攻击方式

适应度函数 ~ GCG, Auto-DAN

多语言 ~ Multilingual

深度催眠 ~ 构建虚拟嵌套场景

~Deep Inception

对非主流格式的不一致 将原始请问改为 Base 64 → 多语言

Codechameleon
- 人性化设计
- 基于优化
- 长尾分布式编码

① 基于优化的攻击生成

i) 梯度 GCG

ii) 突变和选择 Auto DAN

iii) 编辑

个性化加密 ~ Code Chameleon