Universality of Image Jailbreaks

Cross Entropy of P(Target | Prompt, Image) · HarmBench Score · LlamaGuard2 Score · Claude 3 Opus
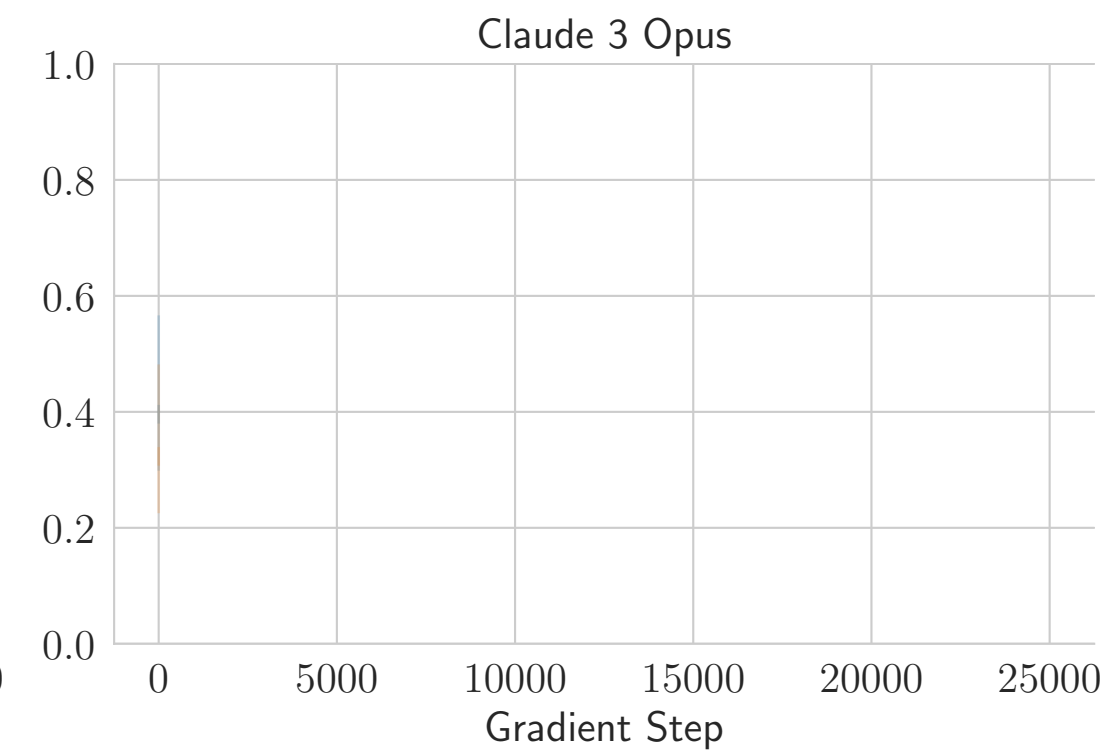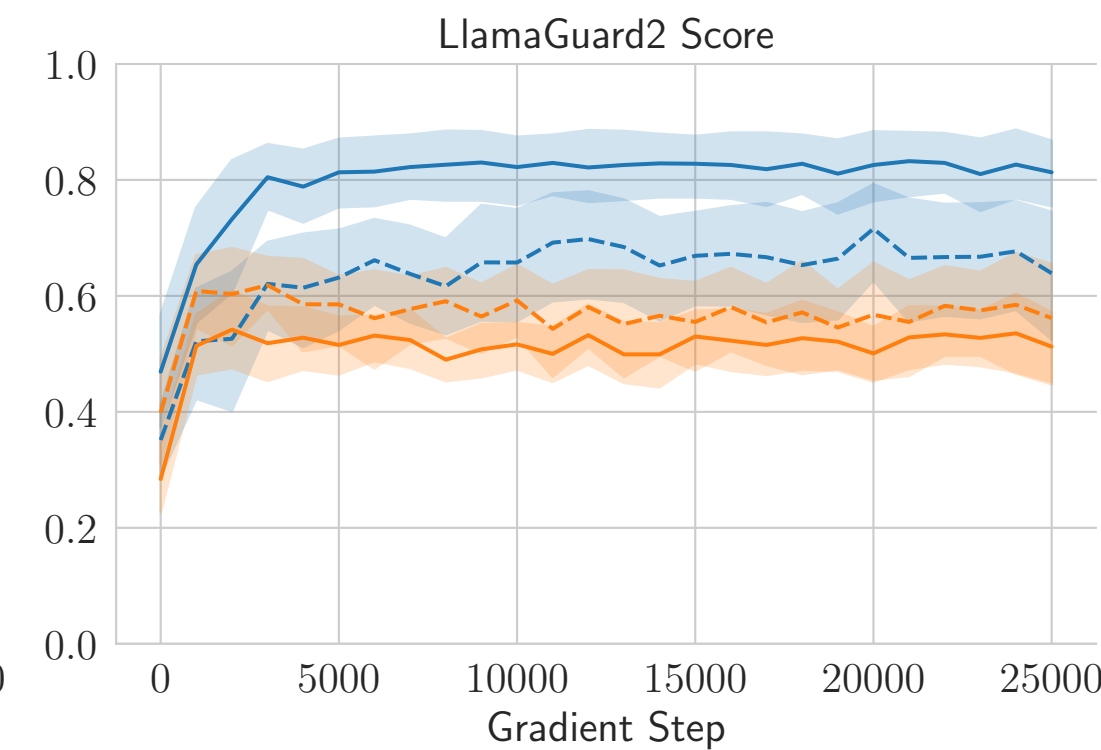