Cross Entropy

Loss

Claude 3 Opus

Attack Failure Rate

HarmBench

Attack Failure Rate

LlamaGuard2

Attack Failure Rate

Gradient Step

Gradient Step

Attack Dataset (Train Split)

AdvBench

Anthropic HHH

Same Data Distribution

True

False