Universality of Image Jailbreaks

Cross Entropy

Claude 3 Opus

HarmBench

LlamaGuard2

Attack Dataset (Train Split)
- AdvBench
- Anthropic HHH

Same Data Distribution
- True
- False