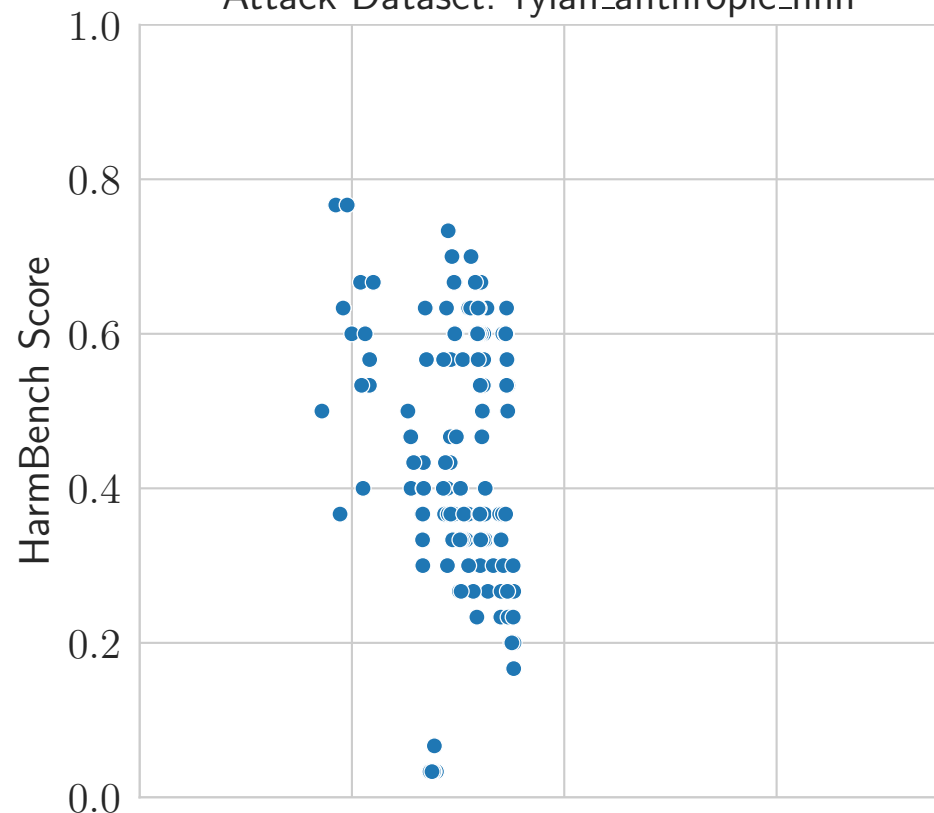
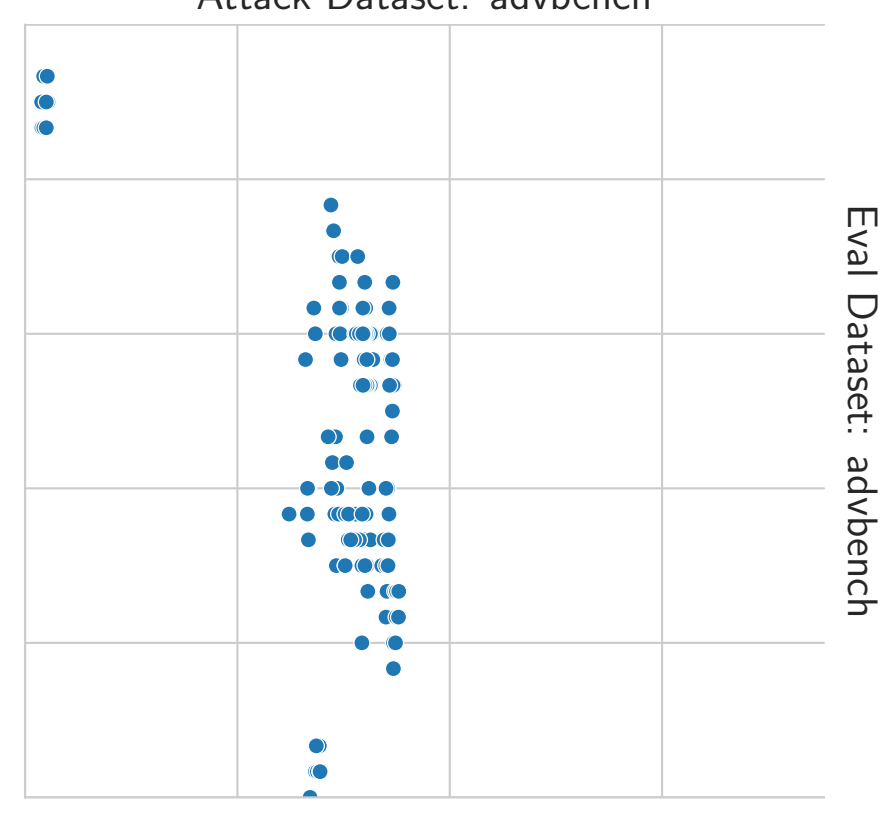