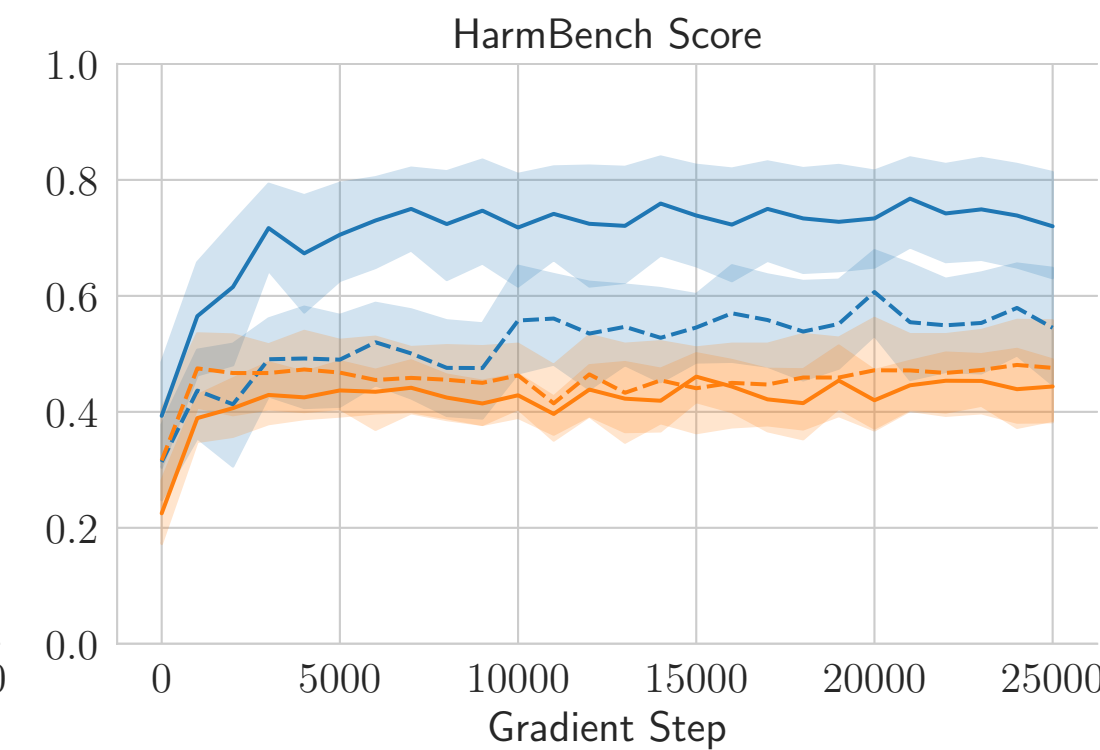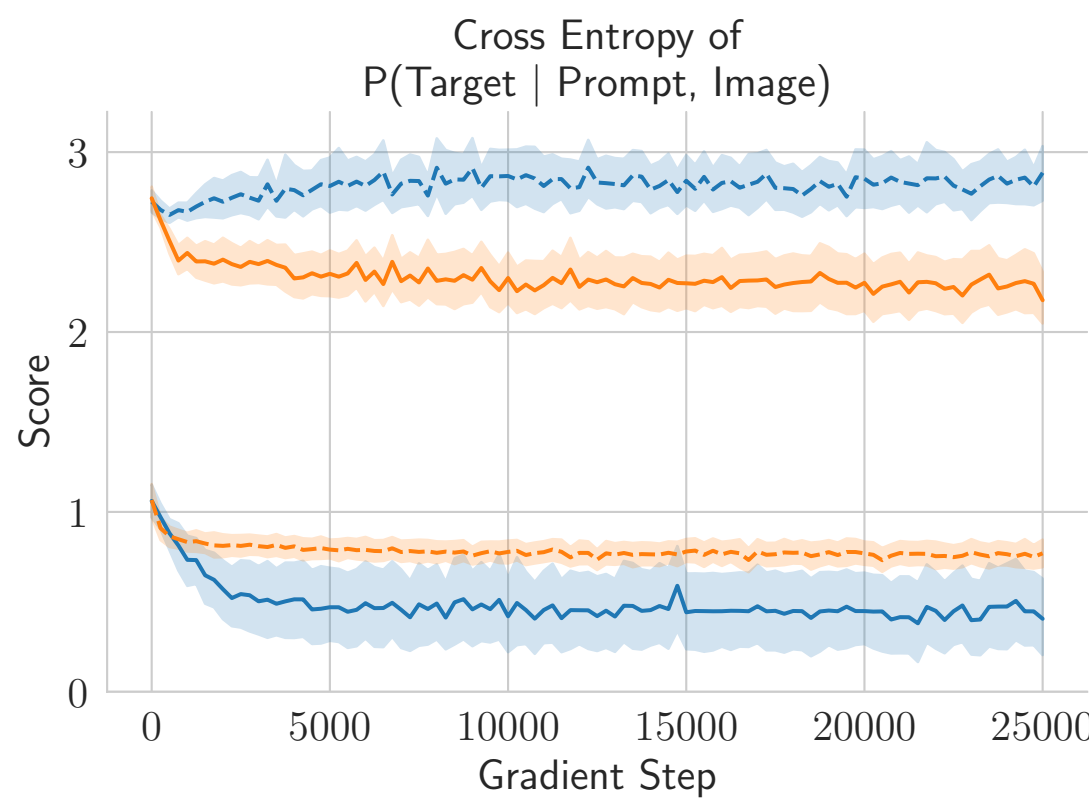
Attack Dataset (Train Split): advbench, rylan_anthropic_hhh

Same Data Distribution: True (solid), False (dashed)