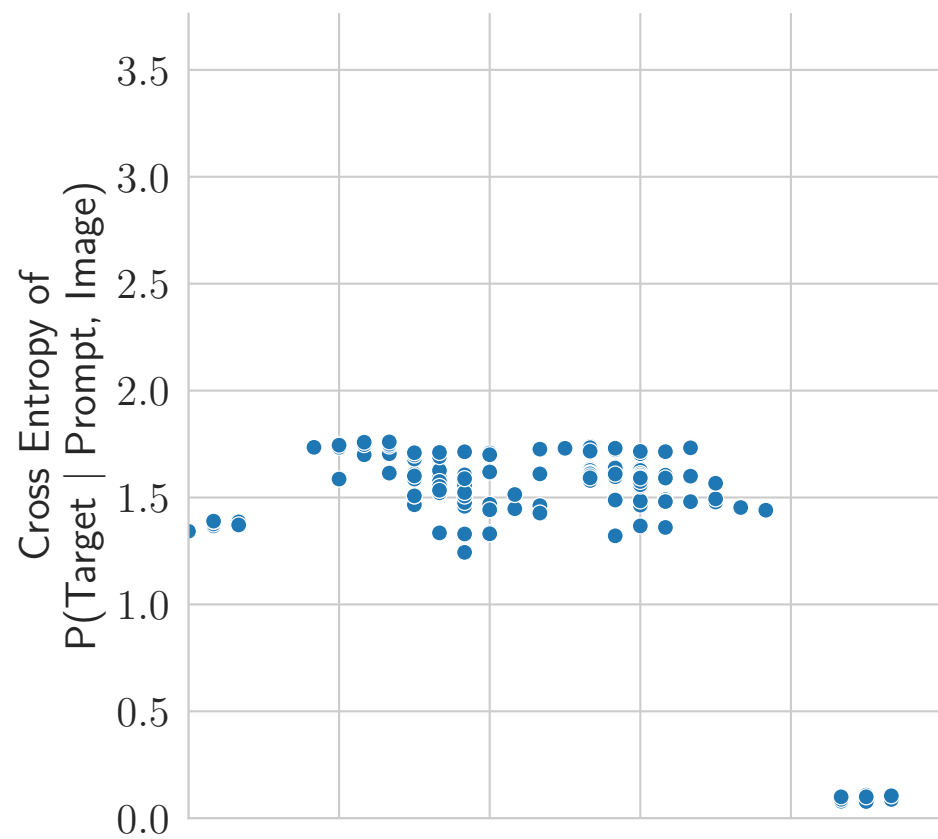
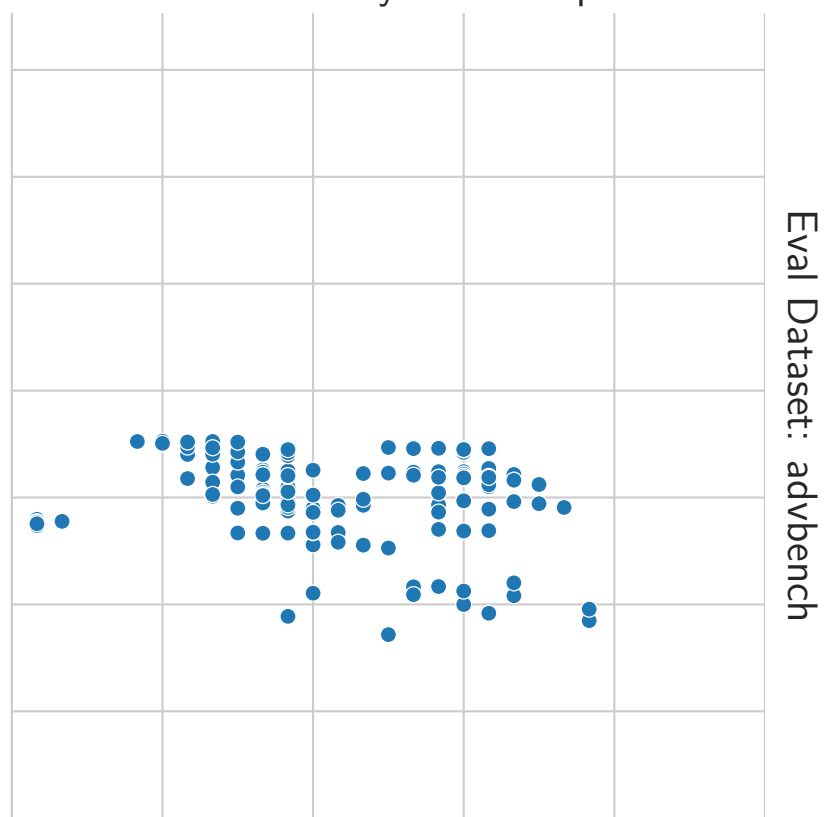


