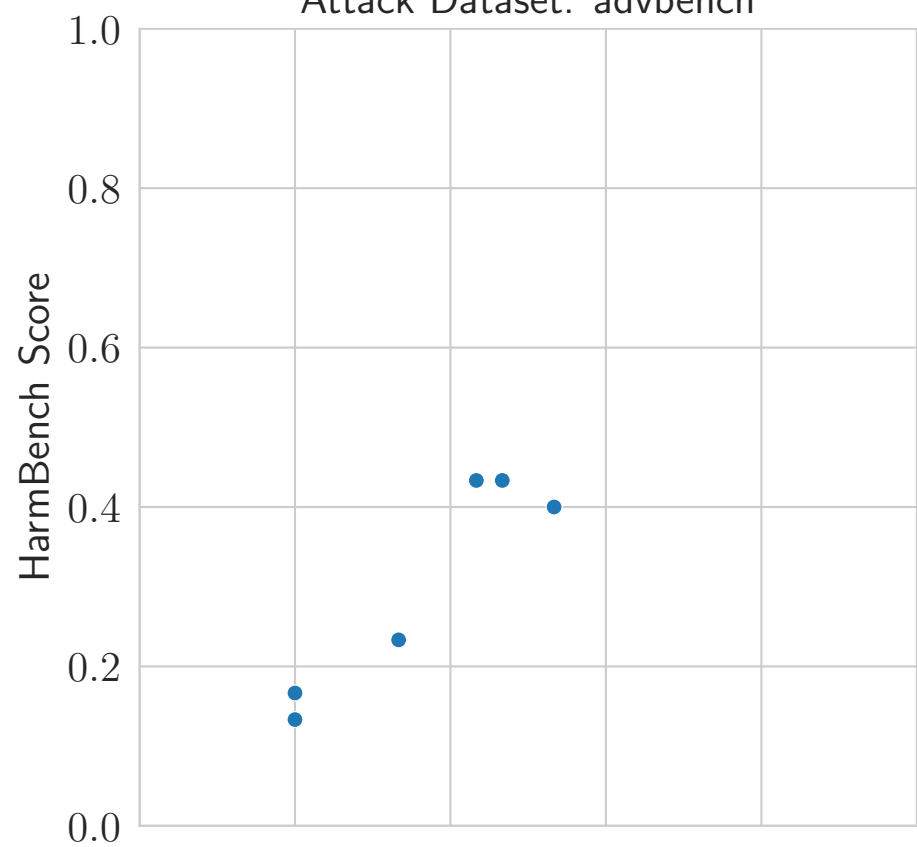
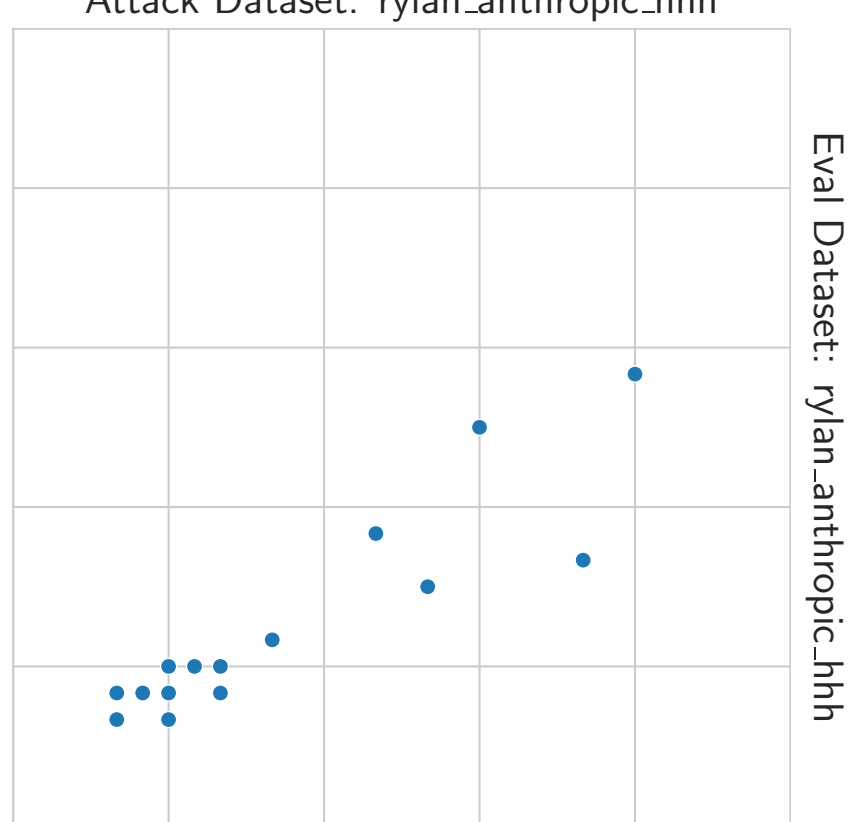