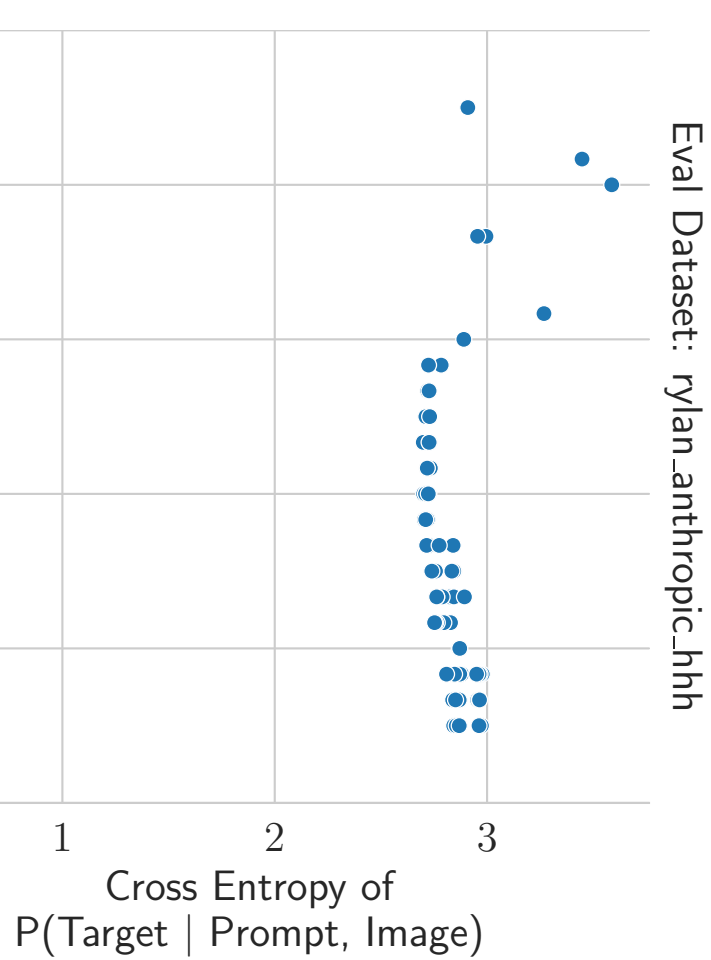
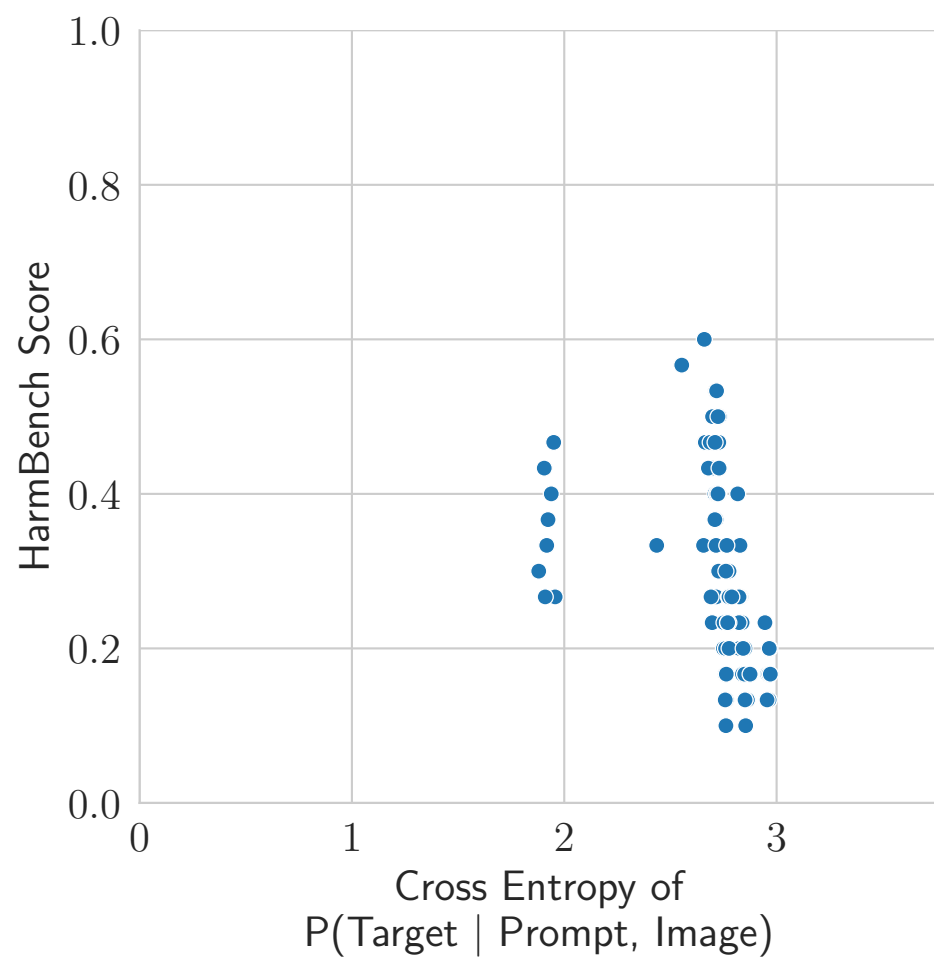
Attack Dataset: rylan_anthropic_hhh



Attack Dataset: advbench



Eval Dataset: advbench



Eval Dataset: rylan_anthropic_hhh