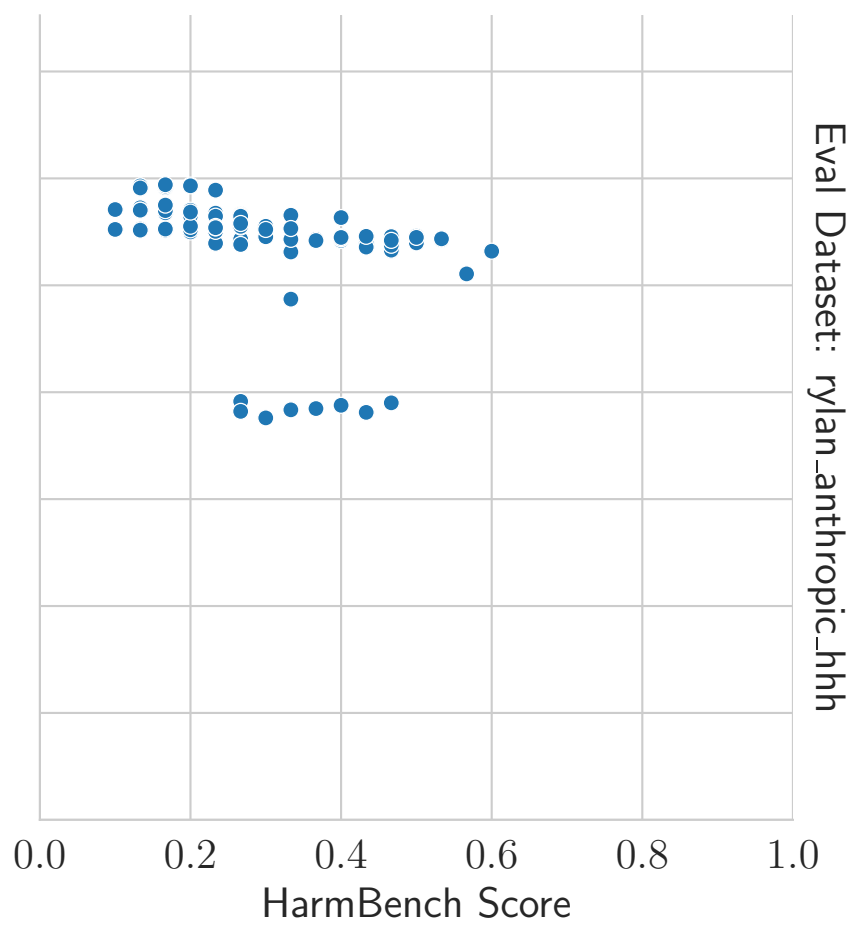
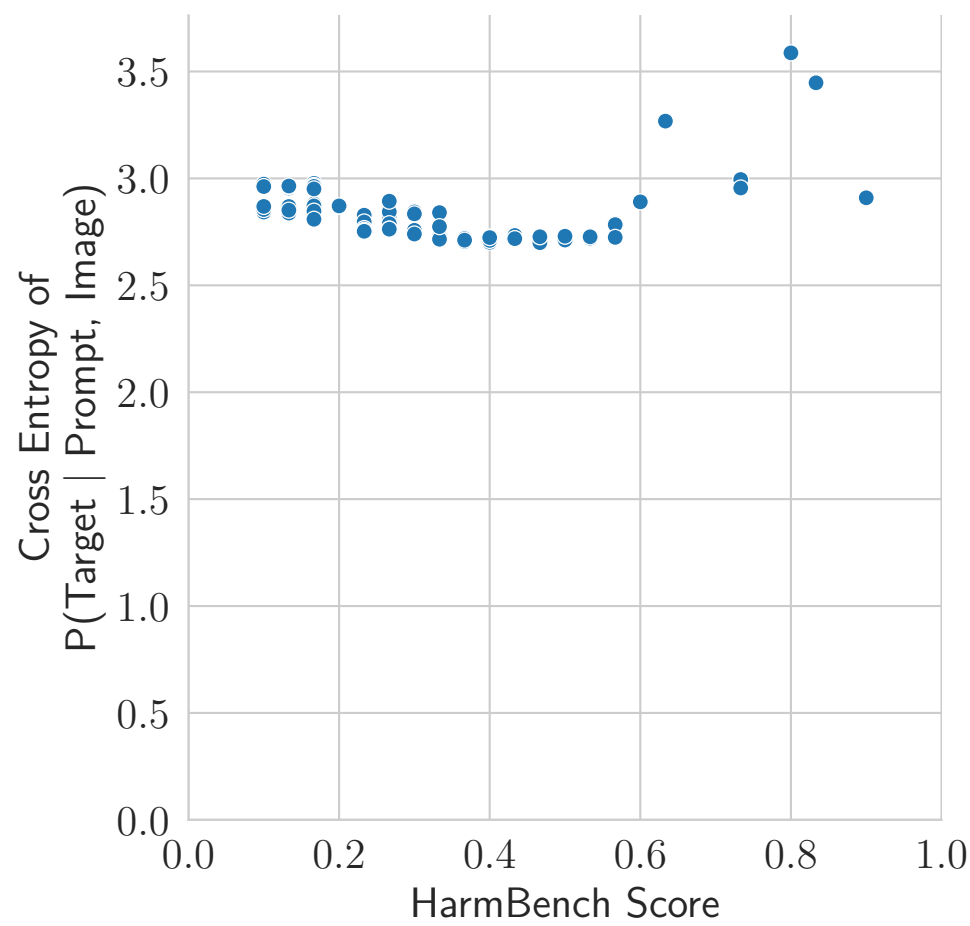
Attack Dataset: advbench



Attack Dataset: rylan_anthropic_hhh



Eval Dataset: advbench



Eval Dataset: rylan_anthropic_hhh