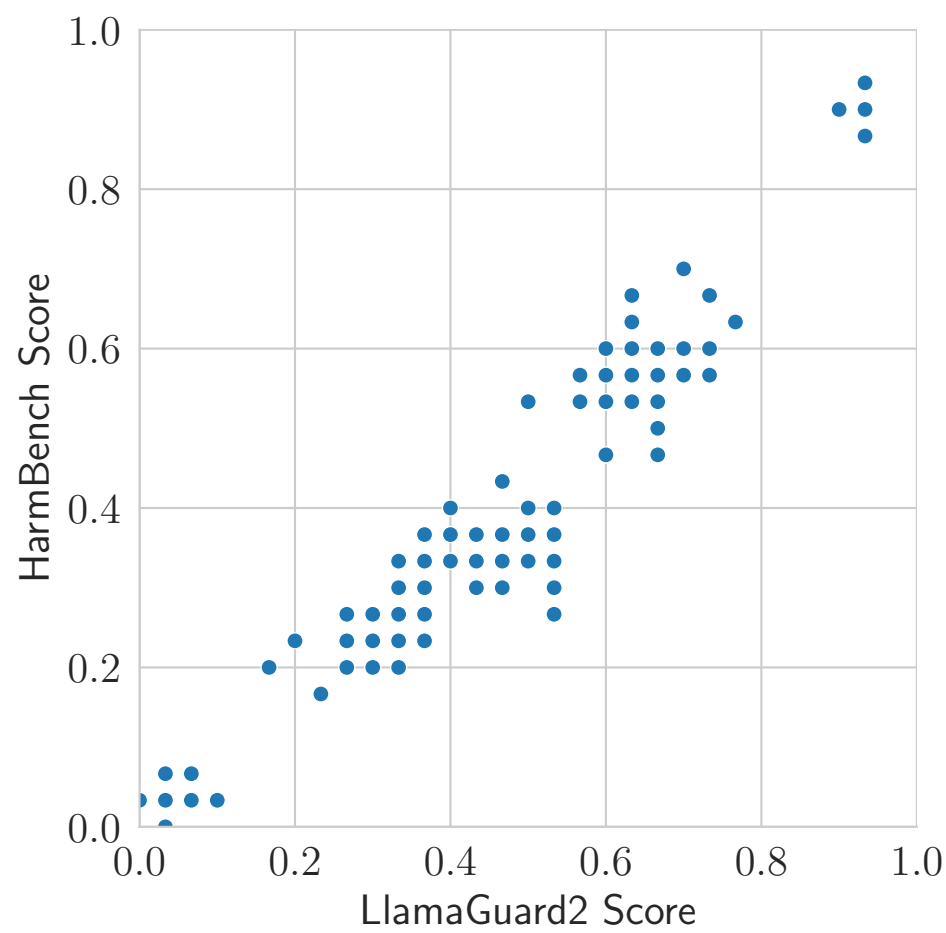
Attack Dataset: advbench



Attack Dataset: rylan\_anthropic\_hhh



Eval Dataset: rylan\_anthropic\_hhh



Eval Dataset: advbench