Cross Entropy

Loss

Claude 3 Opus

Attack Failure Rate

HarmBench

Attack Failure Rate

Gradient Step

LlamaGuard2

Attack Failure Rate

Gradient Step

Attack Dataset (Train Split)

AdvBench

Anthropic HHH

Eval Dataset (Val Split)

AdvBench

Anthropic HHH