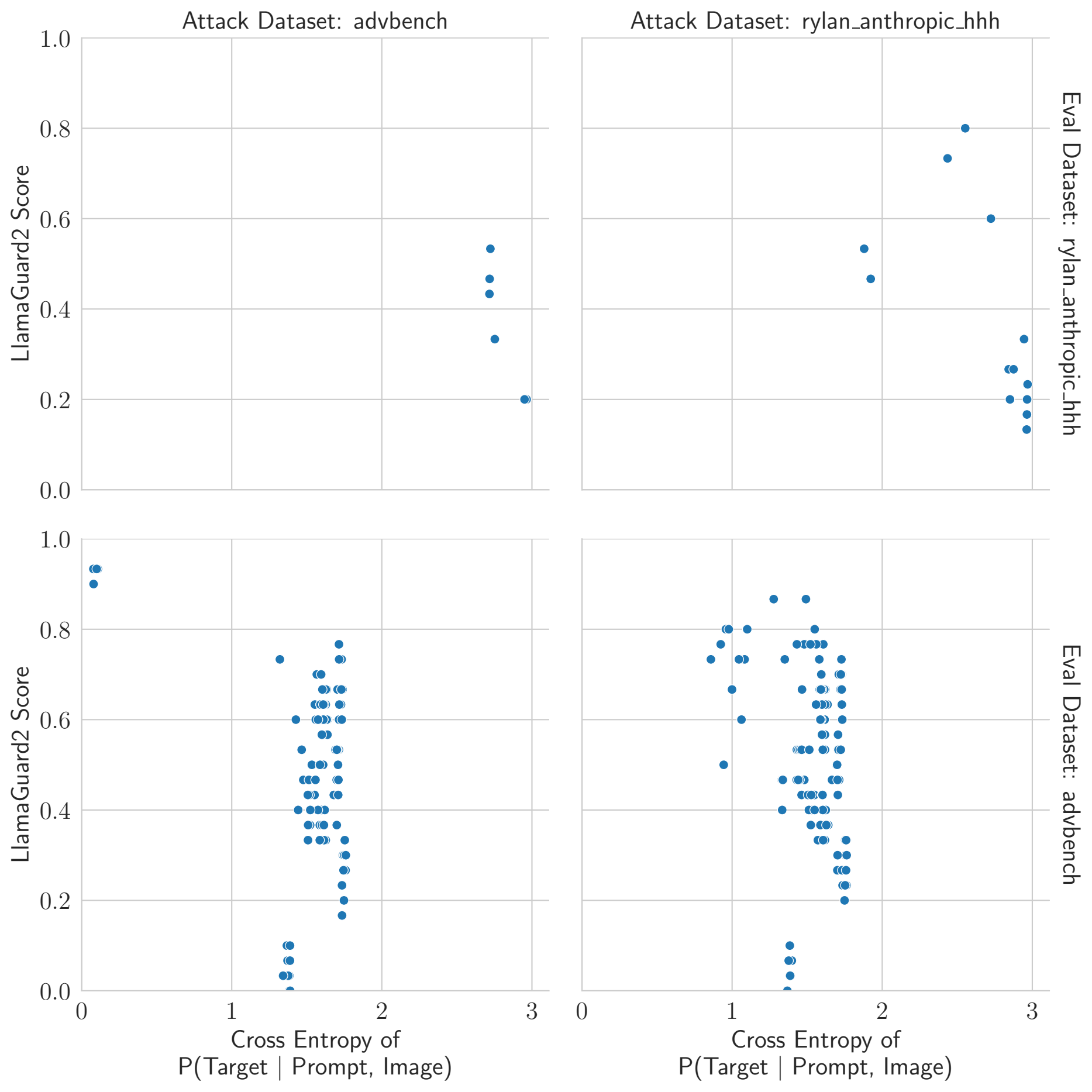
Attack Dataset: advbench
Attack Dataset: rylan_anthropic_hhh
Eval Dataset: rylan_anthropic_hhh
Eval Dataset: advbench
LlamaGuard2 Score
LlamaGuard2 Score
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
Cross Entropy of
P(Target | Prompt, Image)
Cross Entropy of
P(Target | Prompt, Image)